Unveil the Local Universe.

Aurélien Valade

UNIVERSITÉ DE LYON

Université Claude Bernard Lyon 1

THÈSE de DOCTORAT DE L'UNIVERSITE CLAUDE BERNARD LYON 1

En cotutelle avec :
l'Université de Potsdam, Allemagne

Ecole Doctorale 52
Physique et Astrophysique

Discipline : Cosmologie

Soutenue publiquement le 12/06/2023, par :

Aurélien Niels Valentin Valade

Unveil the Local Universe

Devant le jury composé de :

Jounghun LEE **Professeure, Université de Séoul, CORÉE**	**Président.e** **Rapporteure**
Rien VAN DE WEIJGAERT Professeur, Université de Groningen, PAYS-BAS	**Rapporteur**
Sophie CODIS Chargé de recherches, Université de la Sorbonne, FRANCE	**Examinatrice**
Yannick COPIN Professeur associé, Université Claude Bernard Lyon 1, FRANCE	**Examinateur**
Noam LIBESKIND Chargé de recherches, Université de Potsdam, ALLEMAGNE	**Examinateur**
Jenny SORCE Chargé de recherches, Université de Lille, FRANCE	**Examinatrice**
Anne EALET Professeure, Université Claude Bernard Lyon 1, FRANCE	**Directrice de thèse**
Matthias STEINMETZ Professeur, Université de Potsdam, ALLEMAGNE	**Co-Directeur de thèse**

Abstract

Galaxies in the Universe form a gigantic, complex edifice, called the Large Scale Structure (LSS). Still, the vast majority of the matter is thought to be dark, *i.e.*not directly observable by our telescopes and detectable solely through its gravitational interaction with its surrounding. The relationship between the distribution of galaxies and the total matter field is still not fully described and thus, the LSS cannot be reduced to the galaxies that inhabit it.

Unveiling the matter distribution and the associated velocity field in the Local Universe is an extremely difficult task. One approach consists in combining, for each galaxy, a measurement of redshift and estimation of distance to obtain its velocity with respect to its local environment. With the only source of motion on these scales being gravitation, and using the fact that the velocity field is tightly linked to the matter distribution, the two can be reconstructed together.

Yet, estimations of distances, and thus of velocities, are difficult to make: data are sparse, tainted with errors and plagued with observational biases. Only the radial component of the velocity can be measured and the error size grows with the distance. Powerful mathematical methods need thus be employed.

Our method follows the Bayesian inference approach developed in the last decade to over-come the short-falling of the Wiener Filter methodology, whose simplistic modeling of the data requires a somewhat ad-hoc treatment of the data beforehand. The first step in Bayesian inference is the description of the conditional probability of a set of parameters of a given model given a set of observations. The second step is the creation of a series of realizations of this probability law with a Monte Carlo method, on which summary statistics can be computed.

However, this process is computationally very costly, and the previously developed methods were unable to face the growing size of the current and future problems. This work answers this issue thanks to two major innovations. Firstly, with the use of the cutting edge Hamiltonian Monte Carlo method to create the realizations of the posterior (no comma) and secondly, with the implementation of an GPU accelerated code which reduces the computation time by orders of magnitude: the HAmiltonian Monte carlo reconstruction of the Local EnvironmentT (Hamlet).

Hamlet is first applied to mock data consistent with the implemented model. We demonstrate that our method converges properly with the number of constraints and the amplitude of the uncertainties.

Then, Hamlet is ran on another mock catalog extracted from a dark matter only cosmological simulation. In parallel, we apply to the same catalog the most recent instance of the canonical twofold approach: the Bias Gaussianization correction / Wiener Filter (BGc/WF) pipeline. In comparison to the BGc/WF, Hamlet is able to extract more information from the data and produce maps with a higher contrast. However, some new biases appear in its reconstruction.

Finally, we apply Hamlet to the latest release of the Cosmicflows peculiar velocity catalogs: Cosmicflows-4. We compare our reconstruction of the matter distribution to a large compilation of redshift surveys, demonstrating a remarkable matching between the two. We analyze the Basins of Attraction of the velocity field as well as the monopole and dipole, showing a mild tension with ΛCDM .

Résumé (version courte)

Les galaxies de l'Univers forment un édifice gigantesque et complexe, appelé structure à grande échelle (LSS). Cependant, la grande majorité de la matière est considérée comme sombre, c'est-à- dire qu'elle n'est pas directement observable par nos télescopes et qu'elle n'est détectable que par son interaction gravitationnelle avec son environnement. La relation entre la distribution des galaxies et celle de la matière n'étant à ce jour pas entièrement bien bien décrite, la LSS ne peut être réduit aux galaxies qui l'habitent.

Dévoiler la distribution de la matière et le champ de vitesse associé dans l'Univers local est une tâche extrêmement difficile. Une approche consiste à combiner, pour chaque galaxie, une mesure du décalage vers le rouge et une estimation de la distance pour obtenir sa vitesse par rapport à son environnement local. La seule source de mouvement à ces échelles étant la gravitation, le champ de vitesse est étroitement lié à la distribution de matière, et les deux peuvent être reconstruits ensemble.

Cependant, l'estimation des distances, et donc des vitesses, est difficile à réaliser : les données sont rares, entachées d'erreurs et entachées de biais d'observation. Seule la composante radiale de la vitesse peut être mesurée et la taille de l'erreur croît avec la distance. Des méthodes mathématiques puissantes doivent donc être employées.

Notre méthode suit l'approche de l'inférence bayésienne développée au cours de la dernière décennie pour pallier les insuffisances de la méthodologie du filtre de Wiener, dont la modélisation simpliste des données nécessite un traitement quelque peu ad hoc des données préalables. La première étape de l'inférence bayésienne consiste à écrire la probabilité conditionnelle d'un ensemble de paramètres d'un modèle donné, à partir d'un ensemble d'observations. La deuxième étape consiste à créer une série de réalisations de cette loi de probabilité à l' aide d'une méthode de Monte Carlo, sur laquelle des statistiques sommaires peuvent être calculées.

Cependant, ce processus est très coûteux en temps de calcul, et les méthodes développées précédemment n'ont pas pu faire face à la taille croissante des problèmes actuels et futurs. Ce travail répond à cette question grâce à deux innovations majeures. Tout d'abord avec l'utilisation de la méthode de pointe Hamiltonian Monte Carlo pour créer les réalisations du postérieur, et avec l'implémentation d'un code accéléré par le GPU qui réduit le temps de calcul par des ordres de grandeur : la reconstruction HAmiltonian Monte carlo de l'environnement localT (Hamlet).

Hamlet est d'abord appliqué à des données fictives compatibles avec le modèle mis en œuvre. Nous démontrons que notre méthode converge correctement avec le nombre de contraintes et l'amplitude des incertitudes.

Ensuite, Hamlet est exécuté sur un autre catalogue fictif extrait d'une simulation cosmologique de matière noire uniquement. Parallèlement à Hamlet , nous appliquons au même catalogue l'exemple le plus récent de l'approche double canonique : le pipeline Bias Gaussianization correction / Wiener Filter (BGc/WF). Par rapport au BGc/WF, Hamlet est capable d'extraire plus d'informations des données et de produire des cartes plus contrastées. Cependant, de nouveaux biais apparaissent dans sa reconstruction.

Enfin, nous appliquons Hamlet à la dernière version des catalogues de vitesses particulières Cosmicflows : Cosmicflows-4. Nous comparons notre reconstruction de la distribution de matière à une large compilation de relevés de décalage vers le rouge, démontrant une correspondance remarquable entre les deux. Nous analysons les bassins d'attraction du champ de vitesse ainsi que le monopôle et le dipôle, montrant une légère tension avec ΛCDM .

Résumé (version longue)

La cosmologie est un domaine en plein essor depuis le début du XXème siècle. Dans la première moitié de ce siècle, les mathématiques fondamentales nécessaires à la physique moderne et à la description statistique du cosmos ont été développées (*e.g.* Einstein, 1916; Friedmann, 1922), tandis que l'existence de plusieurs galaxies et l'expansion de l'Univers ont été mises en évidence (Hubble, 1929; Lemaître, 1933). Dans la seconde moitié du XX^{rmth} siècle, la quantité et la qualité croissantes des données observationnelles ont conduit à quatre découvertes majeures : le fond diffus cosmologique et ses anisotropies (CMB ; Penzias & Wilson, 1965; Smoot et al., 1977, 1992; Fixsen et al., 1996; Komatsu et al., 2011; Planck Collaboration et al., 2016), la structure à grande échelle formées par les galaxies (*Large Scale Structure* (LSS); *e.g.* Peebles, 1980; de Lapparent et al., 1986), l'expansion accélérée de l'Univers (Λ constante ou énergie noire ; Riess et al., 1998) et le problème de la gravitation, résolu dans le modèle actuel par la présence de matière noire (Zwicky, 1937; Rubin et al., 1980) et de matière noire froide (CDM Blumenthal et al., 1984; Davis et al., 1985). Les trente dernières années ont été marquées par l'expansion rapide des capacités de calcul offertes par des technologies en plein essor, qui ont ouvert la voie à des approches beaucoup plus frontales et computationnelles de problèmes qui devaient auparavant être simplifiés pour être résolus analytiquement.

