

Bruno Régaldo-Saint Blancard

Master 2 Recherche Astronomie, Astrophysique et Ingénierie Spatiale



Rapport de stage de Master 2

Vers une description statistique du milieu interstellaire magnétisé

12 mars 2018 - 22 juin 2018



Laboratoire : Laboratoire de Radio-Astronomie, Ecole Normale Supérieure, Paris
LERMA, Observatoire de Paris
Tuteurs de stage : François Levrier, François Boulanger
Référent Master : Stéphane Corbel

Remerciements

Je tiens à remercier toutes les personnes qui ont permis de rendre ce stage aussi passionnant qu'enrichissant.

Je remercie mes encadrants François Levrier et François Boulanger qui ont su m'accorder leur confiance pour mener à bien ce travail, qui se sont montrés toujours disponibles à chaque étape de mon travail et ont fait preuve d'une extrême bienveillance à mon égard. Je remercie Erwan Allys, le troisième encadrant officieux de mon stage, pour sa grande pédagogie et grâce à qui j'ai pu aller si loin dans l'exploration de mon sujet malgré la courte durée de ce stage. Je remercie Stéphane Mallat et Sixin Zhang, avec qui les discussions ont toujours été extrêmement profitables pour ce travail et m'ont permis de découvrir les problématiques fascinantes de la science des données.

Je remercie mes collègues et amis du département de Physique de l'ENS qui ont rendu mon quotidien des plus agréables. Merci à Thibaud, Jordan, Alice, Mélissa, Alexis, Michaël et Gaurav.

Merci également à toute l'équipe du Laboratoire de Radio-Astronomie et à son directeur Michel Perault, qui m'ont accueilli très chaleureusement au sein du groupe.

Je remercie mon référent de Master Stéphane Corbel pour sa disponibilité tout au long de l'année, ainsi que l'équipe enseignante du Master Astronomie Astrophysique et Ingénierie Spatiale pour leurs conseils et leur accompagnement qui m'ont guidé vers ce stage, et m'ont convaincu de poursuivre ma formation en thèse.

Abstract/Résumé

english

français

Introduction

En 2014, la collaboration BICEP annonçait avoir détecté dans le signal de polarisation du fond diffus cosmologique à 150 GHz des traces d’une prédiction majeure du modèle de l’inflation cosmique : l’existence d’ondes gravitationnelles primordiales [3]. Cette annonce éclatante et aux conséquences profondes sur notre connaissance des origines de l’Univers mettait un terme à une quête ayant occupé les cosmologistes des décennies durant... Mais presque immédiatement, la communauté scientifique a émis des doutes concernant l’évaluation d’un signal parasite présent dans les observations : le signal dû aux émissions polarisées de la poussière galactique, arguant que ce signal avait été largement sous-estimé dans l’analyse. Il a alors fallu un travail de concert des collaborations BICEP et Planck pour finalement confirmer ces doutes et revenir sur l’annonce initiale : la collaboration BICEP ne pouvait affirmer avoir détecté ces ondes gravitationnelles primordiales [4]. Cet épisode houleux de la recherche en cosmologie, outre son caractère pédagogique et incitant à la prudence dans toute analyse scientifique, a notamment permis de mettre le doigt sur une étape indispensable pour l’éventuelle détection d’ondes gravitationnelles primordiales dans le fond diffus cosmologique : la nécessité d’avoir une description statistique fine du signal dû au milieu interstellaire galactique.

Ce signal d’origine galactique demeure largement dominant devant le signal primordial, et il faut donc le caractériser avec une meilleure précision relative que le rapport entre les deux signaux. De nature statistique bien plus riche, sa description complète demeure complexe : le signal galactique est fondamentalement non-gaussien par opposition au signal primordial gaussien. Nous reviendrons dans ce rapport sur les implications de ces non-gaussianités pour la caractérisation du signal, retenons simplement pour l’instant qu’une description statistique d’un signal non-gaussien exige des outils statistiques plus complets que la simple détermination d’un spectre de puissance.

Mon travail de stage de Master 2 s’inscrit dans cette perspective de compréhension de la statistique régissant le milieu interstellaire galactique observé¹. Les implications ne sont pas encore d’ordre cosmologique, mais participent à l’élaboration d’un modèle statistique du milieu interstellaire. Plus précisément l’objectif est de mettre en place une méthodologie de comparaison entre observations et simulations fondée sur des outils statistiques allant au-delà de l’information contenue dans le spectre de puissance. Cette étape est en effet indispensable pour relier la description statistique du milieu interstellaire observé à une physique contrôlée dans les simulations. Ce travail s’appuie sur des cartes en densité de colonne simulées du milieu interstellaire et des observations en intensité du nuage de Polaris avec le télescope spatial Herschel à 500 μm (ce rayonnement est dû à l’émission thermique des poussières qui est au premier ordre proportionnelle à la densité de colonne du gaz).

Le milieu interstellaire a une importance d’autant plus étonnante au sein de « l’écosystème galactique » que sa proportion en masse est négligeable (moins de 1% de la masse totale de notre Galaxie). Notamment au cœur du cycle de formation stellaire, il met en jeu des processus physiques complexes et hautement non-linéaires. On le décrit comme étant un fluide auto-gravitant magnétisé et turbulent constitué de gaz et de poussières, dont la composition chimique est extrêmement riche bien que largement dominée par l’hydrogène sous ses formes ionisée, atomique, et moléculaire. Ce milieu constitue également un système ouvert, ce qui en complique sensiblement la modélisation. Il est soumis à des chocs, à des rayons cosmiques et à un champ de rayonnement aux origines multiples (extragalactique, stellaire, interstellaire).

Face à une telle complexité de phénomènes, une description déterministe des signaux observés paraît fastidieuse voire impossible. La description statistique apparaît alors comme une nécessité, et doit être suffisamment complète sans quoi elle risquerait de passer à côté d’une part importante de la complexité de la physique du milieu. L’ambition qui nourrit ce travail de stage est précisément de développer un modèle à faible nombre de paramètres qui capturerait l’essentiel de l’information statistique pertinente présente dans les observations. Et idéalement, les paramètres de ce modèle pourraient être associés aux paramètres physiques du milieu (champ magnétique B , nombre de Mach $\mathcal{M}$, etc). Nous allons avoir que

1. Dans ce rapport, toute mention du milieu interstellaire fera implicitement référence au milieu interstellaire de notre galaxie, la Voie Lactée. Si la résolution angulaire des observations du milieu galactique est évidemment bien meilleure comparée aux études extragalactiques, notons toutefois qu’étant plongés dans ce milieu nous ne bénéficions pas d’un recul permettant d’observer la globalité du milieu, et notamment de faire des statistiques sur des populations d’objets.

le choix de l'outil statistique est crucial pour l'élaboration de ce modèle, et dans le cadre de ce stage, mon travail se fonde sur un outil issu de la science des données : la *Wavelet Scattering Transform* [7]. Cet outil est développé par le groupe de Stéphane Mallat (Département d'Informatique de l'ENS, Collège de France) qui collabore avec nous sur ce travail et est demandeur de potentielles applications physiques de leurs travaux mathématiques. La maîtrise de la *Wavelet Scattering Transform* et de ses propriétés appliquées au problème de caractérisation statistique du milieu interstellaire est au cœur de mon travail.

Dans ce rapport, je détaillerai dans une première partie l'intérêt de la *Wavelet Scattering Transform* (abrégée WST dans la suite) pour la modélisation statistique du milieu interstellaire. Je décrirai ensuite mon travail sur la modélisation à faible nombre de paramètres des cartes simulées du milieu interstellaire. Dans une troisième partie, je développerai mon analyse statistique du nuage de Polaris observé avec le télescope spatial Herschel. Finalement, j'esquisserai les perspectives données par ce travail sur la comparaison des simulations avec les observations du milieu interstellaire.

Table des matières

1	Décrire statistiquement le milieu interstellaire avec la WST	6
1.1	Physique du milieu interstellaire magnétisé : non-linéarité et interactions entre échelles . . .	6
1.2	Capter l'information statistique d'une image du milieu interstellaire	7
1.3	La représentation <i>Wavelet Scattering Transform</i>	9
1.3.1	Propriétés générales de la représentation WST	9
1.3.2	Définition mathématique	10
1.3.3	Invariance par translation et visualisation des coefficients	11
1.3.4	Capture de l'intégralité de l'information statistique et synthèse	13
1.3.5	Champ d'application de la WST	13
2	Modélisation à faible nombre de paramètres des cartes simulées du milieu inter-	14
	stellaire	
2.1	Cadre de travail : simulations magnéto-hydrodynamique du milieu interstellaire	14
2.2	Symétries et invariances dans les simulations	16
2.3	Modélisation des représentations WST	18
3	Analyse statistique d'une observation : le nuage de Polaris	21
3.1	Propriétés du nuage de Polaris	21
3.2	Analyse standard du nuage : estimateur de spectre de puissance	22
3.3	Partitionnement du nuage : vers l'identification de similarités statistiques	23
3.3.1	Intérêt d'un partitionnement du nuage	23
3.3.2	Dépendance aux paramètres WST	24
3.3.3	Modélisation des coefficients WST sur la carte partitionnée de Polaris	26
4	Vers la définition d'une méthodologie de comparaison entre observations et simu-	28
	lations	
4.1	De la difficulté de simuler le milieu interstellaire	28
4.2	Définition d'une métrique dans l'espace des paramètres du modèle	28
A	Champs gaussiens et spectres de puissance	31
A.1	Définitions préliminaires	31
A.2	Champs gaussiens aléatoires	31
B	Méthodes numériques d'estimation de spectres de puissance	33
B.1	Le périodogramme et ses limites	33
B.2	La méthode de Welch	33
C	Reconstruction du potentiel de lentillage faible de l'Univers : définition d'un esti-	35
	mateur statistique	
C.1	Point de départ et notations	35
C.2	Reconstruire le potentiel de lentillage ψ	35
C.3	Reconstruire le spectre de puissance C_l^ψ	36

1 Décrire statistiquement le milieu interstellaire avec la WST

Le milieu interstellaire est régi par une physique profondément non-linéaire qui rend toute description statistique *a priori* hautement dimensionnelle, c'est-à-dire dépendant d'un grand nombre de grandeurs statistiques. L'élaboration d'un modèle statistique satisfaisant du milieu interstellaire repose sur le pari que malgré cette complexité apparente, un comportement statistique émergent pourrait apparaître dans les cartes du milieu interstellaire.

Partant d'observations du milieu interstellaire, et pour commencer de cartes traçant la densité de colonne du gaz interstellaire ou de la poussière, quelle est la description qui permette de capturer l'essentiel des propriétés statistiques de ces observations ? Le point de départ de mon travail repose sur un outil statistique issu de la science des données et destiné à trouver le meilleur compromis entre réduction de dimensionnalité et capture de l'information statistique pertinente pour des signaux variés (images, sons, etc). Cet outil, appelé *Wavelet Scattering Transform* est particulièrement adapté à la caractérisation statistique de textures, ou de manière plus générale, adapté à la caractérisation de processus statistiques stationnaires². Nous allons voir qu'en un certain sens, les images du milieu interstellaire dont on dispose peuvent être vues comme des réalisations de processus stationnaires et bénéficier d'une description statistique efficace reposant sur la WST.

1.1 Physique du milieu interstellaire magnétisé : non-linéarité et interactions entre échelles

Le milieu interstellaire désigne le milieu galactique dans lequel baignent et évoluent les étoiles. C'est un système ouvert et aux propriétés physiques variées qui joue un rôle essentiel pour le cycle de formation stellaire. Les étoiles naissent de l'effondrement gravitationnel du gaz interstellaire, et peuvent fortement interagir avec leur environnement au début de leur vie par l'intermédiaire de jets. Elles meurent le plus souvent de façon spectaculaire en *supernovæ*, dont les restes alimentent de nouveau le milieu interstellaire en matière et en énergie. A la fois cimetière et berceau des étoiles, ce milieu éclaire donc sur l'histoire de la Galaxie et joue un rôle majeur dans ce que l'on peut appeler « l'écosystème galactique », malgré sa relative discrétion dans le bilan de masse de notre galaxie (le milieu interstellaire représente environ 0,5% de la masse totale de la Galaxie).

Le milieu interstellaire est constitué de gaz et de poussières, composés à majorité d'hydrogène sous ses formes ionisée, atomique, et moléculaire, mais aussi d'éléments lourds jouant un rôle fondamental pour la chimie du milieu. Les grains interstellaires (carbonés et silicatés) sont notamment le lieu de formation de la molécule H_2 . Soumis à un rayonnement environnant important d'origine stellaire et extragalactique, ainsi qu'à des rayons cosmiques accélérés notamment par les restes de *supernovæ*, le milieu interstellaire redistribue cette énergie incidente et s'organise *via* un mécanisme d'instabilité thermique en phases de densités et de température variées (n_H allant de 10^{-3} cm^{-3} à 10^4 cm^{-3} et la température d'environ 50 K à 10^6 K).

En terme de structure, le milieu interstellaire apparaît fragmenté et auto-similaire sur au moins quatre ordres de grandeur (typiquement de 10^{-1} à 10^3 pc), allant des échelles galactiques aux échelles protostellaires [16]. On y trouve aussi des filaments de matière qui connectent des cœurs denses en effondrement gravitationnel. De plus, les largeurs de raies observées ne peuvent s'expliquer uniquement par des dispersions de vitesse d'origine thermique et cette composante supra-thermique varie avec l'échelle : ceci suggère que le milieu est turbulent. On observe alors des lois d'échelle dans les fonctions de structure du champ de vitesse ou encore dans la distribution du nombre de fragments selon les échelles, indiquant une géométrie fractale du milieu interstellaire.

La physique du milieu interstellaire est décrite par les équations de la magnéto-hydrodynamique (MHD), ce milieu s'apparente alors à un fluide magnétisé turbulent auto-gravitant et perturbé par un rayonnement externe, ou encore des rayons cosmiques. Rappelons les équations fondamentales de la MHD idéale couplant le champ de vitesse $\mathbf{v}$, le champ de densité ρ , le champ de pression p , le champ magnétique $\mathbf{B}$ et la densité de courant $\mathbf{j}$ pour un fluide totalement ionisé :

2. Appliquée à des images, la stationnarité est de nature spatiale, et un « processus statistique stationnaire » fera référence à un champ aléatoire bidimensionnel homogène, *i.e.* dont les propriétés statistiques sont invariantes par translation.

Equation de continuité (bilan de masse) : $\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0$

Bilan de la quantité de mouvement : $\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} = -\frac{1}{\rho} \nabla \left(p + \frac{B^2}{2\mu_0} \right) + \mathbf{g} + \mu \Delta \mathbf{v} + \frac{1}{\mu_0} (\mathbf{B} \cdot \nabla) \mathbf{B}$

Equation d'induction : $\frac{\partial \mathbf{B}}{\partial t} = \nabla \wedge (\mathbf{v} \wedge \mathbf{B})$

Ici μ désigne la viscosité cinématique et $\mathbf{g}$ le champ de gravitation. De plus, à ce système se rajoutent notamment une équation d'état et une équation énergétique prenant en compte les multiples processus de chauffage et de refroidissement du milieu interstellaire.

Cette physique est hautement non-linéaire, le couplage entre les échelles est donc important. Mathématiquement, ce couplage entre échelles provient directement des termes non-linéaires. Partant par exemple de l'équation de bilan de la quantité de mouvement en l'absence de champ magnétique et de viscosité nous avons en coordonnées cartésiennes sur l'axe (Ox) :

$$\frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} + v_z \frac{\partial v_x}{\partial z} = -\frac{1}{\rho} \frac{\partial p}{\partial x}$$

Dans l'espace de Fourier, en décomposant nos champs en ondes planes, cette équation conduit à :

$$-\omega \tilde{v}_x(\mathbf{k}, \omega) + \int \tilde{v}_x(\mathbf{k}', \omega) \mathbf{k}' \cdot \tilde{\mathbf{v}}(\mathbf{k} - \mathbf{k}', \omega) d\mathbf{k}' = -\frac{k_x}{\rho} \tilde{p}(\mathbf{k}, \omega)$$

Dans ce cas précis, le couplage apparaît clairement dans l'intégrale et découle directement du terme $(\mathbf{v} \cdot \nabla) \mathbf{v}$ dans l'équation bilan de la quantité de mouvement : les échelles communiquent donc entre elles. La description statistique (pour peu qu'elle soit suffisamment complète) des grandeurs du fluide (champ de densité, champ de vitesse ou champ magnétique par exemple) doit permettre d'identifier ces interactions entre les échelles.

1.2 Capturer l'information statistique d'une image du milieu interstellaire

Comme mentionné en introduction, la capture de l'information statistique des observations du milieu interstellaire apparaît cruciale pour les problèmes de contamination des signaux cosmologiques qui sont déjà traités de manière statistique. Mais plus fondamentalement, la légitimité de l'approche statistique dans l'analyse du milieu interstellaire est aussi à relier au caractère turbulent du milieu, décrit par des moyens statistiques uniquement. Le comportement chaotique turbulent qui émerge des non-linéarités de la physique du milieu interstellaire ne laisse pas de place à une description déterministe. Concrètement, partant d'une image comme celle du nuage de Polaris³ (figure 1, observation avec Herschel à 500 μm [24]), le point de départ est de modéliser cette image comme étant la réalisation x d'un champ aléatoire X de loi inconnue. Que peut-on alors dire sur la loi de X à partir seulement d'une seule réalisation (ou éventuellement d'un petit nombre de réalisations pour des cartes simulées) ?⁴

Ce problème de capture de l'information statistique d'une image du milieu interstellaire est évidemment lié à la physique des processus en jeu dans ce milieu. On a vu que la non-linéarité des processus mène notamment à des interactions entre échelles. Toute description statistique satisfaisante des images observées ne pourra donc se passer de la prise en compte de ces interactions. On peut donc d'ores et déjà exclure une modélisation de X comme étant un champ gaussien homogène. Cette description statistique

3. La description de ce nuage et de son étude fait l'objet de la partie 3 de ce rapport.

4. Notons à ce stade que si la physique du milieu interstellaire évolue dans un espace tridimensionnel, nous n'avons accès qu'à des grandeurs intégrées sur la ligne de visée, *i.e.* projetées sur la sphère céleste. Le transfert radiatif dans le milieu interstellaire est en théorie à prendre en compte dans nos observations. Néanmoins pour la simple émission thermique des poussières dans le nuage de Polaris et aux fréquences d'observation d'Herschel, le milieu est optiquement mince, l'émission est donc simplement proportionnelle à la densité de colonne des poussières et nous n'aurons pas besoin de tenir compte de ces effets de transfert radiatif pour notre analyse. Précisons toutefois, que selon le milieu observé (moléculaire, neutre, ou ionisé), le choix du traceur influence les données observationnelles à modéliser.