Depuis la découverte de la LSS, des catalogues de redshifts de plus en plus précise profonds et couvrant des parties de plus en plus large du ciel ont été effectués (*e.g.* SDSS, York et al. 2000 ; 2dF, Colless et al. 2001 ; 6dF, Jones et al. 2009 ; DESI DESI Collaboration et al. 2016, etc), confirmant l'existence et la structure à grande échelle de la toile cosmique. En parallèle, des catalogues les distances des galaxies ont été contruits, qui, associées à des mesures de redshift, ont permis de construire des catalogues de vitesses radiales particulières (*e.g.* Great-Attractor, Lynden-Bell et al. 1988 ; SAMURAI, Han & Mould 1992 ; Mark III, Willick et al. 1997 ; SFI++ Springob et al. 2007 ; Cosmicflows Tully et al. 2009, 2013; Tully et al. 2016; Tully et al. 2023).

La structure à grande échelle de l'Univers se révèle dans la distribution des galaxies ainsique que dans leurs vitesses. Ce qui a mené les cosmologistes à étudier l'Univers proche au moyen de catalogues de redshifts de galaxies et de catalogues de vitesses radiales. Deux familles de stratégies ont vu le jour pour reconstruire la LSS.

La première vise à récupérer le champ de densité de matière à partir de la distribution des galaxies dans l'espace des redshifts, à commencer par la reconstruction avec filtre de Wiener de l'IRAS Lahav et al. (1994) et Zaroubi et al. (1995), puis un article ultérieur de Erdoğdu et al. (2004) sur la reconstruction du catalogue 2dF à l'aide de la même méthode. L'approche MCMC de la reconstruction bayésienne à partir de catalogues de redshifts a été inaugurée par Kitaura et al. (2009), suivie par Jasche & Wandelt (2013) et Wang et al. (2014). Cependant, son hypothèse centrale, la relation entre la présence de galaxies et la densité de matière sous-jacente, est mal comprise (Kaiser, 1984; Bardeen et al., 1986; Mo & White, 1996), ce qui menace la fiabilité de leurs estimations.

La deuxième famille de méthodes de reconstruction vise à contraindre le champ de vitesse de l'Univers à partir des vitesses des galaxies (Bertschinger & Dekel, 1989; Dekel et al., 1999; Zaroubi et al., 1999; Lavaux, 2016; Graziani et al., 2019). A ces échelles, la seule source de mouvement étant la succion gravitationnelle, le champ de vitesse est censé retracer la distribution complète de la matière. Ces méthodes ne souffrent pas de la mauvaise compréhension de la relation entre la densité des galaxies et la densité de matière. Elles sont cependant limitées par la mauvaise qualité et la rareté des données résultant de la difficulté d'estimer les vitesses des galaxies lointaines.

En effet, non seulement les données ont un rapport signal/bruit très faible mais elles sont également entachées de forts biais (Strauss & Willick, 1995). Ceci a motivé le développement de plusieurs algorithmes pour corriger les données avant la reconstruction du champ de vitesse par un programme séparé (Sorce, 2015; Hoffman et al., 2021). Cependant, les biais sont nombreux et leur modélisation complexe : toutes les méthodes diffèrent dans leur approche et aucune ne résout l'ensemble du problème de manière frontale.

Ce travail consiste à concevoir, tester et appliquer à des données réelles une méthode qui reconstruit le champ de vitesse linéaire à partir de la mesure des vitesses particulières des galaxies. Elle suit l'exemple de Lavaux (2016); Graziani et al. (2019), qui ont respectivement développé et appliqué un algorithme qui *conjointement* corrige les biais dans les données et reconstruire la vitesse en une seule fois, d'une manière probabiliste, bayésienne et cohérente. La nouveauté de ce travail est triple et consiste en (1) l'amélioration de l'algorithme d'exploration Monte Carlo (2) l'implémentation d'un code accéléré par le GPU et (3) la meilleure modélisation des données, à savoir de sa fonction de sélection. Alors que les méthodes précédentes étaient limitées par leur immense cout computationnel, ces innovations nous

permettent de reconstruire la dernière version des données des catalogues Cosmicflows (CF4 ; Tully et al., 2023) et ouvrent la porte à de futurs développements.

L'inférence bayésienne est la réponse mathématique à la question "à condition d'avoir fait certaines observations et d'avoir un modèle dépendant d'un ensemble de paramètres (ou degrés de liberté), quelles sont les valeurs des paramètres du modèle qui peuvent expliquer au mieux les observations". L'approche de la modélisation prospective, détaillée sous l'angle de son application à la reconstruction du LSS dans chapter 2 et publiée dans Valade et al. (2022), est double.

La première étape consiste à écrire la loi de probabilité conditionnelle de l'ensemble des paramètres compte tenu de l'ensemble des observations et du modèle. Cette probabilité est appelée la distribution postérieure. Elle est le produit d'une fonction de vraisemblance, qui est la probabilité que les observations proviennent de l'ensemble des paramètres, et d'une distribution a-priori, qui est la probabilité des paramètres, indépendamment des observations. Dans notre travail, la distribution postérieure modélise les observations de redshifts et de distances à partir d'un champ de vitesse linéaire, dérivé dans le contexte de ΛCDM. Les paramètres libres sont les modes de Fourier du champ de surdensité linéaire projeté sur une grille ainsi que les distances des contraintes.

Bien que la probabilité a posteriori puisse être écrite analytiquement, sa complexité empêche tout calcul analytique ou même numérique simple des statistiques sommaires (champs moyens, monopole, dipôle, etc.). Le deuxième aspect de la modélisation prospective est donc l'exploration de la probabilité a posteriori par une méthode de Monte Carlo. Une telle méthode génère une longue série arbitraire de réalisations de la probabilité postérieure, sur laquelle des statistiques sommaires peuvent être calculées. L'une des innovations de ce travail est le remplacement de l'échantillonnage de Gibbs utilisé par Graziani et al. (2019) par une méthode d'exploration de pointe : le Monte Carlo Hamiltonien (HMC ; Hoffman & Gelman, 2011). Le HMC utilise les équations hamiltoniennes pour intégrer les trajectoires dans l'espace des paramètres, ce qui permet au processus d'exploration de faire de grands pas dans des espaces de paramètres hautement dimensionnels et donc de résoudre partiellement à ce que l'on appelle la "malédiction des dimensions" – l'augmentation en loi de puissance du temps d'exploration nécessaire en fonction du nombre de données. Cette innovation n'est pas seulement technique : elle permet de faire un bond en avant dans l'applicabilité de la méthode à de grandes distances et de réaliser les plus grandes reconstructions à partir de vitesses particulières de l'univers à ce jour .

La deuxième amélioration clé de ce travail est la mise en œuvre d'un code, Hamlet , qui est conçu pour fonctionner sur GPU. Cette accélération réduit le temps d'exécution de plusieurs ordres de grandeur par rapport à un algorithme précédemment conçu qui était limité dans ses capacités en raison de l'utilisation inefficace des ressources de calcul[1]. En outre, les cadres physique et mathématique sont extrêmement souples et permettent d'améliorer encore le modèle physique et les méthodes d'exploration. Cette flexibilité est reflétée par l'extrême modularité du code.

La dernière amélioration majeure de cette thèse est la modélisation des coupures de redshift dans les données suivant Hinton et al. (2017). Elle est présentée dans le chapter 4, avant l'application à Cosmicflows-4.

Dans chapter 2, directement après la présentation de la méthode, celle-ci est testée sur des données fictives qui sont entièrement conformes à ce que le modèle attend (champ linéaire et même description des erreurs). Un catalogue fictif de référence est construit avec une taille de et une modélisation des erreurs qui reproduit le catalogue Cosmicflows-3, qui était au moment de ce chapitre le catalogue de vitesses radiales le plus fourni .

La fonction de sélection de ce catalogue est isotrope et vise un nombre fixe de points par coquille de distance (donc une densité diminuant comme le carré de la distance) qui correspond à peu près à celle de Cosmicflows-3. D'autres catalogues fictifs sont créés en variant le nombre de contraintes (par un facteur 1/2 et 2) et l'amplitude des erreurs (par un facteur 1/2 et 1/10) du catalogue de référence, conduisant à un nombre total de 9 catalogues fictifs. Les résultats se concentrent sur (1) la moyenne et l'écart-type du champ de surdensité (2) la moyenne et l'écart-type du champ de vitesse et (3) les premiers moments du champ de vitesse (*i.e.* monopôle et dipôle).