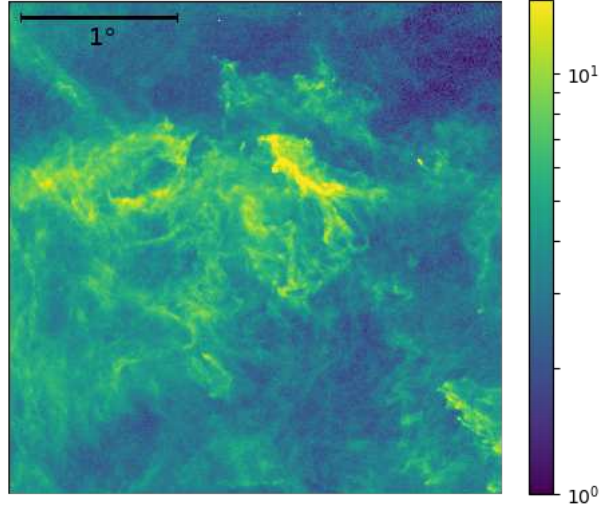


FIGURE 1 – Carte d’intensité du nuage de Polaris (MJy.sr^{-1}). Instrument : *Herschel Spire LW* ($500\ \mu\text{m}$) [22]

prise des cosmologistes pour des phénomènes (quasi-)linéaires (description du fond diffus cosmologique, ou du champ de densité primordial de l’Univers aux grandes échelles notamment), n’encode pas de corrélations entre les échelles : un champ gaussien homogène isotrope est uniquement décrit par son spectre de puissance $P(k)$ et sa valeur moyenne. On se référera à l’annexe A pour un rappel des propriétés élémentaires des champs gaussiens aléatoires homogènes.

Aller au-delà de l’information contenue dans le spectre de puissance est donc essentiel pour notre problème. Différents auteurs se sont essayés à des techniques d’analyse statistique d’images permettant de relier des estimateurs statistiques à la physique du milieu interstellaire. Mentionnons par exemple :

- Burkhart, Collins, et Lazarian ont travaillé sur les liens entre la distribution de probabilités des valeurs de densités de colonne sur des images synthétiques et l’effondrement gravitationnel de nuages moléculaires [9].
- Levrier, Falgarone et Viallefond ou encore Burkhart et Lazarian ont étudié des outils statistiques permettant de capturer les informations de phase dans l’espace de Fourier [18] et [10].
- Makarenko, Bushby *et al.* utilisent une approche topologique pour analyser des images simulées de milieux magnéto-hydrodynamiques compressibles turbulents [20].

Ces différentes approches visent à construire une collection d’estimateurs pertinents reflétant une partie de la physique du milieu interstellaire. Idéalement, les estimateurs construits devraient encoder toute l’information statistique disponible avec le moins de paramètres possible. L’objectif de mon travail s’inscrit dans la continuité de ces travaux mais avec une philosophie différente dans la mesure où la finalité est d’établir un véritable modèle statistique des images du milieu interstellaire, qui contienne donc toute l’information pertinente physiquement pour décrire ces images. Un tel modèle serait la brique de base d’une comparaison satisfaisante entre observations et simulations. Notons que ce travail a une portée bien plus générale que l’unique caractérisation statistique du milieu interstellaire, tant est général le problème de comparaison d’images lorsque leurs statistiques sont complexes.

L’information statistique à extraire des images du milieu interstellaire repose sur le choix d’un outil statistique adapté. Typiquement, une analyse fréquentielle de l’image peut être envisagée, de manière à travailler sur l’information statistique contenue dans la représentation de l’image dans l’espace de Fourier. Mais une utilisation directe de l’information contenue dans l’espace de Fourier n’est pas satisfaisante pour des raisons qui seront précisées dans la suite de ce rapport. Mon travail s’appuie en fait sur l’expérience de la science des données pour le traitement d’images, en utilisant un outil développé par le groupe du mathématicien Stéphane Mallat (Département d’Informatique de l’ENS, Collège de France), qui collabore étroitement avec nous sur ce travail. La description de cet outil, appelé *Wavelet Scattering Transform*, fait l’objet de la sous-partie suivante.

1.3 La représentation *Wavelet Scattering Transform*

On modélise donc une image du milieu interstellaire comme étant la réalisation x d'un champ aléatoire X de dimension deux. Cette réalisation est alors une fonction $x : \llbracket 1, T \rrbracket^2 \rightarrow \mathbb{R}$ où T désigne la taille de l'image en pixels, correspondant à une carte en intensité du milieu interstellaire (pour une observation) ou une carte en densité de colonne de gaz (pour une simulation). Mener une analyse statistique directement sur la fonction x (*i.e.* au niveau des pixels) est en général peu judicieux pour plusieurs raisons. Premièrement, la fonction x est une représentation hautement dimensionnelle du signal, la dimension étant égale au nombre de pixels qui est en général élevé (typiquement de l'ordre du million de pixels pour une image comme celle de Polaris en figure 1). Deuxièmement, cette représentation ne tire parti d'aucune propriété du signal et encode donc potentiellement de l'information redondante ou peu pertinente vis-à-vis de ce que l'on cherche à en tirer. Notamment si l'on suppose que X est un champ homogène, tirer parti de cette homogénéité dans la représentation de notre signal est essentiel pour manipuler une information statistique de plus basse dimension possible.

1.3.1 Propriétés générales de la représentation WST

L'objet de la *Wavelet Scattering Transform* est de transformer le signal original x en un signal de basse dimension $\Phi(x)$ adapté à la classification de textures, ou plus généralement à la classification de processus stationnaires [7]. On dira que $\Phi(x)$ est une « représentation » de x . Si le problème de caractérisation des images du milieu interstellaire n'est pas en soi un problème de classification (comme peuvent l'être les problèmes de reconnaissance de visages ou d'objets), arriver néanmoins à capturer une information statistique suffisante pour distinguer des milieux aux comportements physiques bien distincts et pourtant difficilement distinguables « à l'œil nu » est une étape initiale importante.

En science des données, la classification d'images est un problème actif de recherche qui se heurte à un obstacle de taille : la grande variabilité dans les classes d'images. Par exemple, si l'on cherche à distinguer des images de chats et des images de chiens dans un grand jeu de données en s'intéressant directement aux valeurs des pixels contenus dans ces images, l'immense variabilité de ces valeurs pour chacune des classes (chat ou chien) rend la classification ardue⁵. Pour un tel problème, trouver les caractéristiques statistiques qui distinguent un chat d'un chien semble même impossible. Néanmoins, on peut réduire simplement la variabilité des images représentant un chien ou un chat en remarquant que la nature de l'animal n'est pas altérée si l'on applique par exemple une translation ou une rotation à notre image. Plus généralement, la construction d'un opérateur Φ invariant sous l'action d'un groupe de symétries donné est un moyen simple de réduire la variabilité des images sur lesquelles on travaille. La représentation $\Phi(x)$ de nos images sera donc une représentation s'affranchissant d'une partie de l'information contenue dans x et jugée non pertinente.

Avant d'entrer dans la définition mathématique de la WST, il est important de souligner trois propriétés essentielles qui ont conduit à une telle définition de cette transformée :

- Une invariance par translation des opérateurs de représentation de l'image : le but étant de réduire la variabilité des signaux à représenter, supposés stationnaires. Il est à noter que nativement la WST ne comprend pas d'invariance par rotation ou dilatation, qui n'ont d'ailleurs pour l'instant pas d'utilité directe au problème de caractérisation statistique du milieu interstellaire.
- Une stabilité aux petites déformations de la représentation de nos images : les opérateurs permettant de définir la WST ont pour propriété d'être non expansifs⁶, ce qui conduit à un contrôle des variations de la représentation d'une image que l'on perturberait. Cette stabilité est assurée par l'emploi d'ondelettes dans les opérateurs, elle est à opposer aux instabilités intrinsèques d'une transformation standard de Fourier [21]. Pour des images du milieu interstellaire, la stabilité aux petites déformations de la représentation est essentielle pour pouvoir envisager une modélisation statistique (notamment l'évolution temporelle à court temps est une petite déformation, pourtant la physique ne change pas, les propriétés statistiques des images doivent alors restées (quasi) inchangées).

5. Se fonder sur la distance dans l'espace des valeurs des pixels entre deux images pour déterminer si l'on a affaire à un chien ou un chat est complètement sans espoir.

6. On dira qu'un opérateur $\Phi : E \rightarrow E$, avec E un espace normé, est non expansif si et seulement si $\forall (x, y) \in E^2, \|\Phi(x) - \Phi(y)\| \leq \|x - y\|$

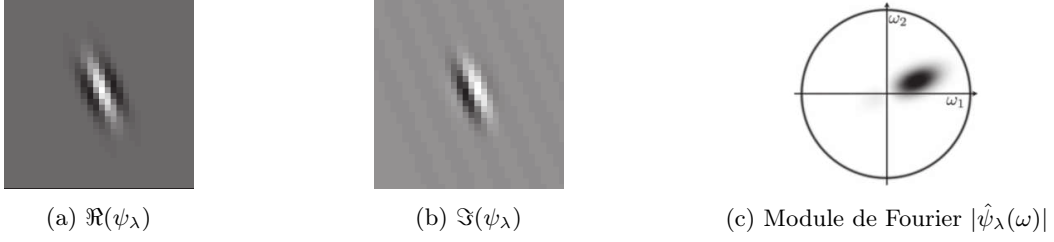


FIGURE 2 – Ondelette de Morlet $\psi(u)$ avec $\sigma = 0.85$ et $\xi = \frac{3\pi}{4}$ (images tirées de [7])

- Une prise en compte de moments statistiques d'ordre supérieur à deux : la discrimination d'images homogènes ne peut se faire généralement avec une information du type spectre de puissance uniquement. De plus, il a été déjà discuté précédemment le caractère fortement non-gaussien des images du milieu interstellaire qui par conséquent ne peuvent être décrites uniquement par une information de type spectre de puissance.

1.3.2 Définition mathématique

La WST s'appuie sur l'utilisation d'ondelettes, fonctions oscillantes à support compact. Plus précisément dans le cadre de ce travail nous utiliserons des ondelettes de Morlet définies dans le plan complexe (muni du produit scalaire usuel) par :

$$\psi(u) = \alpha(e^{iu \cdot \xi} - \beta)e^{-\frac{|u|^2}{2\sigma^2}}$$

α et β sont des paramètres ajustés de manière à normaliser ψ et à rendre la fonction de moyenne nulle ($\int \psi(u)du = 0$). La figure 2 montre une telle ondelette pour $\sigma = 0.85$ et $\xi = \frac{3\pi}{4}$. Nous considérerons des ondelettes dilatées d'un facteur 2^j et tournées d'un angle $\frac{(\theta-1)\pi}{L}$ via la rotation r_θ définies de la manière suivante :

$$\psi_{j,\theta} = 2^{-2j}\psi(2^{-j}r_\theta^{-1}u)$$

Nous retiendrons que L est un paramètre fixant le nombre d'angles à parcourir pour les ondelettes. De plus, nous parlerons abusivement par la suite d'échelle j et d'angle θ , mais gardons en tête que ce sont 2^j et $\frac{(\theta-1)\pi}{L}$ qui correspondent à l'échelle et à l'angle de l'ondelette respectivement.

Mathématiquement, la WST est une série de transformations opérant sur l'image x . Ces transformations s'organisent en couches successives suivant le schéma de construction suivant :

- La **couche 0** du réseau calcule la convolution $x * \phi_J$ où ϕ_J est une gaussienne d'écart-type 2^J qui agit comme une fenêtre de localisation dans l'image, de taille caractéristique 2^J avec $J \in \mathbb{N}^*$. Cette couche calcule donc simplement une moyenne locale de l'image.
- La **couche 1** du réseau calcule la collection de signaux $\{|x * \psi_{j_1, \theta_1}| * \phi_J\}$ avec $j_1 \in \llbracket 0, J-1 \rrbracket$ et $\theta_1 \in \llbracket 1, L \rrbracket$. Cette transformation du signal encode une information semblable à celle de la densité spectrale de puissance, c'est donc une mesure de l'énergie du signal dans une région bien spécifique du plan de Fourier identifiée par une échelle j_1 et un angle θ_1 .
- La **couche 2** du réseau calcule la collection de signaux $\{||x * \psi_{j_1, \theta_1}| * \psi_{j_2, \theta_2}| * \phi_J\}$ avec $(j_1, j_2) \in \llbracket 0, J-1 \rrbracket^2$ et vérifiant $j_2 > j_1$, et $(\theta_1, \theta_2) \in \llbracket 1, L \rrbracket^2$. Cette double convolution par des ondelettes extrait une information statistique correspondant aux interactions entre les échelles j_1 et j_2 dans des directions respectives θ_1 et θ_2 . Ces fonctions encodent donc une information statistique correspondant à des moments d'ordre supérieurs à 2 (*i.e.* au-delà du spectre de puissance).

On pourrait extraire des coefficients provenant de couches ultérieures (en convoluant avec une troisième série d'ondelettes notamment), mais on peut montrer que les trois premières couches contiennent déjà de l'ordre de 99% de l'énergie totale⁷ du signal x [7].

Cette représentation du signal x correspond donc à un ensemble de signaux répartis en couches. On dénombre :

7. L'énergie d'un signal x est à comprendre ici comme la norme au carré du signal $||x||^2$.

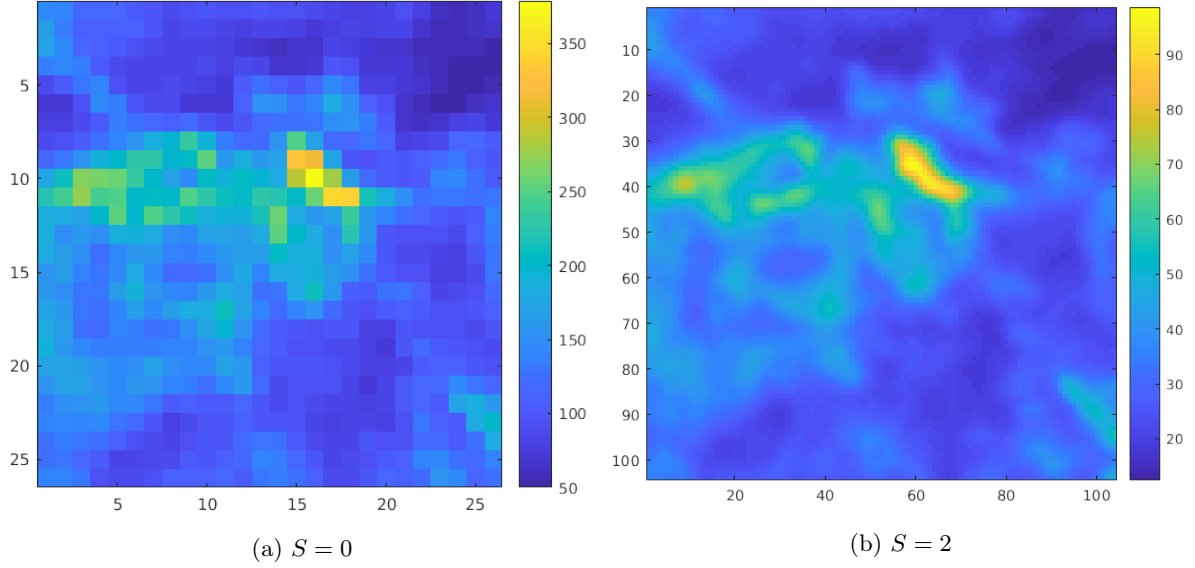


FIGURE 3 – Echantillonnage des coefficients de la couche 0 sur l'exemple de Polaris (unité en MJy.sr^{-1}), image de taille 832×832 pixels avec $J = 5$ et différentes valeurs d'*oversampling*

- pour la **couche 0** : seulement 1 fonction
- pour la **couche 1** : $J \times L$ fonctions
- pour la **couche 2** : $\frac{J(J-1)}{2} \times L^2$ fonctions

Chacun de ces signaux est une fonction de $\llbracket 1, T \rrbracket^2$ dans $\mathbb{R}$, et évidemment si l'un des objectifs de la WST est de réduire la dimension de la représentation du signal x , en multipliant le nombre de fonctions décrivant le signal nous n'avons ici qu'augmenté la dimension de la représentation. En pratique, nous ne conservons pas l'intégralité de ces fonctions, mais nous les échantillonnons sur un sous-réseau régulier de l'image naturellement conditionné par la taille de l'image T et l'échelle J de la fonction de fenêtrage ϕ_J , mais aussi par un paramètre de sur-échantillonnage S permettant une division de l'image en régions pouvant éventuellement se recouvrir. Les paramètres T , J , et S découpent donc l'image en $\left(\frac{T}{2^{J-S}}\right)^2$ régions sur lesquelles sont calculés les échantillons. La figure 3 (a) illustre l'échantillonnage de la fonction $x * \phi_J$ avec un sur-échantillonnage nul, ici la division de l'image correspond donc simplement à un pavage. Nous voyons apparaître $\left(\frac{832}{2^5}\right)^2 = 26 \times 26$ échantillons. Avec $S = 0$ on perd nécessairement de l'information dans le signal. En pratique, choisir un sur-échantillonnage $S = 2$ permet d'avoir un bon compromis en terme de quantité d'information récupérée dans les fonctions et nombre d'échantillons (figure 3 (b)).