Nous montrons que pour ces quantités, notre code converge bien et de manière attendue avec la quantité et la qualité des contraintes. Cette étude nous permet également de donner une première estimation quantitative de l'incertitude sur ces mesures et des prévisions pour les applications futures.

Dans chapter 3, publié dans Valade et al. (2023), Hamlet est testé sur des observations fictives issues

[1]Les différences méthodologiques et numériques entre ce code et celui de Graziani et al. (2019) rendent le calcul exact du facteur d'accélération non trivial. Il semble varier autour de quatre ordres de grandeur. En pratique, un résultat peut être obtenu en quelques minutes avec la présente méthode, alors que des mois ont été nécessaires avec le code de Graziani et al. (2019).

d'un univers simulé. La taille du catalogue et la modélisation des erreurs sont à nouveau liées à celles de Cosmicflows-3, étendu de quelques milliers de points. Alors que la fonction de sélection est très basique dans le premier test, ce catalogue fictif reproduit avec une grande fidélité l'empreinte de Cosmicflows-3 dans l'espace des redshifts, avec notamment une asymétrie hémisphérique et la zone d'ombre de notre galaxie. Les données générée atteignent une 160 Mpc/h. La construction de ce catalogue fictif est réalisée par un algorithme avancé de type Monte-Carlo qui effectue la sélection des halos de matière noire de la simulation (qui représentent des galaxies et des groupes de galaxies).

Deux autres méthodes sont appliquées aux mêmes données fictives afin d'évaluer et de comparer quantitativement la qualité des reconstructions : (1) la Bias Gaussianization correction (BGc) est appliquée aux données, qui sont ensuite données au filtre de Wiener pour construire les champs (pipeline BGc/WF) et (2) les observations fictives *sans* erreurs sont données directement au filtre de Wiener (pipeline Ex/WF). Alors que les deux premières méthodes (Hamlet et BGc/WF) sont des méthodes qui peuvent être effectivement appliquées à des données réelles, la dernière est utilisée comme un scénario hypothétique du meilleur cas, dans lequel la position et la vitesse de chaque contrainte ont été parfaitement récupérées. Il nous permet de quantifier les erreurs et les incertitudes résultant (1) de la rareté des données et (2) de l'application de la théorie linéaire à un univers non linéaire. Les résultats se concentrent à nouveau sur la moyenne et l'écart-type des champs de surdensité et de vitesse, ainsi que sur les moments des champs de vitesse.

Nous démontrons qu'en l'absence d'erreur d'observation, la qualité de la reconstruction est proche de la perfection sur un très grand volume, même avec une description des champs simplifiée et un ensemble limité de contraintes. En présence d'erreur, la situation est cependant différente. Les méthodes Hamlet et BGc/WF donnent toutes deux des résultats très similaires dans un volume de 80 Mpc/h, en dehors duquel Hamlet affiche un contraste plus élevé (et donc a priori meilleur) que la BGc/WF, jusqu'à la fin des données à 160 Mpc/h. Cependant, Hamlet a tendance à "sur-évaluer" le contrast dans entre 80 Mpc/h et 160 Mpc/h, ou la méthode produit des vitesses dont l'amplitude dépasse celles de la simulation cible. Enfin, les moments du champs de vitesse sont mieux reconstruits avec la BGc/WF qu'avec Hamlet .

La conclusion de cette étude est que si Hamlet semble extraire beaucoup plus d'informations à partir des mêmes données, la BGc/WF reste une méthode plus conservatrice. En effet, Hamlet semble être sujet à certains biais qui freinent son potentiel prometteur et qui doivent encore être compris et corrigés.

Après avoir été largement testé et comparé à d'autres méthodes dans chapters 2 and 3, Hamlet est appliqué à des données réelles dans la chapter 4, *i.e.* le catalogue de vitesses particulières Cosmicflows-4 (Tully et al., 2023).

Dans cette section, nous poussons à nouveau Hamlet plus loin : la taille des catalogues Cosmicflows-4 dépasse celle de la version précédente d'un facteur trois, et double approximativement le volume (en étendant la région contrainte de 160 Mpc/h à un peu moins de 300 Mpc/h dans une direction du ciel). Plus que de sortir Hamlet de sa "zone de confort computationnelle", l'application à des données réelles est également éprouvante pour la modélisation physique.

En effet, les données sont extrêmement complexes : le catalogue Cosmicflows-4 n'est pas le résultat d'un seul relevé mais plutôt de la fusion de plusieurs relevés de redshifts, étendus par différents catalogues de distance basés sur des méthodes différentes. L'empreinte spatiale est donc très asymétrique, et les erreurs sur chaque mesure ne sont pas triviales. Bien que ces différentes composantes du catalogue global soient inter-étalonnées, la possibilité d'une erreur dans le processus ou dans les sources demeure.

La distribution réelle de la matière dans l'Univers est inconnue. La validité de notre reconstruction sur ce nouvel ensemble de données ne peut donc pas être évaluée de la même manière que dans chapters 2 and 3. Au lieu d'une véritable étude quantitative, nous avons choisi de comparer qualitativement nos champs reconstruits à la distribution dans l'espace des décalages vers le rouge de la base de données des galaxies extragalactiques de Lyon Meudon (LEDA ; Paturel et al., 2003). Même sans supposer une forme spécifique pour le biais des galaxies, on peut s'attendre à une certaine corrélation entre le champ de surdensité et la distribution des galaxies. Le redshift étant un estimateur assez précis de la distance, cela constitue un premier test satisfaisant pour l'application de notre méthode sur des données réelles.

La correspondance entre les galaxies LEDA et notre estimation du champ de surdensité est satisfaisante, les principales déviations étant imputables à la distorsion de l'espace des décalages vers le rouge. Les principales caractéristiques de l'Univers sont retrouvées, et pour la première fois, la distribution de matière est directement mesurée dans la région du Sloan Digital Sky Survey (SDSS). Ce volume contient notamment le célèbre Sloan Great Wall, qui apparaît comme la surdensité la plus importante de notre reconstruction, plus que la concentration de Shapley et le complexe Hercules–CfA-Great-Wall–Coma. Le champ de surdensité montre une structure filamentaire très riche remplissant le volume nouvellement

cartographié du SDSS.

Le champ de vitesse reconstruit est également examiné de près. Une analyse des bassins d'attraction (BoA) montre que les deux principaux attracteurs dans le volume reconstruit sont Shapley et le Sloan Great Wall dans la région de SDSS. Le Groupe local semble être intégré dans le BoA de Shapley. Le superamas d'Hercule est également l'attracteur d'un BoA relativement grand et bien contraint. Les moments du champ de vitesse sont ensuite discutés. Alors que le monopôle ne montre aucun comportement inattendu et est totalement cohérent avec ΛCDM et le spectre de puissance, le dipôle manifeste un comportement quelque peu surprenant autour de $160\,\mathrm{Mpc}/h$, où la composante le long de l'axe X super-galactique et l'amplitude du dipôle s'écartent de l'espérance de 2-σ de ΛCDM . Ce résultat est confirmé dans la littérature (Hoffman et al., 2015; Magoulas et al., 2016; Howlett et al., 2022; Watkins et al., 2023). L'alignement du dipôle avec la vitesse du CMB est également remarquablement élevé et constant et mérite d'être étudié. Enfin, la V-Web est discutée.

Zusammenfassung

Die Galaxien im Universum bilden ein gigantisches, komplexes Gebilde, die sogenannte Large Scale Structure (LSS). Der größte Teil der Materie gilt jedoch als dunkel, d. h. sie ist mit unseren Teleskopen nicht direkt beobachtbar und kann nur durch ihre Gravitationswechselwirkung mit ihrer Umgebung nachgewiesen werden. Die Beziehung zwischen der Galaxienverteilung und der Materieverteilung ist nach wie vor nicht vollständig erforscht, weswegen die LSS nicht auf die ihr innewohnenden Galaxien beschränkt werden kann.

Die Kartierung der Marterieverteilung und des damit verbundenen Geschwindigkeitsfeldes im lokalen Universum ist eine äußerst schwierige Aufgabe. Eine Alternative besteht darin, für jede Galaxie eine Messung der Rotverschiebung und eine Abschätzung der Entfernung zu kombinieren, um ihre Geschwindigkeit in Bezug auf ihre lokale Umgebung zu ermitteln. Da die einzige Bewegungsquelle in Messungen dieser Größenordnung die Gravitation ist, ist das Geschwindigkeitsfeld eng mit der Materieverteilung verknüpft. Diese beiden Einheiten können im selben Schritt rekonstruiert werden.

Doch die Schätzung von Entfernungen und damit von Geschwindigkeiten ist schwierig: Daten sind rar und mit Bias behaftet. Es kann lediglich die Radialgeschwindigkeit der Galaxien beobachtet werden, wobei die Fehlergröße mit der Entfernung zunimmt. Es müssen daher leistungsfähige mathematische Methoden angewendet werden.

Unsere Methode folgt dem Bayesschen Inferenzansatz, der im vergangenen Jahrzehnt entwickelt wurde, um die Mängel der Wiener-Filter-Methode zu beheben, deren vereinfachte Modellierung der Beoachtungen eine gewisse vorhergehende Behandlung der Daten erfordert. Der erste Schritt des Bayesschen Inferenzansatzes ist die Beschreibung der bedingten Wahrscheinlichkeit eines Satzes von Parametern eines gegebenen Modells bei einer Reihe von Beobachtungen. Der zweite Schritt ist die Erstellung einer Reihe von Realisierungen dieses Wahrscheinlichkeitsgesetzes mit einer Monte-Carlo-Methode, auf deren Grundlage zusammenfassende Statistiken berechnet werden können.

Dieses Verfahren ist jedoch sehr rechenintensiv und die zuvor entwickelten Methoden konnten der wachsenden Komplexität der aktuellen und zukünftigen Herausforderungen nicht gerecht werden. Diese Arbeit löst dieses Problem mithilfe zweier wichtiger Neuerungen. Erstens mit der Verwendung der hochmodernen Hamiltonschen Monte-Carlo-Methode, um die Realisierungen des Posteriors zu erstellen, und zweitens mit der Implementierung eines GPU-beschleunigten Codes, der die Rechenzeit erheblich verringert: die HAmiltonian Monte carlo reconstruction of the Local EnvironmentT (Hamlet).

Hamlet wird zuerst auf Scheindaten angewendet, die mit dem implementierten Modell übereinstimmen. Wir zeigen, dass unsere Methode richtig mit der Anzahl der Nebenbedingungen und der Amplitude der Unsicherheiten konvergiert.

Dann wird Hamlet auf einem anderen Scheinkatalog ausgeführt, der aus einer kosmologischen Simulation nur mit dunkler Materie extrahiert wurde. Gleichzeitig wenden wir die jüngste Instanz des kanonischen zweifachen Ansatzes auf denselben Katalog an: die Bias-Gaussianization correction/Wiener-Filter (BGc/WF)-Pipeline. Im Vergleich zu BGc/WF ist Hamlet in der Lage, mehr Informationen aus den Daten zu extrahieren und Karten mit höherem Kontrast zu erstellen. Bei seiner Rekonstruktion treten jedoch einige neue Bias auf.

Schließlich wenden wir Hamlet auf die neueste Ausgabe der Cosmicflows-Kataloge der Pekuliargeschwindigkeiten an: Cosmicflows-4. Wir vergleichen unsere Rekonstruktion der Materieverteilung mit einer großen Zusammenstellung von Durchmusterungen der Rotverschiebungen und demonstrieren eine bemerkenswerte Übereinstimmung zwischen den beiden. Wir analysieren die Basins of Attraction sowie Monopol und Dipol des Geschwindigkeitsfeldes und zeigen eine leichte Spannung mit ΛCDM .

Contents

An equation free introduction

The problem discussed in this work spreads over very different scales and relies on number of concepts and objects. In this short section, we aim at reviewing them in simple, equation free, descriptive manner. At the end of it, a simple explanation of the problem discussed in the work is given. A more technical and precise detailing of the problem comes next. If what is written here is given as a general truth for readability, the reader has to keep in mind that this description of the Universe is the current state of Science, *i.e.* as accepted by (the majority of) the scientific community, and is always subject to change.

The Earth

Our exploration of the Universe is limited by our relatively stationary position on Earth. While telescopes bring the remote Universe to us, it is (apparently) impossible to physically travel through the universe. Being able to observe the Universe only *now and here* results in a series of limitations.

First and foremost, there is only one Universe to observe. In other words, we have a sample size of one. It is impossible to create smaller universes on demand in laboratories and study them. Thus, each object (be it a particular type of star, a rare merger of galaxies, etc) is fundamentally unique. And although astronomers are able to botanically catalog objects, each class is composed of only a very limited sample size. Further more, environmental conditions that cannot be changed or controlled for. To draw conclusions on the physical properties of the entire universe, from our limited view point, a fundamental assumption has to be made, called the Copernicus principle, which states that we are not sitting in a peculiar place of the Universe. We thus expect to be observing the Universe from a reasonably average location.

Secondly, each telescope can only observe part of the sky. For instance, a telescope in any northern latitude can not observe the South Pole (and vice-versa). Since telescopes are expensive and the locations where they can be built are rare, sets of data tend to be rather sparse and our ability to cover the sky is imperfect. The reader will note that this is not entirely true with space based observations, however these are limited by other more practical factors.

Finally, the Earth is not fixed in the Universe. It orbits the Sun, which orbits around the center of the Milky Way, which moves in its local surrounding, as will be detailed in lengths in the work. This is a problem, as we aim at measuring the velocities of other objects of the Universe. We will however shortly see how this problem can be solved.

Stars

Stars are gigantic balls of hot plasma resulting from the collapse of gas by self-gravitations. The most famous of them is of course the Sun, and at night, thousands of others can be observed by the naked eye.

While stars are much smaller than the scales we intend to study in this work, they are very important to us because they shine. Each star emits light some of which will travel across space from one side of the Universe to the other, and that can be detected in various wave ranges (infrared, visible, ultraviolet, etc) by our telescopes. They are, in some sense, the light houses of the Universe. The light house metaphor is particularly apt because star light also wanrs us where the non-luminous matter resides, as this work demonstrates.

Stars come in many types: different colors (blue, red, white, etc), sizes (from fractions of the Sun to thousands times bigger), ages, stages in their lives, chemical compositions (metalicities), and luminosities.

However, like light houses, the more distant they are from us, the dimmer they seem. First and foremos. the light they emit dilutes in space as flux drops with increasing distance. Secondly, photons are scattered and absorbed by the intergalactic and interstellar media; gas or space dust. Thus, looking

at how bright a star appears in our telescopes, it is *a priori* impossible to know just from its magnitude if we are observing a nearby dim star, or a distant bright star.

Some stars have interesting properties that allow us to estimate how bright they truly are, and therefore to estimate how far they are. Some examples are the Cepheid stars, who's periodic variability depends directly on their luminosity, extremely massive and hot stars who sometimes collapse in exceptionally violent and bright explosions (Super-Novae) whose intrinsic luminosity scales with the speed with which their brightness dims or stars whose intrinsic luminosity peaks at the end of the red giant phase of their evolution.

Galaxies

Stars do not form in random places in the Universe, but rather in islands of matter in an ocean of void: galaxies. Indeed, if stars seem to float distant from one another in an extremely empty space, the void that separates galaxies is order of magnitude bigger and emptier. Galaxies come in various shapes and sizes, but three main families can be drawn.

Irregular galaxies – galaxies whose morphology cant be well defined – are very often small (up to about a tenth of our galaxy, the Milky Way) and make up about a quarter of the galaxies in the Universe. Although they are of great scientific interest, they are not relevant to this work.

Spiral galaxies, are of average size (about the size of the Milky Way, which happens to be a very normal galaxy) and compose the majority of the galaxies in the Universe. They are rather flat rotating disks, populated by hundreds of billions of stars, with often a Super Massive black hole at their center. They tend to have two or more "gran design" spiral arms wound up around their center's - some of them also have co-rotating bars in their centers or small spheroids of stars.

Elliptical galaxies, can be very massive (up to hundred times the size of the Milky Way) and are rarer galaxies that are mostly found in very dense regions of the Universe. Their relatively spherical shape is thought to emerge from the merging of several spiral galaxies. They exhibit little to no internal structure.

Galaxies are the building blocks, the atoms of this work. We will consider them as points and not discuss their internal properties, with some exceptions.

As stars are born, live and die in them, galaxies are also luminous object, thus, the image of the light house stands. Just like for stars, it is a priori impossible to know if an observed galaxy is dim but close or bright but far. This limitation is in fact one of the core problems we have to tackle in order to reconstruct the cosmic web.

The Cosmic Web

Galaxies are also not randomly spread across the Universe. They form the greatest structure of the Universe know to date: the Cosmic Web, a multi-scale three-dimensional spiderweb-like shape that spreads across the entire universe.

Four features are generally distinguished. Clusters are the most dense regions, where hundreds of galaxies can be found. They are the nodes of the Cosmic Web. Filaments are elongated structures, often linking a node to another or flowing into another filament. Walls are like filaments, but flatter. Finally, voids are big empty bubbles lying between the rest of the structure.