En conclusion, la représentation WST d'une image x consiste en une collection de coefficients répartis en 3 couches et correspondant à l'échantillonnage des fonctions $x * \phi_J$, $\{|x * \psi_{j_1, \theta_1}| * \phi_J\}$ et $\{||x * \psi_{j_1, \theta_1}| * \psi_{j_2, \theta_2}| * \phi_J\}$.

1.3.3 Invariance par translation et visualisation des coefficients

Après avoir travaillé sur une image exemple x qui peut être vue comme la réalisation d'un champ aléatoire X , le même raisonnement peut être tenu sur le champ aléatoire X directement. En supposant ce champ homogène, nous nous intéressons aux espérances des fonctions calculées avec la WST, c'est à dire aux fonctions $\mathbb{E}(X * \phi_J)$ pour la couche 0, $\{\mathbb{E}(|X * \psi_{j_1, \theta_1}| * \phi_J)\}$ pour la couche 1 et $\{\mathbb{E}(|X * \psi_{j_1, \theta_1}| * \psi_{j_2, \theta_2}| * \phi_J)\}$ pour la couche 2. Du fait de l'homogénéité de X , ces fonctions sont indépendantes de la position en laquelle elles sont évaluées. Il s'ensuit que l'estimateur naturel de ces espérances à partir d'une unique réalisation x est simplement la moyenne des échantillons de chaque fonction. Ainsi pour décrire la statistique d'une image correspondant à une réalisation d'un processus stationnaire, nous nous baserons sur ces estimateurs naturels des espérances des coefficients WST du champ aléatoire correspondant.

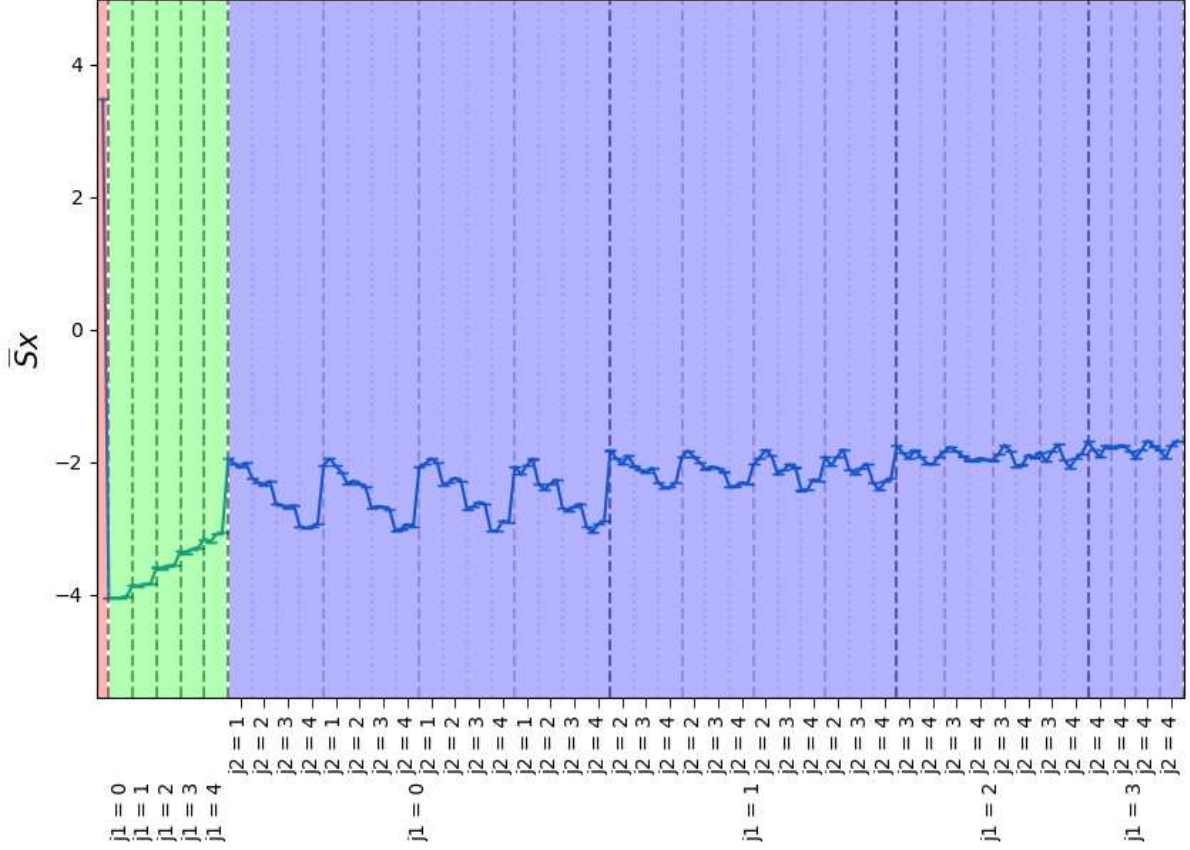


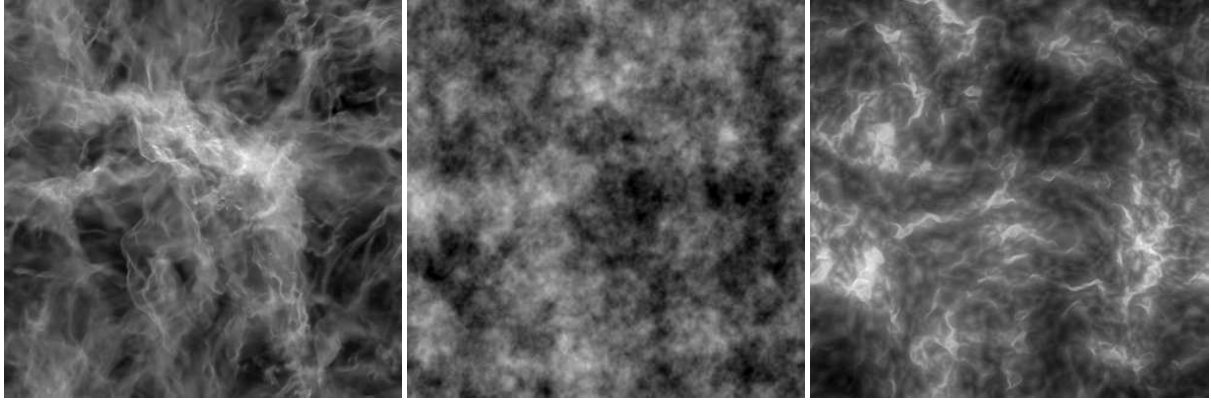
FIGURE 4 – Coefficients WST calculés pour l’image du nuage de Polaris avec $J = 5$, $L = 4$, $S = 2$. L’axe des ordonnées correspond aux valeurs de $\bar{S}x_0$, $\bar{S}x_1$ et $\bar{S}x_2$, classés par numéro de couche croissant et ordre lexicographique sur (j_1, θ_1) pour les coefficients de la couche 1 et ordre lexicographique sur $(j_1, \theta_1, j_2, \theta_2)$ pour les coefficients de la couche 2. Colorié en rouge, vert, bleu les régions contenant les coefficients des couches 0, 1, 2 respectivement.

En pratique, dans le travail rapporté dans ce document, de manière à ramener au même ordre de grandeur les coefficients WST de couches différentes, nous travaillerons systématiquement avec le logarithme népérien de ces coefficients. Nous noterons donc par la suite :

- $\bar{S}_0x$ la moyenne des échantillons de $\ln(x * \phi_J)$
- $\bar{S}_1x(j_1, \theta_1)$ les moyennes des échantillons de $\ln(|x * \psi_{j_1, \theta_1}| * \phi_J)$
- $\bar{S}_2x(j_1, \theta_1, j_2, \theta_2)$ les moyennes des échantillons de $\ln(|x * \psi_{j_1, \theta_1} * \psi_{j_2, \theta_2}| * \phi_J)$

La description statistique d’une image x correspondant au processus aléatoire X sera donc cette collection de $1 + J \times L + \frac{J(J-1)}{2} \times L^2$ coefficients.

Pour représenter ces coefficients calculés pour une image donnée, dans la suite de ce rapport nous adopterons un ordre croissant des coefficients selon leur couche d’appartenance, et un ordre lexicographique sur (j_1, θ_1) pour les coefficients de la couche 1 et sur $(j_1, \theta_1, j_2, \theta_2)$ pour les coefficients de la couche 2. La figure 4 illustre la visualisation adoptée des coefficients WST calculés sur l’exemple de l’image de Polaris. Le coefficient de la couche 0 est le coefficient dans la zone coloriée en rouge (le plus à gauche). Puis viennent dans la zone en vert les coefficients de la couche 1, et dans la zone en bleu, les coefficients de la couche 2. Les tirets verticaux les plus foncés délimitent des coefficients à j_1 fixé pour les couches 1 et 2, les tirets d’intensité intermédiaire délimitent les coefficients à θ_1 fixé pour la couche 2, puis les pointillés clairs délimitent les coefficients de la couche 2 à j_2 fixé.



(a) Image originale extraite d'une simulation (b) Image synthétisée *via* l'information du spectre de puissance (c) Image synthétisée *via* la représentation WST

FIGURE 5 – Synthèses d'une image à partir d'une information de spectre de puissance et à partir des coefficients WST (images fournies par Sixin Zhang, DIENS).

1.3.4 Capture de l'intégralité de l'information statistique et synthèse

Sans prétendre effectuer une étude quantitative précise de l'information statistique capturée dans les images du milieu interstellaire, nous pouvons avoir une appréciation visuelle du gain de la représentation WST par rapport au spectre de puissance pour décrire une image à la statistique complexe. Partant d'une représentation WST donnée, il est en effet possible de synthétiser une nouvelle image ayant la même représentation⁸. La figure 5 montre les résultats de deux synthèses effectuées d'une part à partir de l'information de spectre de puissance contenue dans l'image originale (image (b)), et d'autre part à partir de la représentation WST de l'image originale (image (c)). Cet exemple montre qu'au moins visuellement, la statistique du champ aléatoire responsable de l'image originale semble être capturée de manière plus satisfaisante avec la représentation WST qu'avec le spectre de puissance : on voit notamment apparaître une structure filamenteuse sur l'image (c) qui est totalement inexistante sur l'image (b). Le simple fait de pouvoir reconstituer une image visuellement semblable à l'image originale à partir de sa représentation WST indique que l'information statistique contenue dans les coefficients WST est mieux adaptée à ces images que le seul spectre de puissance.

1.3.5 Champ d'application de la WST

La WST est une transformée adaptée à la caractérisation statistique de processus stationnaires. Un des premiers aspects de mon travail a été d'étudier l'intérêt de cette transformée pour le problème de reconstruction du champ de déflexion de la lumière du fond diffus cosmologique (abrégé CMB), qui peut être vu comme un processus stationnaire. Brièvement (ce travail n'a pas de lien avec la suite de ce rapport), le problème de reconstruction du champ de déflexion du CMB est un problème consistant à extraire le champ de déformation d'un signal observé qui est la déformation d'un signal primordial connu et modélisé comme un signal gaussien. L'étude de l'état de l'art du domaine, dont les grandes idées sont résumées dans l'annexe C, a permis de montrer que la WST n'a rien à apporter de nouveau pour ce type de problème. Ce problème peut en effet être traité de manière analytique *via* un développement perturbatif du signal primordial gaussien. L'identification du signal cible, c'est-à-dire le champ de déflexion, à une échelle donnée, découle alors directement d'un unique estimateur rendu sans biais et de variance minimale.

On prendra donc garde que pour de tels problèmes bénéficiant d'une modélisation simple (perturbation d'un signal gaussien connu *a priori*), et bien décrits sur le plan analytique, l'utilisation d'outils statistiques sophistiqués ne présentera pas nécessairement d'intérêt.

⁸. L'algorithme de synthèse consiste à effectuer une descente de gradient à partir d'un bruit blanc, permettant de converger vers une image ayant les coefficients WST recherchés.

2 Modélisation à faible nombre de paramètres des cartes simulées du milieu interstellaire

Nous cherchons un modèle statistique permettant de caractériser de manière satisfaisante les images du milieu interstellaire, et dont les paramètres seraient reliés aux grandeurs physiques mises en jeu. Si nous disposions d'un tel modèle, nous pourrions l'appliquer à des situations physiques de complexité variable. Il est donc judicieux de commencer par chercher un tel modèle sur un cas pour lequel on contrôle la physique, c'est-à-dire sur des simulations numériques d'écoulements interstellaires turbulents (magnétisés ou non). C'est la première étape vers une modélisation statistique incluant toute la physique non-linéaire régissant la formation des structures interstellaires.

Dans cette partie, je vais détailler mon travail de modélisation sur des cartes en densité de colonne extraites de simulations magnéto-hydrodynamiques avec auto-gravité dans des conditions typiques du milieu interstellaire. L'étude des propriétés statistiques de ces cartes simulées repose sur leur représentation WST. Je vais montrer comment la prise en compte de propriétés simples de symétries et d'invariance dans les simulations permet de développer un modèle statistique de ces cartes à faible nombre de paramètres (en l'occurrence 13 paramètres pour modéliser les coefficients de la couche 2 qui se comptent typiquement en plusieurs centaines).

2.1 Cadre de travail : simulations magnéto-hydrodynamique du milieu interstellaire

Les cartes simulées sont issues de simulations magnéto-hydrodynamiques du milieu interstellaire. Elles ont été fournies par Cédric Colling et Patrick Hennebelle (AIM). Ces simulations résolvent les équations de la magnéto-hydrodynamique idéale avec auto-gravité en régime compressible, sans viscosité et isotherme pour un gaz d'hydrogène partant des conditions initiales suivantes :

- un champ de densité uniforme initialement fixé à 2 cm^{-3}
- un champ magnétique uniforme dont l'intensité B_0 est un paramètre libre de la simulation
- un champ de vitesse initial nul
- un champ de température uniforme fixé à 10 K

La résolution des équations se fait sur une grille cubique de 256^3 cellules. Le côté du cube de simulation correspond à 40 pc et les conditions aux limites sont périodiques : le gaz sortant par une face du cube rentre à nouveau dans la boîte par la face opposée.

On ajoute manuellement une force dans les équations qui va produire de la turbulence. Ce forçage turbulent permet notamment d'injecter une énergie qui permet de garder la dispersion des vitesses turbulentes $v_{\text{RMS}}^{\text{turb}}$ constante au cours du temps. Il compense également une perte d'énergie d'origine numérique et liée au schéma de résolution des équations appelée « dissipation numérique ».

Après un régime transitoire d'une durée de l'ordre du temps d'effondrement $t_{\text{ff}} \sim 1/\sqrt{\rho_H G} \sim 67 \text{ Myr}$, ces simulations entrent dans un régime stationnaire : les propriétés statistiques des grandeurs n'évoluent plus à partir d'un certain temps d'intégration. On se place alors dans ce régime.

Partant de ces cubes de simulations, les cartes sont extraites en intégrant le champ de densité sur une direction perpendiculaire au champ magnétique initial. Pour une simulation donnée, nous extrayons 20 cartes calculées à des temps d'intégration différents (mais toujours en régime stationnaire) et séparés d'un temps pris typiquement de l'ordre du temps de retournement à l'échelle de la taille du cube de simulation ($t_d \sim 40 \text{ pc}/v_{\text{RMS}}^{\text{turb}} \sim 3 - 30 \text{ Myr}$). Nous obtenons donc des cartes simulées en densité de colonne de gaz, que nous considérons statistiquement indépendantes.

Les cartes dont on dispose sont alors regroupées par paquets de 20 en 9 classes correspondant à 9 jeux de conditions initiales différentes. Les classes se distinguent entre elles par l'intensité du champ magnétique initial B_0 et par la valeur de la dispersion des vitesses turbulentes. La table 1 résume les valeurs des paramètres en fonction de la classe considérée.

Nous représentons sur la figure 6 des réalisations de cartes pour chacune des classes considérées. Ces cartes sont notamment très filamenteuses. Les filaments apparaissent de plus en plus marqués et de

Classe	1	2	3	4	5	6	7	8	9
B_0 (μG)	0	0	0	0.5	0.5	0.5	1	1	1
Dispersion des vitesses turbulentes $v_{\text{RMS}}^{\text{turb}}$ (km.s^{-1})	1	4	9	1	4	9	1	4	9

TABLE 1 – Valeurs des paramètres libres pour chaque classe de simulation

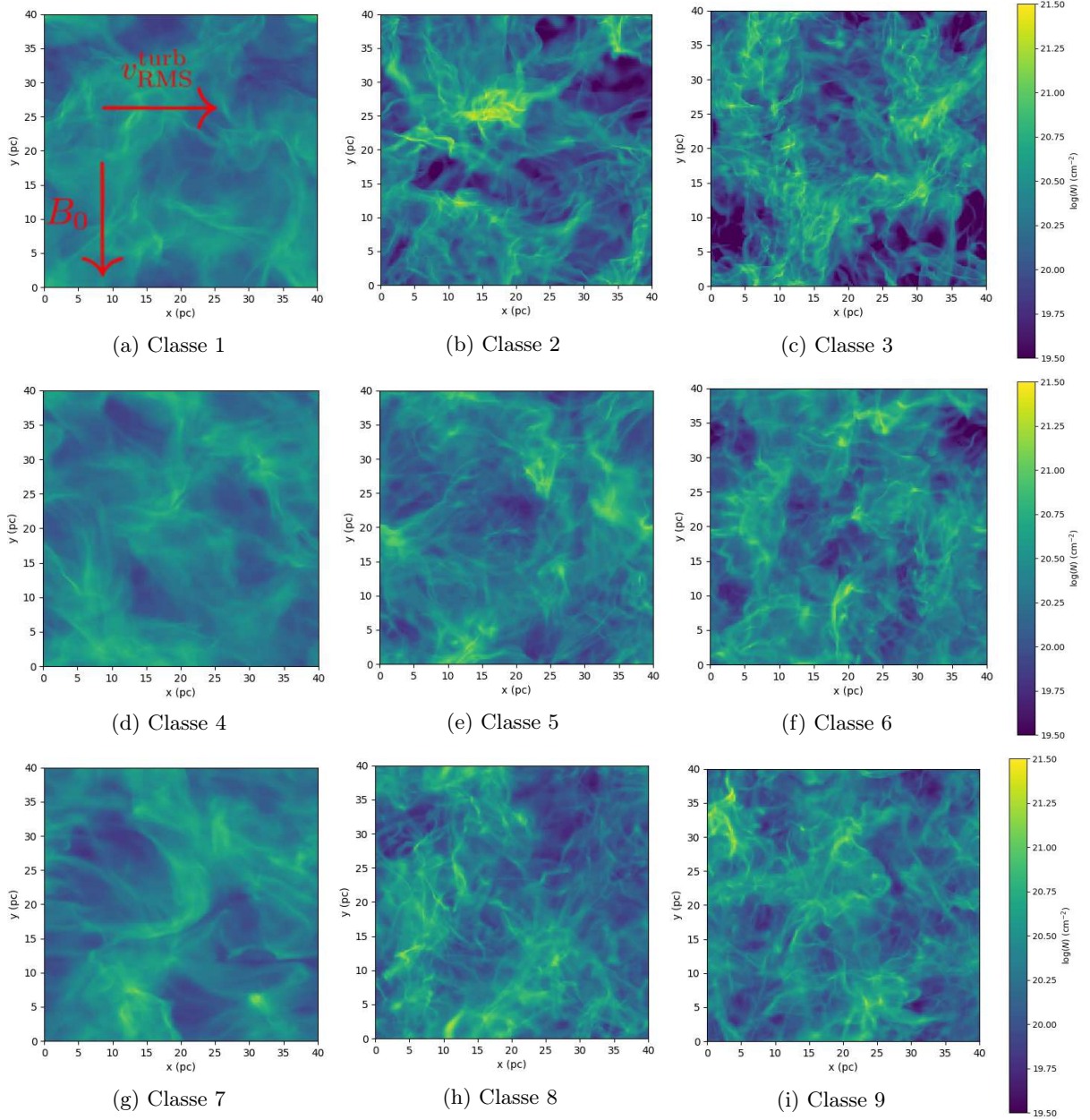
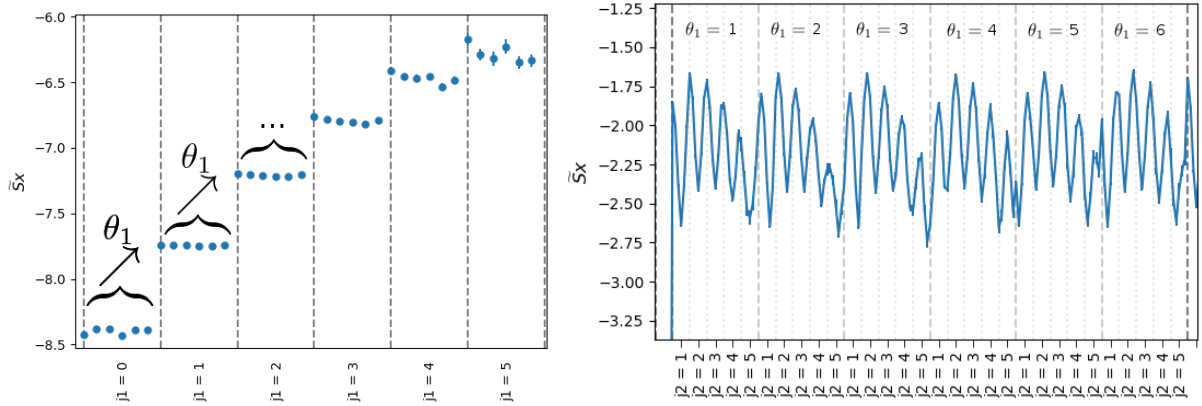


FIGURE 6 – Exemples de réalisations de cartes simulées en densité de colonne de H (en cm^{-2}) pour chacune des classes de paramètres. Les flèches rouges indiquent le sens de croissance des paramètres libres selon les classes.



(a) Coefficients de la couche 1. Pour une échelle j_1 fixée, on observe une invariance des coefficients selon θ_1 et j_2 fixés (délimités par les tirets verticaux clairs) se répètent lorsque l'on prend un autre θ_1 (et avec le même j_2) avec cependant un décalage qui augmente avec θ_1 . Ce motif est une fonction de la variable $\theta_2 - \theta_1$, il apparaît même périodique en $\theta_2 - \theta_1$ de période L

FIGURE 7 – Propriétés d'isotropie dans la représentation WST des cartes simulées de la classe 1 (représentation calculée avec $J = 6$, $L = 6$ et $S = 2$)

moins en moins diffus à mesure que le forçage turbulent augmente. L'influence de la valeur de champ magnétique initiale B_0 sur ces cartes est déjà moins évident visuellement à forçage turbulent fixé. On entrevoit déjà les difficultés qui apparaissent pour discriminer les comportements physiques simplement à partir d'observations à l'œil nu de ce type de cartes en densité de colonne. L'idée de mener une analyse statistique satisfaisante de ces images permettrait idéalement de proposer une méthodologie systématique pour identifier des comportements physiques différents.

Pour une classe donnée de cartes simulées, nous disposons de 20 images $\{x_i\}_{1 \leq i \leq 20}$ correspondant à des réalisations distinctes d'un même processus aléatoire. Nous calculons pour chacune de ces cartes les coefficients moyens $\bar{S}_0x$, $\bar{S}_1x$ et $\bar{S}_2x$ comme expliqué dans la partie précédente, à une nuance près que nous choisissons de normaliser nos coefficients de manière à augmenter leur invariance. Plus précisément nous calculons pour une carte x_i donnée :

- pour la **couche 0**, nous calculons toujours la même quantité, l'estimateur naturel de $\mathbb{E}(\ln(X * \phi_J))$ noté $\bar{S}_0x$ en moyennant les échantillons de la fonction $\ln(x_i * \phi_J)$
- pour la **couche 1**, nous calculons désormais les estimateurs naturels des fonctions $\mathbb{E}\left(\ln\left(\frac{|X * \psi_{j_1, \theta_1}| * \phi_J}{X * \phi_J}\right)\right)$ que l'on notera $\tilde{S}_1x$ en moyennant les échantillons des signaux $\ln\left(\frac{|x_i * \psi_{j_1, \theta_1}| * \phi_J}{x_i * \phi_J}\right)$. Cette normalisation permet de rendre ces coefficients invariants à un éventuel facteur multiplicatif global du signal [8].
- pour la **couche 2**, nous calculons désormais les estimateurs naturels des fonctions $\mathbb{E}\left(\ln\left(\frac{|X * \psi_{j_1, \theta_1} * \psi_{j_2, \theta_2}| * \phi_J}{|X * \psi_{j_1, \theta_1}| * \phi_J}\right)\right)$ que l'on notera $\tilde{S}_2x$ en moyennant les échantillons des signaux $\ln\left(\frac{|x_i * \psi_{j_1, \theta_1} * \psi_{j_2, \theta_2}| * \phi_J}{|x_i * \psi_{j_1, \theta_1}| * \phi_J}\right)$. Cette normalisation permet de mettre en évidence les sursauts d'activité à une échelle j_1 et un angle θ_1 fixés [8].

2.2 Symétries et invariances dans les simulations

Les simulations contiennent des propriétés de symétries et d'invariance qu'il s'agit d'exploiter pour simplifier la description statistique des images projetées. On a vu dans la sous-partie précédente que leur homogénéité conduisait à une moyenne des échantillons pour un signal de la représentation WST donné. Mais ce n'est pas la seule invariance exploitable.

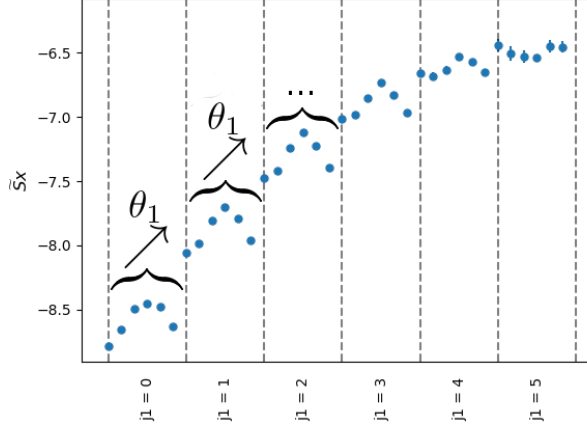


FIGURE 8 – Coefficients de la couche 1 pour les cartes simulées de la classe 7 présentant un champ magnétique initial $B_0 = 1 \mu\text{G}$ (calculés avec $J = 6$, $L = 6$ et $S = 2$). Une direction privilégiée apparaît pour $\theta_1 = 4$ aux plus petites échelles. Aux grandes échelles (*i.e.* à grand j_1) l'anisotropie s'atténue.

Notamment, pour une simulation sans champ magnétique initial (classes 1, 2, 3), nous sommes dans le cas d'une simulation hydrodynamique standard qui ne privilégie aucune direction. Statistiquement ces simulations sont alors isotropes. On doit pouvoir retrouver cette isotropie statistique dans la représentation WST. La figure 7 montre l'apparition de propriétés d'isotropie sur les coefficients WST calculés pour la classe 1 des cartes simulées. Sur les coefficients de la couche 1, qui mesurent une information équivalente à la densité spectrale de puissance, on observe une invariance selon θ_1 des coefficients (à j_1 fixé), au moins aux plus petites échelles. Aux grandes échelles, cette invariance semble se dégrader légèrement, mais remarquons aussi que la variance des coefficients augmente à ces grandes échelles qui se rapprochent de la taille des cartes utilisées (les cartes sont de taille $T = 2^8$ pixels). Cette dégradation des propriétés de symétries aux grandes échelles ne nous préoccupera pas dans notre analyse. On comprend donc que l'isotropie statistique des simulations implique que les coefficients de la couche 1 ne doivent pas dépendre de la variable θ_1 . Concernant les coefficients de la couche 2, l'apparition de l'isotropie statistique est plus subtile dans la mesure où ces coefficients sont calculés par l'intermédiaire d'une double convolution du signal avec des ondelettes orientées respectivement par θ_1 et θ_2 . L'isotropie ici va se traduire par une dépendance angulaire en $\theta_2 - \theta_1$ des coefficients de la couche 2. En effet, l'isotropie statistique implique que les coefficients de la couche 2 ne peuvent dépendre d'un angle de référence θ_1 , seuls les écarts entre les variables angulaires θ_1 et θ_2 peuvent influencer sur la valeur des coefficients.

Pour des simulations présentant un champ magnétique initial non nul (classes 4 à 7), on s'attend au contraire à une brisure de l'isotropie statistique. Le champ magnétique initial impose en effet une direction privilégiée dans les simulations qui influe sur les coefficients WST. La figure 8 montre l'effet de cette anisotropie sur les coefficients de la couche 1 pour les cartes simulées de la classe 7 ($B_0 = 1\mu\text{G}$ et $v_{\text{RMS}} = 1\text{km.s}^{-1}$). Il apparaît clairement une direction privilégiée dans les coefficients : pour $\theta_1 = 4$ les motifs à j_1 fixé présentent un maximum. Cette direction privilégiée correspond à une rotation des ondelettes de $\frac{(4-1)\pi}{L} = \frac{\pi}{2}$ (ici $L = 6$). Cette valeur est liée à la direction du champ magnétique initial. L'anisotropie liée à la physique des simulations apparaît donc clairement dans la représentation WST des cartes extraites de ces simulations.

D'autre part, la présence de symétries miroirs dans les simulations influencera aussi les dépendances angulaires des coefficients. Notamment, dans les simulations isotropes on s'attend à ce que le changement de variable $\theta_1 \rightarrow -\theta_1$ soit sans effet sur la valeur des coefficients de la couche 1. Pour ces coefficients, la bonne variable angulaire sera donc $|\theta_1|$. De la même manière pour les coefficients de la couche 2, le changement de variable $\theta_2 - \theta_1 \rightarrow -(\theta_2 - \theta_1)$ doit être sans effet : l'orientation des écarts angulaires doit être sans importance (propriété d'achiralité). La variable angulaire adaptée pour décrire ces coefficients est donc $|\theta_2 - \theta_1|$. Pour des simulations présentant une anisotropie, il faudra cependant être plus prudent sur les conclusions, dans un cas favorable où une unique direction privilégiée apparaît, on peut s'attendre à des symétries miroirs par rapport à cette unique direction. Le raisonnement se complique dans des situations présentant plusieurs directions privilégiées...

En conclusion, les symétries et invariances présentes dans les simulations ont une influence sur les représentations WST des images extraites de ces simulations. Des motifs émergent de ces représentations, et des raisonnements assez simples sur les propriétés d'isotropie, d'anisotropie, ou de symétries miroirs des images amènent à considérer de nouvelles variables angulaires plus pertinentes pour décrire ces coefficients. Il apparaît donc assez clairement dans ces représentations WST « brutes » de la redondance et des propriétés d'invariance qu'il s'agit d'exploiter pour réduire de manière plus significative la dimensionnalité de la représentation statistique de ces cartes simulées.

2.3 Modélisation des représentations WST

Nous décrivons dans cette partie la modélisation développée à partir des coefficients de la couche 2 des représentations WST de nos classes de cartes simulées. Même si la modélisation des coefficients de la couche 1 reste importante, nous nous concentrons ici sur ceux de la couche 2 pour deux raisons :

1. Premièrement, le nombre de coefficients issus de la couche 2 est en pratique bien supérieur à ceux des couches 0 et 1 ($\frac{J(J-1)}{2} \times L^2$ comparé à 1 et $J \times L$ respectivement), il y a donc simplement beaucoup plus d'enjeux à travailler sur la modélisation à faible nombre de paramètres des coefficients de la couche 2.
2. Deuxièmement, les coefficients de la couche 2 constituent le réel gain de la WST par rapport aux méthodes d'analyse statistique reposant sur le spectre de puissance. On espère donc que leur étude permettra de mettre en évidence quantitativement les propriétés statistiques des observations du milieu interstellaire et de faire le lien entre ces quantités et la physique du milieu.

La modélisation des coefficients de la couche 2 pour ces cartes simulées consiste à trouver une loi analytique reliant les coefficients aux variables descriptives $(j_1, \theta_1, j_2, \theta_2)$. Est-on capable de trouver une formule à petit nombre de paramètres $\tilde{S}_2^{\text{fit}} x(j_1, \theta_1, j_2, \theta_2)$ ajustant les coefficients calculés pour chacune des classes de cartes simulées ?

Pour développer ce modèle, nous allons chercher à identifier différentes contributions additives à $\tilde{S}_2^{\text{fit}} x$ sous la forme suivante :

$$\tilde{S}_2^{\text{fit}} x(j_1, \theta_1, j_2, \theta_2) = \tilde{S}_2^{\text{cst}} x(j_1, j_1 - j_2) + \tilde{S}_2^{\text{iso}} x(j_2, |\theta_2 - \theta_1|) + \tilde{S}_2^{\text{aniso } 1} x(j_2, \theta_1) + \tilde{S}_2^{\text{aniso } 2} x(j_2, \theta_2)$$

Cette décomposition sous forme de contributions additives est justifiée grossièrement par l'utilisation de logarithmes de coefficients WST : on s'attend à avoir décomposé d'éventuels produits en sommes. Et nous allons voir que cette décomposition donne des résultats satisfaisants et nous conforte dans cette approche. Les différents termes permettent de distinguer simplement les contributions ne dépendant pas de la variable angulaire ($\tilde{S}_2^{\text{cst}} x$), isotrope ($\tilde{S}_2^{\text{iso}} x$), ou anisotropes ($\tilde{S}_2^{\text{aniso } 1} x$ et $\tilde{S}_2^{\text{aniso } 2} x$). On reste donc guidé par les éventuelles propriétés d'invariance et de symétries évoquées dans la sous-partie précédente.

On modélise la composante ne dépendant pas des variables angulaires $\tilde{S}_2^{\text{cst}} x$ comme une fonction des variables j_1 et $j_1 - j_2$. La variable $j_1 - j_2$ est préférée à j_2 pour mettre en évidence d'éventuels comportements invariants d'échelle suggérés par l'étude analytique de processus intermittents avec la WST développée dans [8]. On propose un modèle correspondant à un polynôme à deux variables de degré total 2 :

$$\tilde{S}_2^{\text{cst}} x(j_1, j_1 - j_2) = a_0 + a_1 j_1 + a_2 j_1^2 + a_3 (j_1 - j_2) + a_4 (j_1 - j_2)^2 + a_5 j_1 (j_1 - j_2)$$

Les $\{a_i\}$ sont les paramètres du modèle à ajuster. Ce modèle quadratique est l'extension naturelle d'un modèle linéaire en les variables j_1 et $j_1 - j_2$ jugé insuffisant. Qualitativement, cette fonction polynomiale permet d'approcher de manière satisfaisante les tendances ne dépendant pas des variables angulaires, et ce indépendamment des classes de cartes choisies (mais bien sûr les paramètres à ajuster vont dépendre de la classe).