Galaxies are in movement in the Cosmic Web, interacting almost exclusively through gravitation. This movement is extremely slow relative to typical intergalactic distances: a few hundreds kilometers per second – to be compared to the 10^{22} kilometers to our closest neighbor galaxy, for instance. As a result, only a small fraction of galaxies are merging, interacting or colliding. The motion of galaxies through the cosmic web can be very complex and including laminal flows, vortical flows, and chaotic motion.

The Cosmological Microwave Background

When pointing a radio telescope in any direction of the sky, even in the emptiest regions, there is still a very faint signal: The Cosmological Microwave Background (CMB).

Since the speed of light is finite the further away an object is from us, the longer it has taken the light to reach us and thus the younger the Universe was wehn the light was emitted. The CMB is an image of the Universe 13.7 billions light years away from us about 300 000 years after the Big Bang. The signal of the CMB is the first light signal that could ever travel through space since before recombination, photons Thompson scatter off free electrons and the Universe was opaque. What is "behind" the CMB

will therefore forever stay hidden to us using photons; there may be a cosmic Neutrino background that could in theory be detected.

The CMB is the best source of information of the very early Universe we have. We learn from it thatthe Universe was then extremely but not perfectly homogeneous with variations on the order of the millionth of it mean density. We believe these minuscule inhomogeneities are the seeds of the gravitational instabilities that later led to the formation of the Cosmic Web, void regions, clusters, etc.

The sphere of the CMB is however so far and so huge, that the part of the Universe we see in the CMB gives no information about the Universe close to us. Only statistical information can be extracted from it, like its "granularity". We expect our local Universe to have the same statistical properties as this very distant shell. This is gives us some keys to understand the formation of the structures around us.

A Universe in expansion

Observations show unequivocally that the Universe is expanding on very large scales. This means that the underlying frame of the Universe is growing itself. Two objects, locally fixed in their environment, and that are not interacting, are constantly moving away from one another.

According to the Hubble's law, the speed with which two points in the universe recede from each other is proportional to their distances. Edwin Hubble first discovered this phenomenon in the late 1920s and with it the discovery of the extension of the Universe. This growth seems to be the same everywhere, meaning it is a homogeneous and isotropic expansion. Thus objects relatively close to the us recede slowly while objects at cosmological distances recede with speeds approaching he speed of light.

Measuring distances and redshifts

Measuring the distances of galaxies is very difficult. Because of the movement of the Earth around the sun, the direction of any objects (star, galaxy, etc) changes a bit in the sky. The maximum of variation is when the observations are made six months apart. Thanks to a few lines of plain trigonometry, and given that we know the radius of the orbit of the Earth around the sun, the distance to this object can be computed. This method, known as parallax can be used to find the distance to nearby stars.

However, as the distance to extra galactic objects is many orders of magnitude larger than the size of the orbit of the Earth, this variation is too small to be measured by our telescopes. We thus have toresort to other methods to estimate the distances of galaxies.

The first solution is to use the redshift as proxy for the distances, since according to Hubble's law distance and velocity are simply proportional to each other. Galaxies that should appear white in the sky seem in fact yellow, dark orange or even red. This effect due to the shifting of spectral emission lines towards the red end of the spectrum is, aptly, called the redshift and can be interpreted as a Doppler effect. Just as with sound, as a loud object moves away from us its pitch drops as the sound wave frequency decreases due to the added effect of the recessional motion. As the Universe expands, distant galaxies move away from us, leading to an apparent redshift. This relation can be inverted, and the distance estimated from the observed redshift.

The second solution is to estimate the true brightness of the galaxy (or of a star in the galaxy) and compare this with the observed brightness. Indeed, knowing how bright a galaxy *is*, it is possible to reconstruct how far it has to be in order to explain the loss of brightness observed from the Earth. The estimations of distances resulting from this approach often suffer from errors, the extent to which depends on the accuracy and precision of the method used to estimate the intrinsic, or absolute or "true" brightness.

The problems of dark matter and dark energy: ΛCDM

Not only do observations show that the Universe is growing, but also that the rate of growth is speeding up: the expansion is accelerating. This means that the speed at which two independent objects at a fixed distance are moving away from one another increases with time. Fundamental physics associates an energy to this accelerating force. As there is obvious source for that energy – it is virtually "invisible", it earned the name of "dark energy". However many ideas exist, including the concept that dark energy may simply be the vacuum energy, although no particle has been found yet to support one theory of another. In the equations of general relativity, there is a constant, famously introduced by Einstein to prevent the solutions to the field equations from being dynamically unstable. Einstein termed this constant a cosmological constant called Λ, and famously called it his "biggest blunder" since he could have predicted the expansion of the universe without it. Dark energy makes up around 70% of the total

energy of the Universe, that is to say a large majority of it. Understanding dark energy is thus quite central to understanding the Universe.

The problem of dark matter is more local. Let's consider a satellite orbiting around a planet. Kepler's laws of planetary motion, as explained by Newton's theory of gravity (the current model in use today) states that the higher the mass of the planet, the faster a satellite at a a given distance will orbit. Thus a satellite at a given distance from the Moon orbits slower than it would be if it was orbiting at that same distance around the Earth. Using the same equations and observing the rotation curve of galaxies, we can estimate their mass. However, it appears that the gravitational mass is much greater than the luminous mass (gas, stars, etc). An equivalent problem is found for cluster of galaxies: they are orders of magnitude more massive than their inferred from their emitted light. The missing mass is supposed to come from the invisible "dark matter", which composes not less than 80% of the total amount of gravitationally active matter. Many non-baryonic particles have been proposed to be dark matter. If none of the candidates has been discovered yet, constraints points towards a *cold* particle, that is to say, a particle which has a non relativistic energy when it decouples from the radiation field. As such it decouples later and is more massive than say hot dark matter (*i.e.* neutrinos), hence the name, Cold Dark Matter (CDM).

Combining these two important hypothesis leads to the ΛCDM model, which is the current paradigm. The entirety of this work in done within this model. The matter and the energy we know and understand well constitutes only a few percents of the total amount found in the Universe.

This work

Both baryonic matter (that we know) and non-baryonic dark matter (that we do not understand) interact gravitationally. However, we can only directly observe galaxies, made of baryonic matter, which makes up for less than 20% of the total amount of matter.

As we have just seen, placing galaxies in space is not trivial, as measuring their distances is hard. It is nevertheless doable, and the (approximate) distances of millions of galaxies are known. We are therefore able to make maps of the galaxy distribution. On the other hand, mapping the distribution of dark matter is not directly possible. This is because, the relationship between the distributions of baryonic and dark matter is not yet understood – the so-called galaxy "bias" implies that galaxies are biased tracers of the dark matter distribution. Thus the presence of baryons does not necesarilly indicate the presence of a proportional amount of dark matter.

The present work aims at reconstructing the distribution of both baryonic and dark matter in theLocal Universe (up to about one billion light years). This is done by first reconstructing the velocity field: the map of the direction and velocity of each and every point of the Universe. In order to constrain this field, we use measurements of peculiar velocities of galaxies: given a measure of redshift and an independent measure of distance for a given galaxy, it is possible to estimate its peculiar velocity, *i.e.* the component of the galaxy's velocity which is due to the gravitational forces as opposed to the component due to the expansion of the Universe. As galaxies interact with all the matter (baryonic and dark) by gravitation, their peculiar velocities are dictated by the distribution of matter. Under a certain number of assumption this relation can be inverted, and a map of the distribution of the matter in the Local Universe can be inferred from the velocities of the galaxies.

This work is not the first to attack this problem. Its originality lies in the method developed and the strategy of approach. The traditional way to reconstruct the density field from the velocity field is by trying to correct for the imperfections and uncertainties of the data before applying a reconstruction algorithm that retrieves the distributions of interest. Here, we propose a new algorithm that reconstructs the distributions and corrects the data at once, in a self consistent manner. Like most advances in science, this thesis is built on the work of others, among them a similar algorithm put forth by Romain Graziani during his PhD (2015-2018). However string differences exist between his proposed method and the one presented here, differences which, as it turns out, are system-critical for the problem at hand

In this work we tested our idea on fake observations taken from a simulated universe in order to see how accurate it is before applying it to the real universe, whose matter distribution is of course unknown. We also compared how it fares against another completely independent method where the correction of the data and reconstruction of the plans are split. This enabled us to quantify the accuracy and the limits of our techniques. Finally, in the last chapter, we applied our method to the last data release of the Cosmicflow-4 catalogue and present a novel insight on the distribution of dark matter in the Local Universe.