La définition d'un modèle pour le terme isotrope $\tilde{S}_2^{\text{iso}} x$ repose sur les considérations de symétries abordées précédemment. Pour décrire un comportement isotrope, la bonne variable angulaire est $|\theta_2 - \theta_1|$, et nous avons noté de plus une périodicité des coefficients en $\theta_2 - \theta_1$ de période L sur l'exemple des cartes

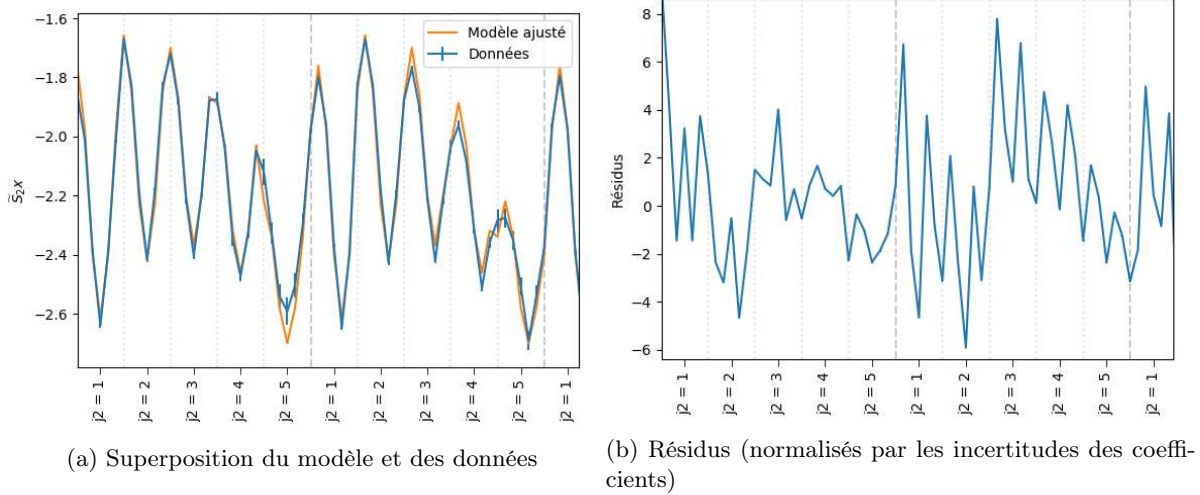


FIGURE 9 – Ajustement de $\tilde{S}_2 x$ par $\tilde{S}_2^{\text{cst}} x + \tilde{S}_2^{\text{iso}} x$ sur une partie des coefficients pour $j_1 = 0$ (calculés pour $J = 6$, $L = 6$, $S = 2$).

de la classe 1 (cf figure 7). En étudiant les cartes des classes isotropes (1, 2, 3) nous nous rendons compte que cette périodicité est très générale. Ceci nous amène à modéliser le terme isotrope de la façon suivante :

$$\tilde{S}_2^{\text{iso}} x(j_2, |\theta_2 - \theta_1|) = (a_6 j_2 + a_7) \cos\left(\frac{2\pi}{L} (\theta_2 - \theta_1)\right)$$

Nous prenons en compte également dans ce terme une modulation du cosinus de loi linéaire en j_2 pour modéliser au plus simple une dépendance en j_2 apparente dans les résidus. On choisit de ne pas faire intervenir la variable j_1 ou $j_1 - j_2$ dans ce terme isotrope, qui semble être moins importante qualitativement.

Ces deux termes $\tilde{S}_2^{\text{cst}} x$ et $\tilde{S}_2^{\text{iso}} x$ permettent de modéliser correctement les coefficients de la couche 2 des classes isotropes avec des valeurs de χ_r^2 ⁹ inférieures à 10. La figure 9 illustre l'adéquation du modèle avec les données. Les résidus restent de l'ordre de l'unité, on ne voit pas apparaître de tendance supplémentaire évidente à prendre en compte. Remarquons que sur ces exemples, il y a 540 coefficients de couche 2, à comparer aux 8 paramètres intervenant dans notre modèle. La réduction de la dimension de la représentation statistique de nos cartes est flagrante sur ces exemples et confirme la pertinence de l'approche.

L'ajustement du modèle aux données repose sur un algorithme de minimisation des moindres carrés (algorithme *Trust Region Reflective* [5]). On fournit des bornes et des valeurs initiales estimées à l'œil sur les données pour les paramètres ajustés de manière à favoriser la stabilité des ajustements : nous voulons que les valeurs ajustées soient comparables entre les différentes classes de cartes simulées. En effet, de manière générale, un problème de minimisation dans un espace de paramètres donné peut présenter plusieurs minima locaux, et il faut donc s'arranger pour que l'algorithme converge vers un minimum cohérent avec les données à ajuster. Notamment si l'on perturbe faiblement les données à ajuster, nous ne voulons pas que l'algorithme converge vers un minimum complètement différent. Notons aussi que les minima locaux vers lesquels l'algorithme peut converger ne sont pas nécessairement globaux, la recherche du minimum global (à supposer qu'il existe !) doit se faire avec des algorithmes plus sophistiqués (méthodes de Monte-Carlo par chaînes de Markov par exemple [6]).

La modélisation des termes d'anisotropie est fondée sur l'observation des résidus des cartes des classes les plus anisotropes (4 et 7 principalement, l'augmentation du forçage turbulent à champ magnétique initial fixé a tendance à gommer les anisotropies). On modélise les termes d'anisotropie par les fonctions

9. On rappelle la définition du χ^2 . Pour un jeu de N données $\{y_i\}_{i \in \llbracket 1, N \rrbracket}$ d'incertitudes $\{\sigma_i\}_{i \in \llbracket 1, N \rrbracket}$ et modélisées par $\{\tilde{y}_i\}_{i \in \llbracket 1, N \rrbracket}$, on définit le χ^2 par $\chi^2 = \sum_{i=1}^N (y_i - \tilde{y}_i)^2 / \sigma_i^2$. Le χ^2 réduit, noté χ_r^2 est alors défini par $\chi_r^2 = \chi^2 / (N - p)$ où p désigne le nombre de paramètres introduits dans le modèle.

Classe	1	2	3	4	5	6	7	8	9
a_0	-2.711	-2.555	-2.504	-2.837	-2.647	-2.622	-2.860	-2.631	-2.611
a_1	0.225	0.169	0.112	0.293	0.204	0.200	0.306	0.211	0.192
a_2	-0.033	-0.024	-0.011	-0.041	-0.030	-0.028	-0.043	-0.033	-0.027
a_3	-0.420	-0.379	-0.343	-0.512	-0.416	-0.415	-0.536	-0.421	-0.405
a_4	-0.080	-0.089	-0.092	-0.103	-0.095	-0.098	-0.106	-0.099	-0.098
a_5	0.104	0.123	0.116	0.140	0.126	0.130	0.144	0.133	0.130
a_6	-0.060	-0.076	-0.084	-0.066	-0.067	-0.070	-0.059	-0.077	-0.071
a_7	0.570	0.557	0.531	0.516	0.509	0.522	0.491	0.520	0.509
a_8	0.007	0.004	0.000	0.007	0.006	0.000	0.018	0.006	-0.001
a_9	-0.005	-0.010	0.001	0.029	-0.005	-0.004	0.053	0.002	0.001
a_{10}	-0.937	0.665	2.538	1.251	0.730	2.210	1.098	1.232	-0.916
a_{11}	0.000	0.005	0.000	0.022	-0.003	-0.002	0.049	-0.003	-0.002
a_{12}	-0.002	-0.016	-0.003	-0.109	0.019	0.022	-0.232	0.007	0.002
χ_r^2	7.01	9.36	12.74	10.37	10.70	11.35	16.85	10.76	10.82

TABLE 2 – Valeurs ajustées des paramètres du modèle et χ_r^2 pour les différentes classes de cartes simulées. Les 6 premiers paramètres correspondent aux paramètres de la fonction $\tilde{S}_2^{\text{cst}}x$. Les paramètres a_6 et a_7 concernent le terme isotrope $\tilde{S}_2^{\text{iso}}x$. Les paramètres restants décrivent l’anisotropie des simulations (termes $\tilde{S}_2^{\text{aniso } 1}x$ et $\tilde{S}_2^{\text{aniso } 2}x$).

suivantes :

$$\begin{aligned}\tilde{S}_2^{\text{aniso } 1}x(j_2, \theta_1) &= (a_8 j_2 + a_9) \cos\left(\frac{2\pi}{L}(\theta_1 - a_{10})\right) \\ \tilde{S}_2^{\text{aniso } 2}x(j_2, \theta_2) &= (a_{11} j_2 + a_{12}) \cos\left(\frac{2\pi}{L}(\theta_2 - a_{10})\right)\end{aligned}$$

On n’intègre pas de dépendance par rapport à la variable j_1 qui est négligeable par rapport à la dépendance en j_2 . On introduit une unique direction d’anisotropie paramétrée par a_{10} et commune aux deux termes d’anisotropie. La dépendance angulaire en θ_1 et θ_2 est rendue périodique de période L , et les cosinus sont modulés par des lois linéaires en j_2 . Ces fonctions modélisent de manière satisfaisante et relativement simple les anisotropies qui apparaissent dans les classes 4 et 7.

Finalement l’ajustement du modèle complet $\tilde{S}_2^{\text{fit}}x$ fournit une description à 13 paramètres de l’information de couche 2 des cartes simulées. On représente sur la table 2 les valeurs ajustées des paramètres pour les différentes classes de cartes simulées. Les ajustements sont évalués par les valeurs de χ_r^2 associées : selon les classes χ_r^2 est compris entre 7 et 17. Ces valeurs montrent que la qualité de l’ajustement est raisonnable bien qu’imparfaite. Les imperfections sont à relativiser lorsque l’on regarde la réduction de dimensionnalité de la description de l’information de couche 2 : nous sommes passés de 540 coefficients à 13 coefficients décrivant cette information pour les cartes simulées !

L’anisotropie des classes apparaît assez clairement dans ces valeurs ajustées. Le degré d’anisotropie en θ_1 des classes est quantifié par les paramètres a_8 et a_9 : en particulier, pour les classes 4 et 7, qui sont visuellement les plus anisotropes (*cf* figure 6), le paramètre a_9 est le plus important (il vaut 0.029 pour la classe 4 et 0.053 pour la classe 7, à comparer aux valeurs des autres classes qui sont plus faibles d’au moins un ordre de grandeur). Le paramètre a_{10} permet alors d’identifier une direction privilégiée dans les simulations et reliée à la direction du champ magnétique de la simulation. Le paramètre a_{12} est lui aussi un traceur de l’anisotropie en θ_2 , les valeurs les plus importantes en valeur absolue sont toujours celles des classes 4 et 7. Ces paramètres sont donc directement des indicateurs des propriétés d’anisotropie des cartes simulées.

3 Analyse statistique d’une observation : le nuage de Polaris

La finalité de ce travail d’analyse statistique d’images du milieu interstellaire est de développer un modèle statistique décrivant les propriétés statistiques des observations du milieu interstellaire magnétisé dans toute leur complexité. Ces observations peuvent sonder différents traceurs (poussières, hydrogène neutre, molécule CO, etc) et une grande gamme de longueurs d’onde allant des rayons gamma aux fréquences radio. Pour mener à bien une comparaison entre simulations magnéto-hydrodynamiques du milieu interstellaire et observations, nous devons nous intéresser à des observations adaptées aux simulations tant du point de vue du traceur choisi que des conditions physiques du milieu observé.

Pour ce stage, j’ai travaillé avec une observation de l’émission thermique des poussières au sein du nuage moléculaire du Polaris Flare, fournie par le télescope spatial Herschel qui a observé le ciel en infrarouge lointain. Après une présentation des propriétés du nuage de Polaris, je décris une analyse statistique standard du nuage à travers l’étude du spectre de puissance de l’émission thermique de la poussière. Je développe ensuite une analyse reposant sur la représentation WST de l’image et mettant en œuvre un partitionnement de l’image en régions aux comportements statistiques distincts. A partir de ce partitionnement, j’étudie la généralité du modèle développé pour les cartes simulées en l’appliquant à Polaris.

3.1 Propriétés du nuage de Polaris

Nous choisissons de nous intéresser aux observations du nuage de Polaris, un nuage moléculaire diffus dans lequel il n’y a pas de formation stellaire. Ce nuage a été abondamment observé : les structures de matière ont été étudiées à travers l’émission thermique de la poussière [22], le champ de vitesse du gaz moléculaire a été étudié en observant les raies rotationnelles de la molécule CO ([11] et [14]), et le champ magnétique a été sondé grâce à des données de polarisation stellaire optique [23] et de polarisation des poussières [27]. Sa distance n’est pas connue avec précision, elle est estimée entre 350 pc et 500 pc [29]. D’autre part, du fait de l’absence de formation stellaire, la dynamique du gaz et celle de la poussière sont dominées par des processus turbulents magnétiques. Nous sommes donc dans des conditions physiques relativement comparables aux simulations étudiées dans la partie précédente. Nous discuterons plus en détail ces similarités avec les simulations dans les processus considérés dans la partie suivante.

Notre travail sur les observations reposera uniquement sur les mesures de l’émission thermique de la poussière du nuage fournies par l’instrument SPIRE du télescope Herschel [24]. L’instrument SPIRE (*Spectral and Photometric Imaging REceiver*, [12]) est constitué d’une matrice de bolomètres permettant des observations photométriques centrées sur les longueurs d’onde 250 μm , 350 μm , et 500 μm . Nous menons dans la suite une étude monochromatique et nous concentrons alors sur l’observation du nuage à 500 μm uniquement qui trace les plus grosses poussières. Nous utilisons des données pré-traitées décrites dans [22], dont nous sélectionnons une région carrée de 832×832 pixels représentée sur la figure 10.

A 500 μm , soit $\nu = 600$ GHz, en supposant que la poussière est la seule source de rayonnement S_ν modélisée comme étant celle d’un corps noir de température T_d supposée uniforme dans le nuage, l’équation de transfert se résout simplement et permet d’écrire l’intensité spécifique $I_\nu(\tau)$ observée comme valant :

$$I_\nu(\tau) = \int_0^\tau S_\nu e^{-(\tau-\tau')} d\tau'$$

τ désigne l’épaisseur optique. A cette longueur d’onde, l’épaisseur optique vaut $\tau = \sigma_\nu N_H$ où σ_ν désigne la section efficace d’absorption de la poussière par atome d’hydrogène et N_H la densité de colonne d’hydrogène. [26] et [27] permettent d’estimer une valeur typique de la section efficace d’absorption de la poussière : $\sigma_{600 \text{ GHz}} \approx \sigma_{353 \text{ GHz}} \left(\frac{600}{353}\right)^{1.6} \approx 2.8 \times 10^{-26} \text{ cm}^2$. N_H est majoré par $5 \times 10^{21} \text{ cm}^{-2}$ [27], il vient alors $\tau \approx 1.4 \times 10^{-4} \ll 1$. Ceci montre que le milieu est optiquement mince à 500 μm et nous permet d’écrire que :

$$I_\nu(\tau) \approx S_\nu \tau \propto N_H$$

L’observation de l’émission thermique de la poussière est donc directement proportionnelle à la densité de colonne d’hydrogène N_H . Ceci légitimera notamment une comparaison avec les cartes simulées décrites dans la partie précédente.

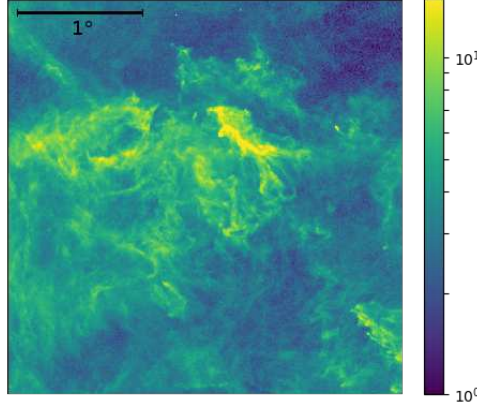
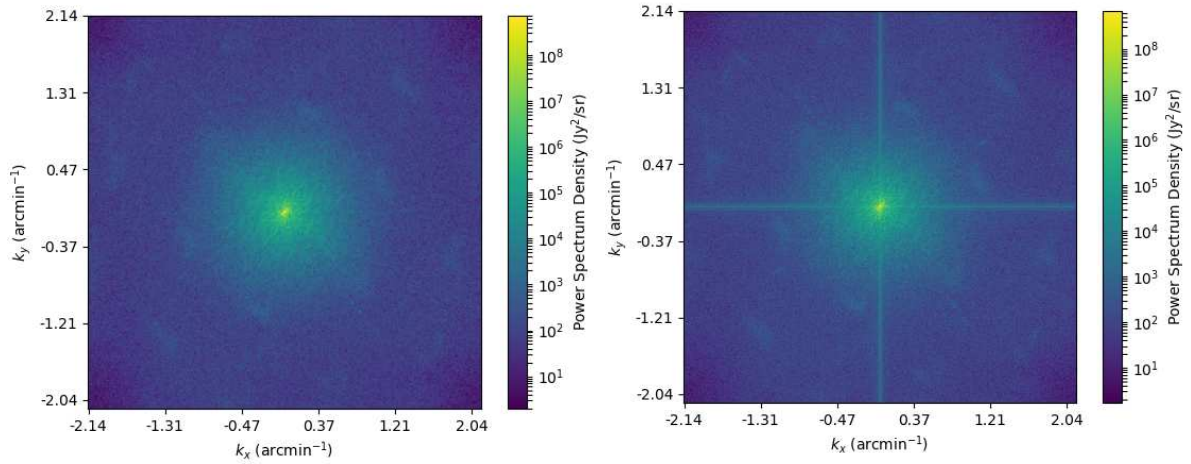


FIGURE 10 – Carte de l’émission thermique de la poussière dans le nuage de Polaris (MJy.sr^{-1}). Instrument : *Herschel Spire LW* ($500 \mu\text{m}$). [22]. Un pixel correspond à 14.00 secondes d’arc dans le ciel, le lobe instrumental fait lui 35 secondes d’arc.



(a) Calcul avec un fenêtrage de Hanning. Les effets de bord sont atténués. (b) Calcul sans fenêtrage. Une croix de diffraction apparaît.

FIGURE 11 – PSD 2D du nuage de Polaris obtenu avec la méthode Welch ($M = 512$, *oversampling* $S = 2$). Le fenêtrage permet d’atténuer les effets de bord inhérents à tout signal de taille finie.

3.2 Analyse standard du nuage : estimateur de spectre de puissance

Comme le milieu est optiquement mince à $500 \mu\text{m}$, l’analyse du spectre de puissance de l’émission thermique de la poussière sonde directement le spectre de puissance de la densité volumique de poussière en trois dimensions [22]. Dans [22], Miville-Deschênes, Martin *et al.* montrent que le spectre de puissance suit une loi de puissance avec un exposant valant -2.7 ± 0.1 sur des échelles comprises entre $k = 0.025 \text{ arcmin}^{-1}$ et $k = 0.86 \text{ arcmin}^{-1}$. Ceci leur suggère que sur ces échelles, la cascade turbulente est responsable de la structure du milieu.

Nous reproduisons ici une partie de ces résultats à partir de l’observation de Polaris à $500 \mu\text{m}$. Nous calculons la densité spectrale de puissance (abrégée PSD par la suite) associée à l’image représentée figure 10 dans un but purement pédagogique, l’objectif étant de se familiariser avec les problématiques de l’analyse spectrale. Nous choisissons d’utiliser la méthode de Welch pour estimer la PSD de l’image de Polaris, que nous implémentons en langage Python. Cette méthode diffère de la méthode d’estimation de la PSD par périodogramme, car elle permet notamment de réduire la variance de l’estimateur en faisant un compromis sur la résolution de la PSD estimée. L’annexe B décrit les méthodes numériques d’estimation de PSD mentionnées ici. On prend comme paramètre (conformément aux notations de

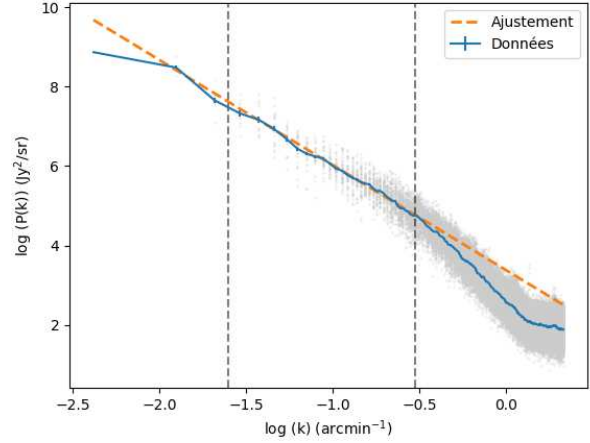
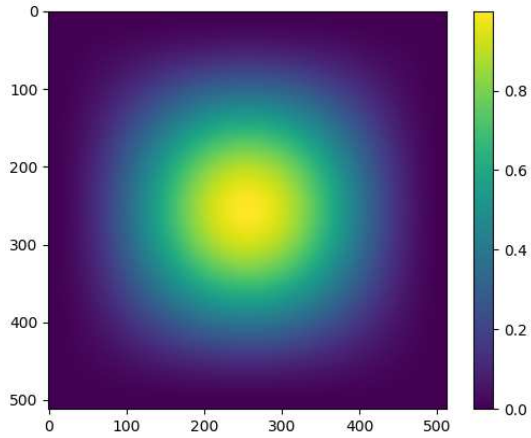


FIGURE 12 – Fenêtre de Hanning de 512×512 pixels. FIGURE 13 – PSD isotrope du nuage de Polaris. L’ajustement en loi de puissance est fait sur la plage de valeurs délimitée par les pointillés, entre $k = 0.0025 \text{ arcmin}^{-1}$ et $k = 0.3 \text{ arcmin}^{-1}$.

l’annexe), $M = 512$ et un *oversampling* de 2. La méthode de Welch repose notamment sur le choix d’une fonction de fenêtrage appelée « fenêtrage d’apodisation » qui permet de s’affranchir des effets de bord mis en évidence sur la figure 11. Sans fenêtrage, une croix de diffraction d’origine purement numérique et liée à la taille finie de l’image apparaît, il s’agit d’un biais de l’estimateur. Nous choisissons une fenêtrage de Hanning pour atténuer ce biais (représentée sur la figure 12).

A partir de la PSD bidimensionnelle associée au signal de l’image, nous calculons la PSD isotrope en effectuant un groupement des valeurs 2D par intervalles réguliers en nombre d’onde k . Pour chacun des groupes (« bins ») nous calculons la moyenne et les incertitudes associées sur les logarithmes décimaux des valeurs. La figure 13 montre en bleu les moyennes et leurs incertitudes¹⁰. Nous représentons en gris le nuage de points pour chaque bin. Nous procédons ensuite à un ajustement linéaire de ces valeurs moyennes de logarithmes de PSD $\log(P(k))$ en fonction des logarithmes de nombres d’onde $\log(k)$. Une minimisation des moindres carrés donne sur la plage de valeurs de k comprises entre $0.0025 \text{ arcmin}^{-1}$ et 0.3 arcmin^{-1} la pente -2.64 ± 0.02 avec un χ_r^2 de 3.2. Cet ajustement est satisfaisant et permet de retrouver la pente calculée dans [22].

Notons que nous nous sommes placés sur une gamme de nombres d’onde plus réduite dans les petites échelles que l’analyse menée dans [22], ceci car nous ne prenons pas en compte dans notre analyse la fonction de transfert de l’instrument qui modifie les valeurs de PSD de manière importante aux petites échelles.

3.3 Partitionnement du nuage : vers l’identification de similarités statistiques

3.3.1 Intérêt d’un partitionnement du nuage

La modélisation de l’information contenue dans le spectre de puissance ne suffit pas à décrire l’ensemble des propriétés statistiques des images du milieu interstellaire (*cf* figure 5). Nous exploitons de nouveau la WST sur l’image de Polaris dans le but de développer un modèle statistique plus complet. L’analyse menée dans la partie précédente sur les cartes simulées présupposait une homogénéité statistique des cartes. On peut se demander si cette hypothèse est valable pour une observation d’un nuage tel que Polaris. Sur la figure 10, on distingue notamment des régions qui semblent nettement plus filamenteuses que d’autres. La variété des structures qui apparaissent dans cette image nous laisse penser

10. Les incertitudes sont les incertitudes associées aux moyennes, elles valent donc $\frac{\sigma_i}{\sqrt{n_i}}$, où σ_i est l’écart-type des valeurs dans le bin i et n_i le nombre de valeurs dans le bin.

qu'il pourrait y avoir des régions aux propriétés physiques distinctes qu'il serait malvenu de confondre dans une modélisation statistique globale.

En supposant que l'hypothèse d'homogénéité statistique dans l'image n'est plus valable, comment alors identifier des morceaux de l'image présentant des propriétés statistiques homogènes dans une région et distinctes d'une région à l'autre ? En d'autres termes, si l'image de Polaris ne peut être modélisée comme étant la réalisation x d'un unique champ aléatoire homogène X , comment identifier des régions partitionnant l'image et pouvant être modélisées comme des réalisations $\{x_i\}_{i \in \llbracket 1, k \rrbracket}$ de k champs aléatoires homogènes $\{X_i\}_{i \in \llbracket 1, k \rrbracket}$ distincts ? Pour répondre à ce problème nous allons nous appuyer sur l'échantillonnage local de la représentation WST de l'image de Polaris. Comme expliqué dans la première partie de ce rapport, le calcul de la représentation WST d'une image est local et repose donc sur un découpage de l'image dépendant de la taille de l'image T , de l'échelle J du filtre gaussien ϕ_J et du paramètre d'*oversampling* S . Nous allons exploiter ce découpage de l'image (illustré sur la figure 3) pour partitionner le nuage de Polaris en régions aux propriétés statistiques distinctes.

Pour un jeu de paramètres T , J et S fixé, l'image x est découpée en $N = \left(\frac{T}{2^{J-S}}\right)^2$ cases de taille 2^J et pouvant éventuellement se recouvrir (selon la valeur de S). Pour une case indexée par $i \in \llbracket 1, N \rrbracket$ et centrée sur la position u_i dans l'image, nous disposons alors de $d = 1 + J \times L + J(J-1)/2 \times L^2$ coefficients WST que nous normaliserons de la même façon que pour les simulations, il s'agit des coefficients :

Pour la **couche 0** : $\ln(x * \phi_J)(u_i)$

Pour la **couche 1** : $\left\{ \ln\left(\frac{|x * \psi_{j_1, \theta_1}| * \phi_J}{x * \phi_J}\right)(u_i), j_1 \in \llbracket 0, J-1 \rrbracket, \theta_1 \in \llbracket 1, L \rrbracket \right\}$

Pour la **couche 2** : $\left\{ \ln\left(\frac{||x * \psi_{j_1, \theta_1}| * \psi_{j_2, \theta_2}| * \phi_J}{|x_i * \psi_{j_1, \theta_1}| * \phi_J}\right)(u_i), (j_1, j_2) \in \llbracket 0, J-1 \rrbracket^2 \mid j_2 > j_1, (\theta_1, \theta_2) \in \llbracket 1, L \rrbracket^2 \right\}$

Ainsi, à une position u_i dans l'image correspond un vecteur noté $\mathbf{y}_i$ (correspondant aux d coefficients WST) dans un espace $\mathbb{R}^d$ de dimension d . Le problème de partitionnement du nuage de Polaris revient à regrouper en k paquets (aussi appelés *clusters*) les vecteurs étant proches les uns des autres dans cet espace, et au sens d'une distance que l'on choisira euclidienne.

Pour résoudre ce problème de partitionnement de données en k *clusters* dans un espace de dimension d , nous utilisons un algorithme standard appelé « algorithme des k -moyennes » et déjà implémenté dans la bibliothèque Python *scikit-learn*¹¹. Cet algorithme est utilisé pour faire de l'apprentissage non supervisé sur des données. Considérant un ensemble de N vecteurs $\{\mathbf{y}_i\}_{i \in \llbracket 1, N \rrbracket}$ d'un espace $\mathbb{R}^d$ muni de la norme euclidienne, le principe de cet algorithme est de partitionner $\llbracket 1, N \rrbracket$ en k sous-ensembles notés $S_1, \dots, S_k$, ce afin de minimiser la somme des variances dans chaque *cluster* $\{y_i\}_{i \in S_j}$. Formellement il s'agit de trouver un partitionnement $\{S_1, \dots, S_k\}$ qui minimise :

$$\sum_{j=1}^k \sum_{i \in S_j} ||\mathbf{y}_i - \mu_j||^2$$

avec μ_j le barycentre des points $\{\mathbf{y}_i\}_{i \in S_j}$ (ie $\mu_j = \frac{1}{\text{Card}(S_j)} \sum_{i \in S_j} \mathbf{y}_i$, aussi appelé *centroïde*).

Cet algorithme nécessite donc de donner *a priori* le nombre k de *clusters* à identifier. Il suppose aussi implicitement une convexité des *clusters* et une isotropie des données dans l'espace $\mathbb{R}^d$, dès lors un partitionnement *via* l'algorithme des k -moyennes identifiera mal les structures allongées ou concaves.

3.3.2 Dépendance aux paramètres WST

Nous appliquons l'algorithme des k -moyennes aux représentations WST locales du nuage de Polaris comme expliqué ci-dessus. Il apparaît raisonnable qualitativement de partitionner Polaris en 4 *clusters*, ce choix sera néanmoins discuté en détail ci-après. Nous fixons donc $k = 4$ et nous intéressons à la dépendance du partitionnement aux paramètres WST J et L et aux coefficients WST pris en compte pour chaque case de l'image (couche 0, 1, 2).

11. <http://scikit-learn.org/stable/index.html>

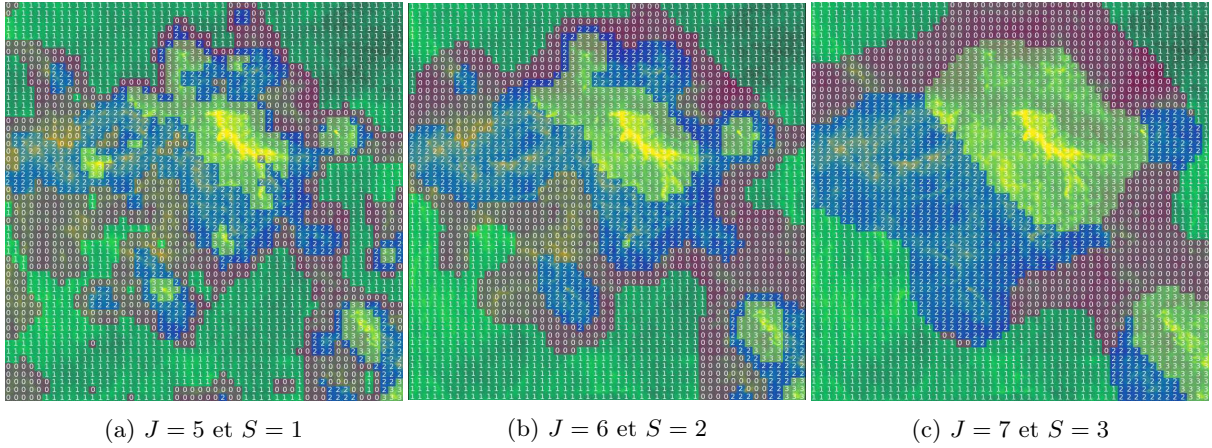
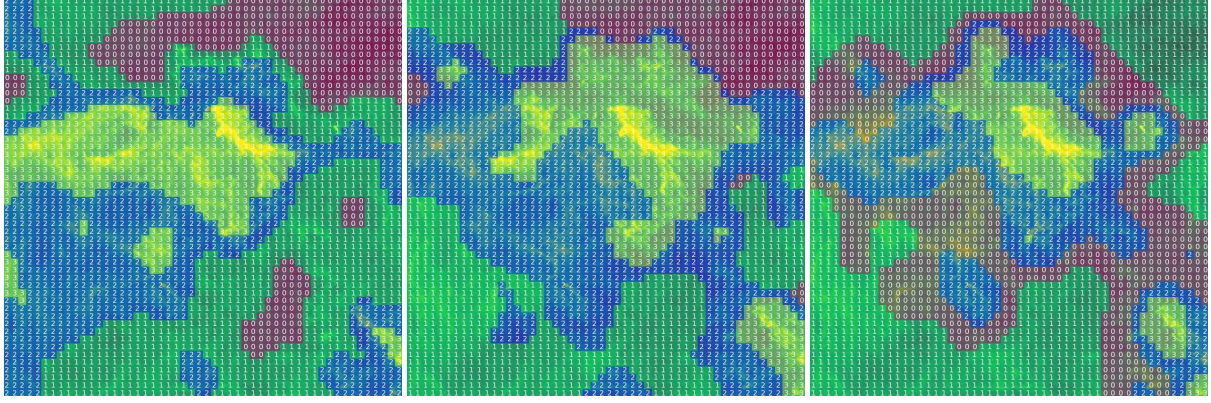


FIGURE 14 – Évolution du partitionnement du nuage de Polaris en fonction de la gamme d'échelles considérée. Les coefficients WST sont calculés avec $L = 6$ et selon différentes valeurs de J et de S de manière à garder le nombre de cases découpant l'image constant. Une case sur l'image fait 16×16 pixels et est à comparer à la taille caractéristique des filtres gaussiens ϕ_J valant 2^J (allant donc de 32 à 128 pixels sur ces images). Le chiffre visible sur chaque case correspond simplement à l'index du *cluster* auquel elle appartient.

La figure 14 montre le résultat du partitionnement de Polaris en 4 *clusters* en considérant des gammes d'échelles différentes J pour le calcul des coefficients WST. Les *clusters* sont représentés sur l'image de Polaris *via* un masque transparent de couleur rouge, vert, bleu ou jaune. Nous adaptons le paramètre d'*oversampling* S pour chaque J considéré de manière à garder le nombre de cases découpant l'image constant pour chaque partitionnement. Remarquons toutefois que les cases représentées sur ces images sont simplement des repères des positions u_i en lesquelles sont évaluées les signaux WST, leur taille n'a pas de lien avec les échelles intervenant dans les calculs des signaux (qui peuvent d'ailleurs dépasser largement la taille de la case valant dans ce cas 16 pixels). Sur ces images, les *clusters* rassemblent majoritairement des cases contiguës (il y a très peu de cases isolées du reste du *cluster* auquel elles appartiennent), ceci correspond à ce qu'on attend d'un bon partitionnement de l'image : les cases aux propriétés statistiques similaires ont plus de chance d'être voisines car soumises à la même physique. Il apparaît qu'en augmentant la gamme d'échelles à considérer pour calculer la représentation WST de l'image, le nombre de composantes connexes par *cluster* diminue. Ceci n'est pas étonnant dans la mesure où les coefficients WST encodant une information à grande échelle seront proches pour des cases séparées par une distance suffisamment inférieure à la taille de l'échelle sondée. Nous choisissons de garder les paramètres $J = 6$ et $S = 2$ pour la suite de cette analyse, ils correspondent en effet à un compromis qualitativement correct en regard de ces remarques : il n'y a pas de case isolée et la gamme d'échelles prises en compte (jusqu'à $2^{J-1} = 32$ pixels pour les ondelettes) reste comparable à la taille apparente des cases (16 pixels).

Une analyse similaire de l'évolution du partitionnement pour différentes valeurs de L montre que le partitionnement n'est pas sensible au nombre d'angles considéré pour les ondelettes, du moins à partir d'une valeur suffisamment grande pour encoder de l'information potentiellement anisotrope (typiquement à partir de $L = 4$).

Nous pouvons également choisir d'effectuer le partitionnement sur une partie seulement des coefficients WST. La figure 18 montre le résultat de différents partitionnements en ne tenant compte que des coefficients de la couche 0 (en (a)), de la couche 1 (en (b)) et de la couche 2 (en (c)). Le partitionnement à partir des coefficients de la couche 0 correspond à un partitionnement sur les moyennes locales des pixels : nous regroupons simplement les pixels les plus brillants entre eux et les pixels moins brillants entre eux. L'intérêt physique de ce type de partitionnement est limité au regroupement de régions ayant des densités de colonne comparables. Le partitionnement à partir des coefficients de la couche 1 correspond qualitativement au regroupement de régions ayant des spectres de puissance comparables, tandis que le partitionnement à partir des coefficients de la couche 2 regroupe des régions ayant des interactions entre échelles comparables. On note que le partitionnement fait avec les coefficients de la couche 2 uniquement



(a) Partitionnement sur les coefficients de la couche 0 (b) Partitionnement sur les coefficients de la couche 1 (c) Partitionnement sur les coefficients de la couche 2

FIGURE 15 – Évolution du partitionnement du nuage de Polaris en fonction de la couche de coefficients WST considérée pour faire ce partitionnement. Les coefficients WST sont calculés avec $J = 6$, $L = 6$ et $S = 2$. Une case sur l'image fait 16×16 pixels et est à comparer à la taille caractéristique des filtres gaussiens ϕ_J valant 2^J .

est identique à celui fait pour l'intégralité des coefficients (figure 14 (b)). Ce sont donc les coefficients de la couche 2 qui sont discriminants pour ce type de partitionnement. Remarquons toutefois que les coefficients de la couche 2 sont généralement en bien plus grand nombre que les coefficients des couche 0 et 1 combinés selon les paramètres J et L choisis. Leur importance pour le partitionnement est donc à rapprocher de leur « poids dimensionnel » dans l'espace $\mathbb{R}^d$, même si ce n'est pas nécessairement le seul facteur.