Chapter 1

Introduction

The Universe is filled with matter in a magnificent variety of states, temperatures and composition. From the emptiness and vastness of the intergalactic vacuum to the infinite density in the centers of Black Holes. Physics as a discipline is dedicated to the study of matter and energy in the universe. This thesis is focused on one specific scale in the vast spectrum of scales available to the natural scientist: Galaxies and the large scale distribution of matter. Since, as will be described below, galactic light traces the full matter distribution in the Universe, this thesis begins its journey by examining galaxies: what they are, how we measure their properties and how we employ them as tools to characterize the nature of of our Universe.

1.1 Observing Galaxies

Galaxies are – with some exceptions – the only constraints we have on the large scale of the Universe. After a brief definition of what a galaxy is, this section goes into detail of how to recover the necessary data needed for our work, namely the distance and the redshift.

1.1.1 What are galaxies?

In the vaguest sense galaxies are relatively dense agglomerations of stars, gas, and dark matter. Galaxies have a variety of shapes and sizes, colors and magnitudes, masses and kinematics. The earliest classification of galaxies - the so called "tuning fork" is due to Hubble (1926). Although there is intense debate in the community regarding what exactly constitutes a galaxy (*e.g.* Forbes & Kroupa, 2011) and how – for example, these differ from star clusters, perhaps one their most important defining quantities is that the are bright enough to be seen clearly across the Universe. Even as this thesis is being written some studies claim that the James Webb Space Telescope has spied galaxies so far away that their light was emitted when the Universe was a mere 100 million years old (Naidu et al., 2022).

Thanks to theirs stars, galaxies (and objects therein or thereabout) are the most commonly observed objects in the cosmos that are emitting in the optical wave band. They have been the first *cosmological* objects known by humankind, imagined by Kant (1755) under the terms of "islands Universes" and whose scientific existence was the subject of the so-called "great debate" of 1921 that opposed between H. Curtis and H. Shapley (*e.g.* Smith, 1982). The hot and cold gas of galaxies can be observed at radio, infrared, ultraviolet or even X-ray wavelengths. Super-massive black holes have been directly and indirectly observed at the centers of most galaxies (*e.g.* Fabian, 2012; Event Horizon Telescope Collaboration et al., 2019).

A large quantity of dust (as in small particles of matter) are found in the interstellar regions of most galaxies. Cosmic dust is the product of the stellar life cycle (*e.g.* Weingartner & Draine, 2001; Draine, 2003). The dust of our own galaxy obstructs our view and makes it virtually impossible to observe through the disk of our galaxy, creating what is called a Zone of Avoidance (ZoA): a region of $\sim 20\,\mathrm{deg}$ around the plane of the Milky Way in which little to no galaxies can be observed (Hubble, 1929). This is because cosmic dust is opaque to most star light, absorbing it and re-emitting it as heat in the infra-read (*e.g.* Draine, 2003).

1.1.2 Measuring distances

Measuring the distances of galaxies is extremely hard. Considering a generic nebula on the sky, the question comes down to not knowing if it is big bright and far or if it is small, dim but close. Direct measures such as parallax or measuring the proper motion of an object against a fixed background is, in general, not applicable to extra galactic sources[1]. Therefore other indirect methods must be used.

Distance moduli

With the exception of proper motion measurements, which are available only in the local group (*e.g.* Sohn et al., 2013), all methods for determining extra galactic distances relies on comparing the *observed* luminosity of an object (also called apparent luminosity) and its *estimated* true luminosity (also called absolute or intrinsic luminosity)[2]. Since the true, absolute luminosity of a galaxy is not known, it must be modeled. If an object has a luminosity L, the flux f received at a distance d_L is

$$f = \frac{L}{4\pi {d_L}^2}. \tag{1.1}$$

The quantity d_L is called the *luminosity* distance, as it is derived from a discussion on the luminosity of objects, as seen in section 1.4.5, it is close to but is not the *true* distance of this galaxy.

Historically, astronomers prefer to use the logarithm of the flux: the magnitude[3]

$$m = -2.5 \log\left(\frac{f}{f_0}\right) = -2.5 \log\left(\frac{L}{4\pi {d_L}^2 f_0}\right) \tag{1.2}$$

where f_0 is a flux of reference (considered as a technical parameter for this work) for example from a reference star like Vega. The absolute magnitude is defined as the magnitude of an object *if* it was observed from 10 pc of distance, *i.e.* setting $d_L = 10\text{pc}$ [4]:

$$M = -2.5 \log\left(\frac{L}{4\pi 10^2 f_0}\right). \tag{1.3}$$

Taking the difference between the apparent and the absolute magnitude, we obtain the distance modulus:

$$\mu = m - M = 5 \log\left(\frac{d_L}{10\,\text{pc}}\right). \tag{1.4}$$

The apparent magnitude can be directly observed with great precision, but the absolute magnitude can only be indirectly estimated based on a model. The uncertainty on the intrinsic magnitude M can thus be directly reported on the distance modulus μ: $\sigma_\mu \equiv \sigma_M$. In other words, a poor estimator of the intrinsic magnitude yields a poor estimator of the distance. An extensive discussion on the uncertainties of distances measurements and their implications is presented in section 1.5.

In this work, we exclusively write distances in Mpc, we thus rewrite eq. (1.4) to

$$\mu = 5 \log\left(\frac{d_L}{1\,\text{Mpc}}\right) + 25. \tag{1.5}$$

Standard candles and scaling relationships

Objects whose true intrinsic brightness can be deduced from other properties are called standard candles. When a galaxy is the host of such an object, its distance can be relatively precisely observed. The standard candles most used in this work are:

[1] By measuring the position of some Local Group galaxies (the nearest ones) a decade or more apart, the proper motion of these objects can be measured by aligning the background high redshift quasars. At the moment this is not possible beyond 1 Mpc (*e.g.* Sohn et al., 2013)

[2] A recent new technique makes use of gravitational wave sirens. However this is still too recent to be suitable for this work

[3] The first classification of stars by magnitude was done by the ancient Greeks with the naked eye. The mathematical definition of the magnitude is based on this classification.

[4] This value is arbitrary and is in fact much smaller than the size of any galaxy.

— **Cepheid stars.** Due to the so-called κ-mechanism (an oscillating obstruction of the light by the ionization of the rejected Helium), their brightness varies in time with a period that can directly be linked to their true (average) brightness. Cepheids are the first observed standard candles (Leavitt & Pickering, 1912) and were used by Hubble (1929) to demonstrate that the Andromeda's nebula (a galaxy in fact) is not part of the Milky Way. They are used up to 10 Mpc and yield measurements of M with an uncertainty $\sigma_M \sim 0.1$ (*e.g.* Riess et al., 2018).
— **Tip of the red-giant branch (TRGB).** When a Sun-like star approaches the end of its life, its luminosity increases while its temperature decreases until a so-called "helium flash" (Deupree & Wallace, 1987), after which it virtually retraces its steps in terms of temperature and luminosity. When plotted on a Hertzsprung-Russel diagram, red-giant stars constitute a prominent branch-like feature whose tip is very well defined. Because of nature of the process, any start undergoing a helium-flash has the same luminosity Ferrarese et al. (2000). Thus, stars that form the "Tip of the red-giant branch" (TRGB) are expected to have the same luminosity in every galaxy: this feature of the Hertzsprung-Russell diagram can be used as a standard candle. Technically, every galaxy is the host of red-giant stars, but for these to be observed, the galaxy needs to be close enough so that its stellar population can be resolved (*i.e.* stars can be separately observed). This method is used up to 10 Mpc and yields measurements of M with an uncertainty $\sigma_M \sim 0.1$ (*e.g.* Lee et al., 1993; Riess et al., 2018).
— **Super Novae Ia (SNIa).** In the very rare situation where a star and a white dwarf orbit together, at end of its life, the white dwarf collapses in an gigantic explosion whose luminosity rivals the entire host galaxy.The Nova is a flash followed by a gradual decrease in magnitude. The speed with which the nova fades can be charactersized in a "light curve". It turns out that the shape of the light curve is directly related with the peak brightness of the nova. Therefore by timing the rat rate at which the nova fades, it is possible to estimate the true brightness of the super nova. Although they are very rare events (rouhgly one per galaxy per century in the late universe), they can be used until ~ 4000 Mpc and yield measurements of M with an uncertainty $\sigma_M \sim 0.14$ (*e.g.* Betoule et al., 2014).

Standard candles are not the only way distances from galaxies can be estimated. Other methods, the *scaling methods*, although less precise, can also be used. Scaling relations are often empirically driven "laws" which correlate bulk properties of a galaxy with their intrinsic magnitude.

— **Tully-Fisher relation (TF).** As can be naively derived from Newton's gravitational law, the more massive a spiral galaxy is, the faster it rotates. Meanwhile, the more massive it is, the more stars are formed. Building on this idea, Tully & Fisher (1977) showed through observations that there is a direct (but noisy) relation between the asymptotic rotation velocity of a spiral galaxy and its intrinsic luminosity. This method can be widely applied to many galaxies. It yields measurements of M with an uncertainty $\sigma_M \sim 0.4$ (*e.g.* Kourkchi et al., 2020; Tully et al., 2023).
— **Fundamental Plane / Faber-Jackson relation (FP).** Faber & Jackson (1976) found another relation of the same type of the Tully-Fisher relation, but for elliptical galaxies. Since elliptical galaxies are supported by isotropy – stars orbit the galactic center on all possible orbits, the velocity dispersion of the stellar spheroid is correlated with the diameter of the galaxy. The size is in turn correlated with the mass. Faber & Jackson (1976) thus linked the Intrinsic luminosity of elliptical galaxies their stellar velocity dispersion. Further work showed these quantities (stellar velocity dispersion and Luminosity) to also correlate with effective radius, impling the existence of a "fundamental plane" which all elliptical galaxies sit on in stellar velotiy dispersion - Luminosity-effective radius space Thus measuring either the velocity dispersion or the effective radius give an estimate for the intrinsic brighness. It yields measurements of M with an uncertainty $\sigma_M. \sim 0.4$ (*e.g.* Djorgovski & Davis, 1987).
— **Surface Brightness Fluctuation.** The surface (as seen from the Earth) of a galaxy is not perfectly smooth, there is some granularity due to the inhomogeneous distribution of stars within the galaxy. When looking at the image of a galaxy on a CCD captor, as found in every modern telescope, there is thus a random fluctuation of the number of photons that landed on each pixel. The farther away the galaxy is, the smoother it seems at a given pixel resolution, and the smaller the amplitude of this fluctuation is. Thus, once the method has been properly calibrated, and with some modeling of the galaxy, the distance of a galaxy can be estimated from its measured *surface brightness fluctuation*. This method yields measurements of M with an uncertainty $\sigma_M \sim 0.4$ (Tonry, 1997; Blakeslee et al., 1999; Tonry et al., 2000, 2001).

1.1.3 Measuring redshifts

When observing a distant galaxy well known sets of emission or absorption lines are routinely shifted towards the red end of the spectrum. For example, the series of wavelength emitted by a Hydrogen atom when electrons fall into the ground state, known as the Lyman Series, has a well measured wavelengths for each transitions. Yet in distant objects the Lyman Series (as well as other series) are routinely shifted toards the red end of the spectrum from where they are measured on earth. This effect is called the redshift and is defined as

$$z_{\text{obs}} = \frac{\Delta\lambda}{\lambda_{\text{source}}} \tag{1.6}$$

where λ_{source} is the known (or estimated) wavelength of the source and $\Delta\lambda = \lambda_{\text{obs}} - \lambda_{\text{source}}$ is the difference between the observed and expected spectral line. Note that if the redshift is negative, we refer to a "blue shift". As this effect is comparable to a Doppler effect[5], the redshift is very often transformed into a velocity by a simple multiplication by the speed of light $v_{\text{rec}} = cz_{\text{obs}}$, called the *recession velocity*. This velocity is not necessarily physical, as it can exceed the speed of light. By abuse of notation, redshifts are thus often given in km/s.

There are two methods to measure the redshift in use today:

— **Spectroscopy.** Once the spectrum of a galaxy has been obtained, emission or absorption lines in it are compared to their expected value. Instead of one couple λ_{obs} and λ_{source}, several are used, yielding a quite precise results with an error of only 50 km/s. Yet, obtaining the spectrum is time consuming, and thus this method only applies to a subset of the observed galaxies.

— **Photometry.** Using two pictures of a galaxy where two different colored filters have been put in the optic of the telescope (usually red and blue), it is possible to estimate its redshift. This technique allows for many galaxies to be observed at once but is much less precise and has errors up to about $\sigma_{cz}/cz \approx 50\%$, which reaches to 15 000 km/s for the range of redshifts of matter to this work (Bolzonella et al., 2000).

1.1.4 Hubble's Law

In 1929, shortly after the extra-galactic nature of nebulae was confirmed[6] by Hubble (1926), Hubble demonstrated that there is a direct linear relation between the redshift of galaxies and their distances (measured with Cepheid stars Hubble, 1929). This relation is called the Hubble or Hubble-LeMaitre Law and reads

$$cz = H_0 d \tag{1.7}$$

where H_0 is known as Hubble's constant. This law has been confirmed in the Local Universe by later measurements (*e.g.* Riess et al., 2018), however, the value of Hubble's constant is the source of a controversy. Estimations of H_0 from the CMB are found around $H_0 = 67.8 \pm 0.9$ km/s/Mpc (Planck Collaboration et al., 2016), which is much lower than estimations from the local universe, *e.g.* $H_0 = 74.6 \pm 0.8$ km/s/Mpc (Tully et al., 2023). If the redshift is interpreted as a Doppler effect, this law means that a object at a distance d recedes with a velocity $H_0 d$.

1.2 The homogeneous Universe

In this section, we review the theoretical bases on which this work – and indeed modern cosmology – is built. We start by a general description of the dynamics of a homogeneous, isotropic universe.

1.2.1 The Cosmological principle

Modern cosmology relies on a fundamental hypothesis, the Cosmological principle, namely the assumption that *the Universe is homogeneous and isotropic on large scales*. The Cosmological principle is a

[5] The Doppler effect is the shift of the frequency perceived by a fixed observer with respect to a moving emitter. Acoustically this is simply a drop in pitch, but it can also be applied to light sources as well, both being waves. However it is noted that the redshift has various interpretations including the stretching of wavelengths due to metric expasion of space. see https://arxiv.org/abs/1605.08634

[6] The term nebula was then, used to designate any extended source of light, which notably included galaxies.

generalization of the Copernican principle stating that *we, on Earth, are not privileged observers of the Universe*. Such an hypothesis is motivated not just by the philosophy that it is highly unlikely that we inhabit a special place or time in the Universe, but also from observations that the universe indeed appears homogeneous and isotropic.

1.2.2 The glue of the Universe: gravitation

At great distances, objects interact primarily through gravitation. According to the Newtonian interpretation, gravity is a force that every massive object applies to every other massive object. It is an intrinsic property of all mass, regardless of form. The force exerted by two masses on each other is proportional to the mass of each object m_a and m_b and inversely proportional to the square of the distance between the objects r_{ab}:

$$F_{ab} = -G\frac{M_a M_b}{r_{ab}^2} \tag{1.8}$$

where $G = 6.674 \cdot 10^{-11}\,\mathrm{m^3 kg^{-1} s^{-2}}$ is the gravitational constant. This law as the value of G are empirical and have been at first designed by Newton to explain movements in the solar system.

1.2.3 The large scale dynamics of the Universe: Friedmann equations

The theory of modern cosmology starts with general relativity, which extends or replaces Newtonian gravity. In Newton's view of gravity, the Universe is the stage on which gravitating objects act. In general relativity, however space-time is considered to by a dynamical quantity that can change due to the presence of mass-energy. These changes are described by the Field equations of general relativity: (Einstein, 1914, 1915b,a; Einstein, 1916):

$$\underbrace{R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu}}_{\text{Curvature of space}} = \underbrace{\frac{8\pi G}{c^4}T_{\mu\nu}}_{\text{Content of space}} \tag{1.9}$$

where $g_{\mu\nu}$ is the metric tensor[7] from which one can compute the self contracted Riemann Tensor $R^{\lambda}_{\mu\lambda\nu} = R_{\mu\nu}$, known Ricci tensor, as well as the Ricci scalar R. Λ is the "cosmological constant" whose physical meaning is to be discussed in the next section. The energy-momentum tensor $T_{\mu\nu}$ describes the content space-time, and c is the velocity of light. A more precise understanding of this equation is not necessary for the purpose of this work. In the non-relativistic limit of weak fields and low velocities, general relativity, simplifies to the Newtonian description of gravitation.

The problem of finding solutions to eq. (1.9) was immediately posed after Einsteins seminal 1917 Fiel equation paper. After K Schwarzschild famously found the solution to the gravitational field around a point mass, Friedmann (1922, 1924) and Lemaître (1927, 1931) were the first to independently find solutions for a homogeneous isotropic distribution of matter, namely the solutions which describe the dynamics of the Universe as a whole. This is known as the Friedmann-LeMaiter-Robertson-Walker (FLRW) metric (Robertson, 1933, 1935, 1936a,b; Walker, 1937). The line element can be written in spherical coordinates (r, θ, ϕ):

$$d\tau^2 = c^2 dt^2 - a^2(t)\left(\frac{dr^2}{1-kr^2} + r^2 d\theta + r^2 \sin^2\theta d\phi^2\right) \tag{1.10}$$

where $a \in [0, 1]$ is a time dependent global scale factor of the Universe and k is its curvature[8]. This metric, which forms the basis of modern cosmology, describes the solution to the Einstein Field equations under the Cosmological principle. The resulting equations describe a homogeneous and isotropic universe of density ρ and internal pressure p whose evolution is described by the two independent so-called "Friedmann equations":

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\rho - \frac{kc^2}{a^2} + \frac{\Lambda c^2}{3}, \tag{1.11}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\rho + 3p) + \frac{\Lambda c^2}{3}. \tag{1.12}$$

[7] The core quantity that describes the "shape" ofspace-time.