3.3.3 Modélisation des coefficients WST sur la carte partitionnée de Polaris

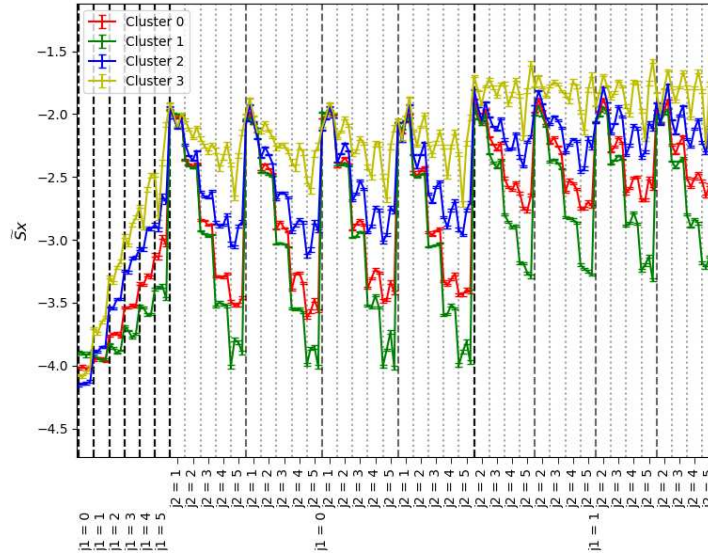
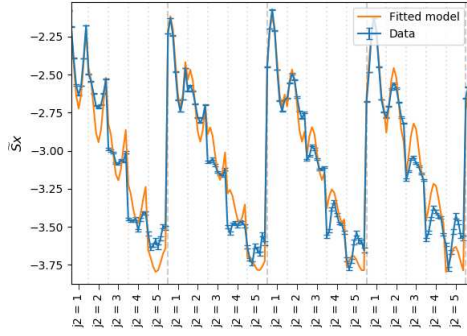


FIGURE 16 – Tracé d'une partie des coefficients WST pour les 4 *clusters* identifiés sur Polaris (coefficients calculés avec $J = 6$ et $S = 2$, la valeur de L n'étant pas discriminante, nous choisissons $L = 4$ par souci de clarté de la représentation). On montre seulement les coefficients de la couche 1 et ceux de la couche 2 pour $j_1 = 0$ et $j_1 = 1$. Les *clusters* se distinguent nettement dans ces courbes.

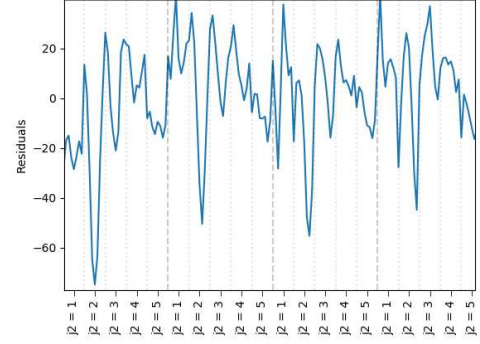
Cluster	0	1	2	3
a_0	-2.192	-2.241	-2.359	-2.457
a_1	-0.076	-0.037	0.086	0.221
a_2	0.008	0.001	-0.023	-0.044
a_3	0.213	0.126	0.066	-0.091
a_4	-0.020	-0.054	-0.019	-0.025
a_5	-0.126	-0.053	-0.073	-0.039
a_6	-0.054	-0.053	-0.066	-0.052
a_7	0.355	0.342	0.430	0.394

Cluster	0	1	2	3
a_8	-0.012	0.008	-0.013	-0.008
a_9	0.018	-0.010	0.022	0.026
a_{10}	2.588	-0.621	1.931	2.813
a_{11}	-0.009	-0.006	-0.024	-0.044
a_{12}	0.007	0.018	0.022	0.044
χ_r^2	236.2	267.1	134.4	55.4

TABLE 3 – Valeurs ajustées des paramètres du modèle $\tilde{S}_2^{\text{fit}}x$ et χ_r^2 pour les 4 *clusters* identifiés sur le nuage de Polaris comme représenté sur la figure 14 (b) (avec $J = 6$, $L = 8$, $S = 2$)



(a) Superposition du modèle et des données



(b) Résidus (normalisés par les incertitudes des coefficients)

FIGURE 17 – Ajustement de $\tilde{S}_2x$ pour le *cluster* 0 (en rouge sur la figure 14 (b)) avec le modèle $\tilde{S}_2^{\text{fit}}x$ défini en partie 2 sur une partie des coefficients pour $j_1 = 0$ (calculés pour $J = 6$, $L = 8$, $S = 2$). L'adéquation du modèle avec les données est bien moins bonne que pour les cartes simulées, on a $\chi_r^2 = 236.2 \gg 1$.

L'image étant partitionnée en 4 *clusters* aux propriétés statistiques distinctes, nous étudions désormais la pertinence du modèle développé dans la deuxième partie de ce rapport pour modéliser les propriétés statistiques de ces 4 *clusters*. Si un tel modèle, développé pour des cartes simulées, est généralisable aux images d'observations, nous aurons fait un grand pas vers la comparaison statistique d'images simulées et d'images observées.

La figure 16 trace les valeurs d'une partie des coefficients WST pour chacun des *clusters* identifiés sur l'image de Polaris. Remarquons déjà que la séparation des coefficients est nette, et fait apparaître des caractéristiques statistiques bien distinctes. En s'intéressant aux coefficients de la couche 1, on remarque que la croissance des coefficients avec l'échelle est bien plus abrupte pour le *cluster* 3 que pour le *cluster* 1. Ceci traduit que la puissance spectrale est beaucoup plus importante aux grandes échelles pour le *cluster* 3 que pour le *cluster* 1. De même, les variations sur les coefficients de la couche 2 montrent que grandes et petites échelles sont faiblement corrélées pour les *clusters* 0 et 1 (chute des coefficients avec la croissance $j_2 - j_1$), à opposer à des interactions plus marquées entre grandes et petites échelles pour les *clusters* 2 et 3 (chute moins abrupte des coefficients pour des grands $j_2 - j_2$). Ces propriétés suggèrent que la physique diffère entre chaque *cluster* et corrobore la légitimité d'un partitionnement.

Nous procédons à un ajustement des coefficients WST calculés pour chaque *cluster* en utilisant le modèle $\tilde{S}_2^{\text{fit}}x$ défini dans la partie précédente. Les valeurs des paramètres ajustés sont reportées dans la table 3. Les valeurs de χ_r^2 sont comprises entre 55 et 268 et montrent alors que ce modèle n'est pas généralisable de façon satisfaisante à ce type d'observations. La figure 17 illustre les difficultés du modèle à ajuster les coefficients du *cluster* 0. Les résidus normalisés par les incertitudes des coefficients atteignent des valeurs de l'ordre de 50 (en valeur absolue), il apparaît de plus des motifs récurrents (par exemple l'apparition de valeurs très négatives pour $j_2 = 2$ et lorsque θ_1 varie), ceci suggère que le modèle est incomplet ou inadapté pour décrire les propriétés statistiques des observations.

4 Vers la définition d’une méthodologie de comparaison entre observations et simulations

Le travail de ce stage est axé sur la définition d’une méthodologie de comparaison entre simulations et observations du milieu interstellaire. La direction prise dans ce rapport est de fonder cette comparaison sur un modèle statistique à faible nombre de paramètres et capturant l’essentiel des propriétés statistiques des simulations / observations. Une comparaison peut alors être envisagée dans l’espace des paramètres du modèle. Nous discutons dans cette partie quelques questions soulevées par la définition d’une telle comparaison.

4.1 De la difficulté de simuler le milieu interstellaire

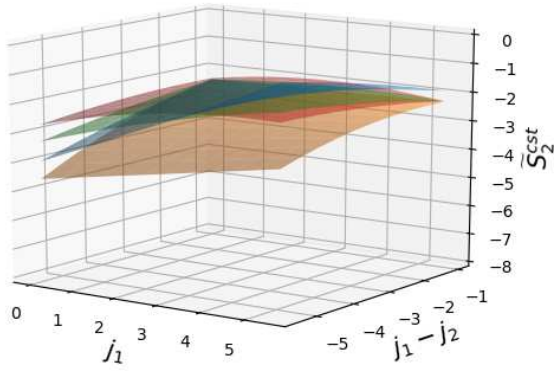
Nous avons vu dans la première partie de ce rapport que la physique du milieu interstellaire met en jeu de nombreux processus hautement non-linéaires. Simuler une telle physique est complexe, et il est en général impossible de reproduire totalement les conditions physiques du milieu interstellaire.

Dans ce travail nous utilisons une observation de l’émission thermique des poussières à $500\ \mu\text{m}$ dans le nuage de Polaris proportionnelle à la densité de colonne de gaz du milieu. Ce traceur est donc comparable aux densités de colonne des cartes simulées présentées dans la deuxième partie de ce rapport.

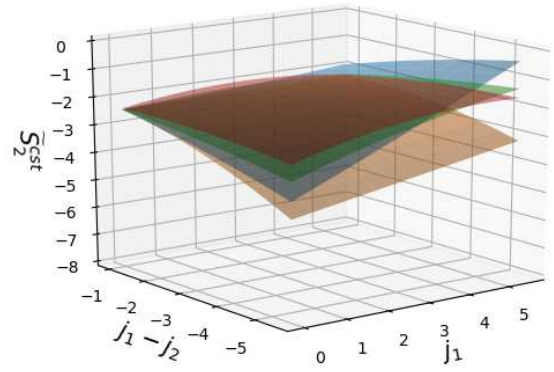
4.2 Définition d’une métrique dans l’espace des paramètres du modèle

Mener une comparaison dans l’espace des paramètres du modèle nécessite de définir une métrique. Une simulation et une observation pourront alors être dites similaires (respectivement différentes) si la distance entre les paramètres modélisant la simulation et l’observation est suffisamment petite (respectivement grande). On peut penser à utiliser une simple distance euclidienne, traitant de manière équitable les différents paramètres du modèle. Néanmoins le modèle développé dans la deuxième partie de ce rapport sur les cartes simulées montre qu’il serait peu judicieux de s’y prendre de la sorte. En effet, pour des cartes isotropes les ajustements des paramètres liés à l’anisotropie des données peuvent se révéler instables. Notamment le paramètre a_{10} qui correspond à l’identification d’une direction privilégiée dans les données est un paramètre non pertinent pour des simulations isotropes, et ne doit donc pas être impliqué dans la métrique dans ce cas précis. Cet exemple de la non-prise en compte des paramètres d’anisotropie pour des données isotropes montre que la métrique à définir ne peut être simplement euclidienne.

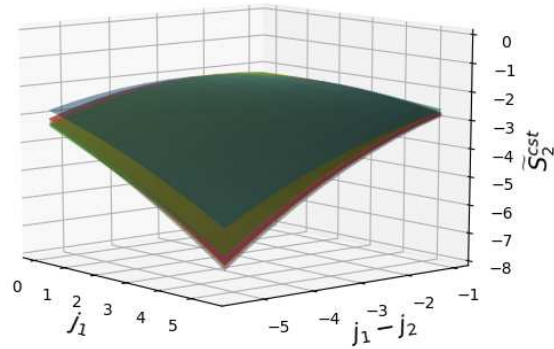
Le modèle développé pour les simulations est un modèle satisfaisant pour décrire les cartes simulées présentées en partie 2, néanmoins il apparaît incomplet voire inadapté pour modéliser les propriétés statistiques de l’observation de Polaris décrite en partie 3. Supposons cependant que le terme $\hat{S}_2^{\text{cst}}x(j_1, j_1 - j_2)$ ajusté sur les simulations et sur l’observation de Polaris modélise au moins de manière satisfaisante les contributions ne dépendant pas des variables angulaires θ_1 et θ_2 , comment s’y prendre alors pour comparer ces contributions ? La figure 18 montre visuellement l’incompatibilité des ajustements entre les 9 classes de simulations et les 4 *clusters* identifiés dans l’observation de Polaris (comme décrit dans la partie précédente). Nous représentons la surface $\hat{S}_2^{\text{cst}}x$ sur la gamme de valeurs de j_1 et $j_1 - j_2$ utilisée pour calculer l’ajustement sur les données. On note la similarité des surfaces pour les différentes classes de cartes simulées. Ces surfaces restent néanmoins visuellement distinctes de celles des *clusters* de Polaris, ceci nous suggère que les contributions non angulaires du modèle pour les simulations ne peuvent être rapprochées à celles des *clusters* de Polaris.



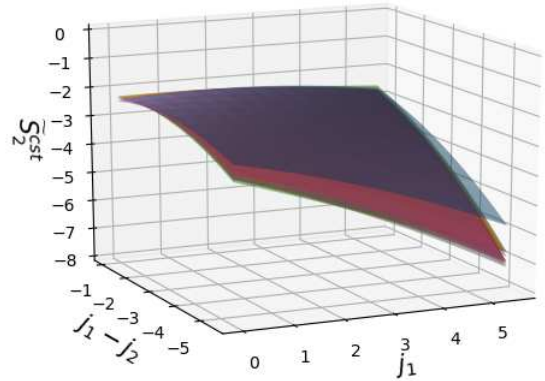
(a) $\tilde{S}_2^{cst}$ ajustée sur les 4 clusters de Polaris



(b) $\tilde{S}_2^{cst}$ ajustée sur les 4 clusters de Polaris (vue alternative)



(c) $\tilde{S}_2^{cst}$ ajustée sur les 9 classes de cartes simulées du MIS



(d) $\tilde{S}_2^{cst}$ ajustée sur les 9 classes de cartes simulées du MIS (vue alternative)

FIGURE 18 – Comparaison des modèles ajustés pour la fonction $\tilde{S}_2^{cst}$ appliqué aux 9 classes de simulations et aux 4 clusters identifiés sur Polaris. Les surfaces ajustées semblent visuellement incompatibles entre simulations et observations.

Conclusion

La portée de ce travail va même au-delà de la caractérisation du milieu interstellaire, tant le problème de caractérisation d'images à la statistique complexe est général et se retrouve dans de nombreux domaines des sciences de l'Univers.

A Champs gaussiens et spectres de puissance

Cette annexe rappelle les notions élémentaires caractérisant un champ gaussien aléatoire homogène et précise les notations et conventions employées dans ce rapport.

A.1 Définitions préliminaires

Plaçons nous dans un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ donné.

Définition 1 (Variable aléatoire). *On appelle variable aléatoire toute application mesurable X de $(\Omega, \mathcal{F}, \mathbb{P})$ dans $\mathbb{R}$ muni de la tribu borélienne $\mathcal{B}(\mathbb{R})$. On note $\mathcal{V}$ l'ensemble des variables aléatoires.*

Définition 2 (Champ aléatoire). *On appelle champ aléatoire de dimension d toute application $X : \mathbb{R}^d \rightarrow \mathcal{V}$. On note les fonctions de distribution multivariées d'un champ aléatoire X sous la forme $F_p(x_1, \dots, x_p; \mathbf{r}_1, \dots, \mathbf{r}_p) = \mathbb{P}(X(\mathbf{r}_1) \leq x_1, \dots, X(\mathbf{r}_p) \leq x_p)$*

Définition 3 (Homogénéité). *On dira qu'un champ aléatoire est homogène (ou stationnaire au sens strict) si ses fonctions de distribution multivariées vérifient pour tout $\delta \in \mathbb{R}^d$:*

$$F_p(x_1, \dots, x_p; \mathbf{r}_1 + \delta, \dots, \mathbf{r}_p + \delta) = F_p(x_1, \dots, x_p; \mathbf{r}_1, \dots, \mathbf{r}_p)$$

Définition 4 (Autocorrélation). *On définit la fonction d'autocorrélation d'un champ aléatoire X (ou fonction à deux points) par : $A_X(\mathbf{r}_1, \mathbf{r}_2) = \mathbb{E}(X(\mathbf{r}_1) X(\mathbf{r}_2))$. Dans le cas d'un champ aléatoire homogène $A_X(\mathbf{r}_1, \mathbf{r}_2)$ ne dépend que de la différence $\mathbf{r}_2 - \mathbf{r}_1$, on notera alors $A_X(\mathbf{r}_1, \mathbf{r}_2) = A_X(\mathbf{r}_2 - \mathbf{r}_1)$.*

Pour un champ aléatoire X donné, on notera ses composantes de Fourier (en supposant qu'elles sont bien définies) :

$$\tilde{X}(\mathbf{k}) = \int_{\mathbb{R}^d} X(\mathbf{r}) e^{-i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r}$$

La convention de Fourier choisie conduit à la relation inverse :

$$X(\mathbf{r}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \tilde{X}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{r}} d\mathbf{k}$$

Définition 5 (Spectre de puissance). *Pour un champ aléatoire homogène X donné, on appelle spectre de puissance associé au champ aléatoire X la fonction :*

$$P_X(\mathbf{k}) = \int_{\mathbb{R}^d} A_X(\mathbf{r}) e^{-i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r}$$

On montre alors :

Proposition 1. *Le spectre de puissance d'un champ aléatoire homogène X vérifie avec ces conventions :*

$$\mathbb{E} \left(\tilde{X}(\mathbf{k}_1) \tilde{X}(\mathbf{k}_2) \right) = (2\pi)^d \delta_d(\mathbf{k}_1 + \mathbf{k}_2) P(\mathbf{k}_2)$$

Le spectre de puissance est une quantité qui est donc directement reliée aux corrélations entre échelles à deux points du champ aléatoire homogène considéré.

Notons que cette quantité est l'équivalent continu de la densité spectrale de puissance définie dans l'annexe B.

A.2 Champs gaussiens aléatoires

Définition 6 (Champ gaussien aléatoire). *Un champ aléatoire X est dit gaussien si toute combinaison linéaire de la forme $\alpha_1 X(\mathbf{r}_1) + \dots + \alpha_p X(\mathbf{r}_p)$ suit une loi gaussienne.*

Proposition 2 (Caractérisation des champs gaussiens homogènes). *Un champ gaussien aléatoire homogène X est entièrement défini par sa moyenne $\mathbb{E}(X)$ (indépendante de la position à cause de l'homogénéité) et son spectre de puissance P_X .*

Notons que ce dernier résultat est une conséquence du théorème de Wick appliqué aux fonctions de corrélations à n points dans l'espace de Fourier du champ (on se référera à [2], chapitre 3, pour plus de détails).

B Méthodes numériques d'estimation de spectres de puissance

Nous décrivons dans cette annexe deux méthodes standards d'estimation de la densité spectrale de puissance (*Power Spectral Density* ou PSD) d'un signal discret aléatoire stationnaire et unidimensionnel $\{y(t), t = 1, \dots, N\}$, échantillonné à une fréquence $f_s = \frac{1}{T_s}$ et supposé de moyenne nulle ($\mathbb{E}(y(t)) = 0$). Ici la variable t s'exprime en unités de périodes d'échantillonnage T_s . Le vocabulaire employé fera référence à une évolution temporelle du signal mais reste transposable à un signal évoluant spatialement (à deux dimensions ou plus). Cette annexe résume les idées développées dans [30] (chapitres 1 et 2).

B.1 Le périodogramme et ses limites

On définit la densité spectrale de puissance du signal $\{y(t)\}$ par :

$$\phi(\omega) = \lim_{N \rightarrow +\infty} \mathbb{E} \left(\frac{1}{N} \left| \sum_{t=1}^N y(t) e^{-i\omega t} \right|^2 \right)$$

La méthode du périodogramme repose sur l'estimateur suivant :

$$\hat{\phi}_p(\omega) = \frac{1}{N} \left| \sum_{t=1}^N y(t) e^{-i\omega t} \right|^2$$

Cette expression découle directement de la définition de la PSD dans la mesure où l'espérance et la limite ne peuvent être obtenues avec un nombre fini d'échantillons. Cependant elle montre vite ses limites lorsqu'on s'intéresse au risque quadratique de cet estimateur $\mathcal{R}(\hat{\phi}_p(\omega), \phi(\omega)) = \mathbb{E} \left(\left| \hat{\phi}_p(\omega) - \phi(\omega) \right|^2 \right)$. En développant ce risque quadratique on a :

$$\mathcal{R}(\hat{\phi}_p(\omega), \phi(\omega)) = \left[\mathbb{E}(\hat{\phi}_p(\omega)) - \phi(\omega) \right]^2 + \text{Var}(\hat{\phi}_p(\omega)) = \text{Biais}(\hat{\phi}_p(\omega))^2 + \text{Var}(\hat{\phi}_p(\omega))$$

Ainsi la qualité de l'estimateur du périodogramme (au sens du risque quadratique) repose sur la valeur de son biais et de sa variance. On peut montrer que même si en pratique l'estimateur du périodogramme est légèrement biaisé, asymptotiquement (pour $N \rightarrow +\infty$) ce n'est plus le cas, mais que par contre sa variance est importante. Plus précisément on montre (*cf* [30]), que pour une classe raisonnablement large de signaux (bruit blanc gaussien filtré linéairement précisément), la variance vérifie :

$$\lim_{N \rightarrow +\infty} \text{Var}(\hat{\phi}_p(\omega)) = \phi^2(\omega)$$

Cette variance asymptotique est donc comparable au carré du signal à estimer, ce qui signifie que le périodogramme fluctue largement autour de la valeur théorique de la PSD.

B.2 La méthode de Welch

La méthode de Welch se fonde sur la méthode du périodogramme et vise à offrir plus de contrôle sur le biais de l'estimateur et à permettre de réduire la variance *via* un compromis sur la résolution de la PSD estimée.

Plus précisément, cette méthode introduit :

- un fenêtrage du signal (appelé encore « apodisation ») qui agit sur le biais et la variance de l'estimateur. Notamment, la taille de la fenêtre permet de contrôler le compromis résolution / variance de l'estimateur.

- un sous-échantillonnage du signal : en découpant le signal en sous-segments contenant M échantillons et pouvant éventuellement se recouvrir (on parlera d'*oversampling*), la variance peut être réduite.

Formellement, pour un sous-ensemble du signal $\{y_j(t), t = 1, \dots, M\}$ (j fait référence au sous-ensemble considéré) on définit :

$$\hat{\phi}_j(\omega) = \frac{1}{MP} \left| \sum_{t=1}^M y_j(t) e^{-i\omega t} \right|^2$$

avec $\{v(t)\}$ la fenêtre temporelle et $P = \frac{1}{M} \sum_{t=1}^M |v(t)|^2$ sa « puissance ».

L'estimateur de Welch est alors défini comme valant :

$$\hat{\phi}_W(\omega) = \frac{1}{S} \sum_{j=1}^S \hat{\phi}_j(\omega)$$

avec S le nombre de sous-ensembles du signal considéré. C'est cet estimateur que nous utiliserons dans les calculs de PSD menés dans ce rapport.

C Reconstruction du potentiel de lentillage faible de l'Univers : définition d'un estimateur statistique

Nous décrivons ici succinctement l'état de l'art de la reconstruction du champ de déflexion du CMB tel qu'employé pour la mission Planck. Nous construisons un estimateur du potentiel de lentillage non biaisé et de variance minimale reposant sur les moments d'ordre 2 des fluctuations de température. Nous en déduisons un moyen de construire un estimateur du spectre de puissance du potentiel de lentillage directement.

C.1 Point de départ et notations

Le fond diffus cosmologique (CMB par la suite) prend la forme numérique d'une carte en température et en polarisation du ciel. La suite de ce document ne prend en considération que la carte en température, délaissant donc toute l'information contenue dans la carte de polarisation, par simple souci de simplification. Nous nous plaçons dans l'approximation du ciel plat (*flat-sky approximation*) qui consiste à considérer une petite partie du ciel assimilable à un plan.

On définit le champ de fluctuations de température Θ par la relation : $\Theta(\mathbf{x}) = \frac{T(\mathbf{x}) - \bar{T}}{\bar{T}}$.

Le long de son trajet, la lumière du CMB a été perturbée avant de nous parvenir aujourd'hui. Notamment, elle a été perturbée par la distribution de masse dans l'univers (effet de relativité générale). La lumière observée aujourd'hui apparaît donc comme un traceur de la distribution de masse dans l'Univers.

On note ψ le potentiel de lentillage de l'Univers, et $\tilde{\Theta}$ le champ de fluctuations de température lentillé par la distribution de masse. On a la relation fondamentale suivante entre les grandeurs :

$$\tilde{\Theta}(\mathbf{x}) = \Theta(\mathbf{x} + \nabla\psi(\mathbf{x}))$$

Θ est un champ gaussien homogène isotrope dont on connaît théoriquement le spectre de puissance C_l^Θ . ψ est aussi un champ aléatoire, approximativement gaussien pour des échelles suffisamment grandes en pratique (sans incidence sur la suite du raisonnement).

C.2 Reconstruire le potentiel de lentillage ψ

Les déformations induites par le champ de déflexion $\nabla\psi$ sur la carte des fluctuations de température sont petites et un développement perturbatif est possible :

$$\tilde{\Theta}(\mathbf{x}) = \Theta(\mathbf{x}) + \nabla\psi(\mathbf{x}) \cdot \nabla\Theta(\mathbf{x}) + o(\nabla\psi(\mathbf{x}))$$

Nous disposons de $\tilde{\Theta}$ et de ses dérivées et souhaitons alors isoler le potentiel de lentillage de manière à l'exprimer en fonction des quantités connues.

On trouve au plus simple [13] :

$$\langle \tilde{\Theta}(\mathbf{x}) \nabla \tilde{\Theta}(\mathbf{x}) \rangle_\Theta = \nabla\psi(\mathbf{x}) \langle \Theta(\mathbf{x}) \Delta \Theta(\mathbf{x}) \rangle_\Theta + o(\nabla\psi(\mathbf{x}))$$

De manière duale, dans l'espace de Fourier on trouve :

$$\langle \tilde{\Theta}(\mathbf{l}) \tilde{\Theta}^*(\mathbf{l} - \mathbf{L}) \rangle_\Theta = \delta(\mathbf{L}) C_l^\Theta + \psi(\mathbf{L}) \times f(\mathbf{l}, \mathbf{L}) + o(\psi(\mathbf{L}))$$

avec $f(\mathbf{l}, \mathbf{L}) = \frac{1}{2\pi} ((\mathbf{L} - \mathbf{l}) \cdot \mathbf{L} C_{|\mathbf{l} - \mathbf{L}|}^\Theta + \mathbf{l} \cdot \mathbf{L} C_l^\Theta)$

On cherche alors à construire un estimateur permettant d'extraire cette information au premier ordre. On part de la classe d'estimateur suivante :

$$\hat{\psi}(\mathbf{L}) = N(\mathbf{L}) \int \frac{d^2\mathbf{l}}{2\pi} \tilde{\Theta}(\mathbf{l}) \tilde{\Theta}^*(\mathbf{l} - \mathbf{L}) g(\mathbf{l}, \mathbf{L})$$

La fonction N apparaît ici comme un facteur multiplicatif global de notre estimateur, et la fonction g comme une fonction de poids. On impose deux contraintes pour fixer ces fonctions :

1. On veut un estimateur non biaisé : $\langle \hat{\psi}(\mathbf{L}) \rangle_{\Theta} = \psi(\mathbf{L})$ (la moyenne est faite sur les réalisations du CMB)
2. On veut un estimateur de variance minimale. Plus précisément, on a :

$$\langle |\hat{\psi}(\mathbf{L}) - \psi(\mathbf{L})|^2 \rangle_{\Theta, \psi} = \langle |\hat{\psi}(\mathbf{L})|^2 \rangle_{\Theta, \psi} + \mathcal{O}(C_l^\psi)$$

On cherche à minimiser l'ordre zéro de ce développement (seul terme sur lequel on peut influencer).
Finalement avec ces contraintes on obtient l'estimateur non biaisé et de variance minimale suivant :

$$\hat{\psi}(\mathbf{L}) = N(\mathbf{L}) \mathbf{L} \cdot \int \frac{d^2 \mathbf{l}}{2\pi} \frac{C_l^\Theta \tilde{\Theta}(\mathbf{l}) \tilde{\Theta}^*(\mathbf{l} - \mathbf{L})}{\tilde{C}_l^{tot} \tilde{C}_{|\mathbf{L}-\mathbf{l}|}^{tot}}$$

avec $N(\mathbf{L})^{-1} = \int d^2 \mathbf{l} \frac{(f(\mathbf{l}, \mathbf{L}))^2}{2 \tilde{C}_l^{tot} \tilde{C}_{|\mathbf{L}-\mathbf{l}|}^{tot}}$ et $\tilde{C}_l^{tot} = C_l^\Theta + N_l$ (N_l correspond à un bruit instrumental).

C.3 Reconstruire le spectre de puissance C_l^ψ

La reconstruction du spectre de puissance repose sur l'estimateur précédent. On part d'un estimateur du type :

$$\hat{C}_L^\psi \propto \sum_{|\mathbf{l}|=L} |\hat{\psi}(\mathbf{l})|^2$$

Cet estimateur fait intervenir la fonction de corrélation à 4 points de $\tilde{\Theta}$. A noter que tel que défini cet estimateur est biaisé, il faut donc retrancher des termes pour enlever les biais.

Références

- [1] J. P. Bernard, A. Abergel, I. Ristorcelli, F. Pajot, J. P. Torre, F. Boulanger, M. Giard, G. Lagache, G. Serra, J. M. Lamarre, J. L. Puget, F. Lepeintre, and L. Cambrésy. PRONAOS observations of MCLD 123.5 + 24.9 : cold dust in the Polaris cirrus cloud. *Astronomy & Astrophysics*, 347 :640–649, July 1999.
- [2] F. Bernardeau. Cosmologie : Des fondements théoriques aux observations. *EDP Sciences, CNRS Editions*, 2007.
- [3] BICEP2 Collaboration. Detection of B-Mode Polarization at Degree Angular Scales by BICEP2. *Physical Review Letters*, 112(24) :241101, June 2014.
- [4] BICEP2/Keck Collaboration and Planck Collaboration. Joint Analysis of BICEP2/Keck Array and Planck Data. *Physical Review Letters*, 114(10) :101301, March 2015.
- [5] M.A. Branch, T.F. Coleman, and Y. Li. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM Journal on Scientific Computing*, 21(1) :1–23, 1999.
- [6] S. Brooks, A. Gelman, G.L. Jones, and X.L. Meng. Handbook of Markov Chain Monte Carlo. *Chapman & Hall / CRC*, 2011.
- [7] J. Bruna and S. Mallat. Invariant Scattering Convolution Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8) :1872, August 2013.
- [8] Joan Bruna, Stéphane Mallat, Emmanuel Bacry, and Jean-François Muzy. Intermittent process analysis with scattering moments. *Ann. Statist.*, 43(1) :323–351, 02 2015.
- [9] B. Burkhart, D. C. Collins, and A. Lazarian. Observational Diagnostics of Self-gravitating MHD Turbulence in Giant Molecular Clouds. *The Astrophysical Journal*, 808 :48, July 2015.
- [10] B. Burkhart and A. Lazarian. The Phase Coherence of Interstellar Density Fluctuations. *The Astrophysical Journal*, 827 :26, August 2016.
- [11] E. Falgarone, J.-F. Panis, A. Heithausen, M. Perault, J. Stutzki, J.-L. Puget, and F. Bensch. The IRAM key-project : small-scale structure of pre-star-forming regions. I. Observational results. *Astronomy & Astrophysics*, 331 :669–696, March 1998.
- [12] M. J. Griffin, A. Abergel, A. Abreu, et al. The Herschel-SPIRE instrument and its in-flight performance. *Astronomy & Astrophysics*, 518 :L3, July 2010.
- [13] D. Hanson, A. Challinor, and A. Lewis. Weak lensing of the CMB. *General Relativity and Gravitation*, 42 :2197–2218, September 2010.
- [14] P. Hily-Blant and E. Falgarone. Intermittency of interstellar turbulence : parsec-scale coherent structure of intense, velocity shear. *Astronomy & Astrophysics*, 500 :L29–L32, June 2009.
- [15] W. Hu and T. Okamoto. Mass Reconstruction with Cosmic Microwave Background Polarization. *The Astrophysical Journal*, 574 :566–574, August 2002.
- [16] Lequeux James, Edith Falgarone, and Charles Ryter. *Le milieu interstellaire*. EDP Sciences, Savoirs actuels et CNRS Editions, 2002.
- [17] M. Kesden, A. Cooray, and M. Kamionkowski. Lensing reconstruction with CMB temperature and polarization. *Physical Review D*, 67(12) :123507, June 2003.
- [18] F. Levrier, E. Falgarone, and F. Viallefond. Fourier phase analysis in radio-interferometry. *Astronomy & Astrophysics*, 456 :205–214, September 2006.
- [19] A. Lewis and A. Challinor. Weak gravitational lensing of the CMB. *Physics Reports*, 429 :1–65, June 2006.
- [20] I. Makarenko, P. Bushby, A. Fletcher, R. Henderson, N. Makarenko, and A. Shukurov. Topological data analysis and diagnostics of compressible mhd turbulence. *arXiv*, 1804.04688, 2018.
- [21] S. Mallat. Group Invariant Scattering. *Communications in Pure and Applied Mathematics*, 65(10) :1331–1398, October 2012.
- [22] M.-A. Miville-Deschênes, P. G. Martin, et al. Herschel-SPIRE observations of the Polaris flare : Structure of the diffuse interstellar medium at the sub-parsec scale. *Astronomy & Astrophysics*, 518 :L104, July 2010.

- [23] G. V. Panopoulou, I. Psaradaki, and K. Tassis. The magnetic field and dust filaments in the Polaris Flare. *Monthly Notices of the Royal Astronomical Society*, 462 :1517–1529, October 2016.
- [24] G. L. Pilbratt, J. R. Riedinger, T. Passvogel, G. Crone, D. Doyle, U. Gageur, A. M. Heras, C. Jewell, L. Metcalfe, S. Ott, and M. Schmidt. Herschel Space Observatory. An ESA facility for far-infrared and submillimetre astronomy. *Astronomy & Astrophysics*, 518 :L1, July 2010.
- [25] Planck Collaboration. Planck 2015 results. I. Overview of products and scientific results. *Astronomy & Astrophysics*, 594 :A1, September 2016.
- [26] Planck Collaboration, A. Abergel, P. A. R. Ade, N. Aghanim, M. I. R. Alves, G. Aniano, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, and et al. Planck 2013 results. XI. All-sky model of thermal dust emission. *Astronomy & Astrophysics*, 571 :A11, November 2014.
- [27] Planck Collaboration, P. A. R. Ade, N. Aghanim, D. Alina, et al. Planck intermediate results. XX. Comparison of polarized thermal emission from Galactic dust with simulations of MHD turbulence. *Astronomy & Astrophysics*, 576 :A105, April 2015.
- [28] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, and et al. Planck 2015 results. XV. Gravitational lensing. *Astronomy & Astrophysics*, 594 :A15, September 2016.
- [29] E. F. Schlafly, G. Green, D. P. Finkbeiner, H.-W. Rix, E. F. Bell, W. S. Burgett, K. C. Chambers, P. W. Draper, K. W. Hodapp, N. Kaiser, E. A. Magnier, N. F. Martin, N. Metcalfe, P. A. Price, and J. L. Tonry. A Large Catalog of Accurate Distances to Molecular Clouds from PS1 Photometry. *The Astrophysical Journal*, 786 :29, May 2014.
- [30] P. Stoica and R.L. Moses. *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.