[8] There are three possibilities here: either the Universe is lat ($k = 0$), spherical ($k > 0$), or hyperbolic ($k < 0$). Accordingly the sum of the internal angles in a triangle is either exactly 180°, greater than 180°, or less than 180° respectively.

A key characteristic of such an universe is its possible expansion (or collapse) with time. This was at first a pure mathematical conjecture when the equations of Friedmann were first written by Lemaitre, but was evidenced some years later by Hubble (1929) (see section 1.4.1 for an observational description of the Hubble Law). The time dependent expansion rate is named the Hubble parameter $H(t) = \dot{a}/a$ and converges at modern times to the so-called Hubble constant $H(0) = H_0$.

A choice of mean density, pressure and cosmological constant thus provides a description with how the Universe expands - a testable prediction of the theory. Among the predictions are that there exists a "critical density" ρ_c, namely the density needed to halt the expansion of the Universe.

$$\rho_c = \frac{3c^2 H_0^2}{8\pi G}. \tag{1.13}$$

ρ_c is in general a function of time (but can be defined today) and derived by solving eq. (1.11) assuming $k = 0$ and $\Lambda = 0$. In general, if $\rho > \rho_c$, the Universe has ample matter and thus positive curvature and will expand to a maximum extent, and then collapse under its own gravity. If $\rho < \rho_c$ there is insufficient matter to halt the expansion, the universe has negative curvature and expands eternally. Lastly, if $\rho = \rho_c$ the universe is flat, and there is just enough matter to balance the expansion which slows down, never reversing. Measuring the rate of expansion or the rate of deceleration (or the rate of change of the expansion) is thus a kin to measuring its matter content.

Note that the growth factor is often substituted to the cosmological time (*i.e.* time dependent functions are given as functions of a instead of t). These two quantities are bound by

$$t(a) = \int_0^a \frac{\mathrm{d}a}{\dot{a}}. \tag{1.14}$$

1.2.4 The content of the ΛCDM Universe

Such a universe, although homogeneous and isotropic, can have several constituents i such that $\sum \rho_i = \rho$. We define $\Omega_i = \rho_i/\rho_c$ which is preferred over the ρ_i, such that $\sum \Omega_i = 1$ in a flat Universe. Several models have been proposed to describe how the Friedmann equations evolve and thus the history and fate of the universe (*e.g.* Einstein & de Sitter, 1932). The model developed over the last decades to explain the observations is named ΛCDM ("Lambda Cold Dark Matter"). It comprises four components.

— **Baryonic Matter.** Baryonic matter is what the common use of "matter" implies[9] It has been well observed, on Earth and in the Universe. It composes all atomic matters including the gas and the stars that make up galaxies. It constitutes only about 5% of the total content of the Universe ($\Omega_b = 0.04859$; Planck Collaboration et al., 2016).

— **Cold Dark Matter.** Dark matter was introduced in the 1980s to explain the failure of newtonian gravitation (see eq. (1.8)) on large scales. Indeed, the amount of baryonic matter we detect in galaxies does not explain their rotation curves, which instead of dropping off at great distances from their centers appear flat: stars at a galaxy's edge have the same circular speed as stars much closer to the center of the galaxy. Similarly, galaxies in clusters move too fast, compared to the amount of matter inferred by the observed light (Zwicky, 1937; Rubin et al., 1980) In other words the typical speeds of galaxies in clusters is higher than the escape velocity inferred by mass estimates based on the visible light (*i.e.* number of galaxies). Under the assumption that these objects are not flying apart, the sensible conclusion is that the escape velocity is much higher than that predicted by simply counting the galaxies. There thus must be a non-luminous component to these clusters that binds the cluster members and prevents them from flying apart. No dark matter particle has been directly detected, despite many detection experiments. Constraints, coming from observations and cosmological simulations, show that dark matter has to be cold or at least warm and that it cannot be hot, namely relativistic at decoupling[10]. The main evidence for this is due to a phenomenon known as "free streaming" (Blumenthal et al., 1984; Davis et al., 1985) In brief: if dark matter was very hot (*i.e.* fast) then it would be able to easily escape the potential wells that confined it in the early Universe. The shallow potential wells would be erased where as deeper ones could survive. As such the minimum galaxy mass, namely the smallest potential well that

[9] We note that the term "Baryonic Matter" is a misnomer since it also refers to Hadronic matter such as leptons like the electron or mesons like the pion. However, cosmologists somewhat incorrectly refer only to baryons since these are many orders of magnitude more massive than leptons and mesons.

[10] The "temperature" of the matter has to be understood under the eye of statistical physics and refers to the velocity of the particles.

can survive free-streaming, sets a limit on the nature of the dark matter particle (at decoupling). Given the existence of structures on smaller and smaller scale the free streaming process appears to have been unimportant and thus dark matter must have been cold (*i.e.* slow). Hence the name: Cold Dark Matter (CDM). Dark matter makes up for 25% of the total content of the Universe ($\Omega_{\rm DM} = 0.2603$; Planck Collaboration et al., 2016).

— **Dark Energy.** Dark energy, or energy of the void is the source of the accelerated expansion of the Universe. In the Einstein and Friedmann equations, it is represented by the Λ, the cosmological constant[11]. There are many suggestions as to what Λ could be including the vacuum zero point energy, quintessence, dynamical dark energy among others. It is the largest constituent of Universe with $\approx 70\%$ of the total ($\Omega_\Lambda = 0.6911$; Planck Collaboration et al., 2016).

— **Curvature.** Even though curvature is not really a "constituent" of the Universe, it is treated similarly in the Friedmann equations 1.11 and 1.12. Modern measurements indicate that our Universe is flat ($\Omega_k = 8\,10^{-4}$; Planck Collaboration et al., 2016).

— **Radiation.** Due to how radiation and baryons loses energy as the Universe expands and their number density drops, radiation can be neglected in the matter dominated era, *i.e.* $\Omega_r = 0$.

Usually all matter terms are grouped together such that $\Omega_m = \Omega_b + \Omega_{\rm DM}$. Planck Collaboration et al. (2016) gives $\Omega_m = 0.3085$. The acceleration of the growth factor can be written as a function of the constituents of the Universe:

$$\dot{a} = H_0 \sqrt{\Omega_m \left(\frac{1}{a} - 1\right) + \Omega_\Lambda (a^2 - 1) + 1}. \tag{1.15}$$

Similarly as the universe expands, the self gravity of all the gravitating mass in it, decelerates the expansion leading to a deceleration parameter, which today is

$$q_0 = \frac{1}{2}\Omega_m - \Omega_\Lambda \tag{1.16}$$

Measuring the deceleration parameter can thus constrain the amount of matter and dark energy in the Universe.

1.3 The Large Scale Structure

Galaxies form a large structure named the Cosmic Web (Bond et al., 1996), or less originally, the Large Scale Structure (LSS) of the Universe, which was first discovered in the 1980s by the CfA survey (de Lapparent et al., 1986). Four main types of features are found in the LSS: void regions, flat walls of galaxies, elongated filaments and clusters; the filaments and the walls form an irregular web where clusters are found at each intersections (*e.g.* Zeldovich, 1978; Bond et al., 1996). This structure is the largest observed today (hence the name, *the* Large Scale Structure). In fact the transition from cosmic web to homogeniety has yet to be found and estimates on the "scale of homogeniety" differ from 70 (Scrimgeour et al., 2012) to hundreds of Mpc (Peebles, 1980)

1.3.1 The local dynamics of the Universe: structure formation

Even if the Universe is homogeneous and isotropic on the largest scales, it displays rich structures on small(er) scales, from the size of a star to the so-called Cosmic Web, or Large Scale Structure (LSS), a structure coherent over hundred of millions of light years. These small scale inhomogeneities are often characterized by means of density perturbations, or more generally, of a density field. The evolution of this density field constitutes one of the main problems in cosmology and, within the context of the hot big bang, it is determined by assuming the universe is an ideal fluid that behaves according to well known conservation laws. We define the comoving cosmological reference frame in which the expansion of the universe is omitted. Namely two points in the universe whose only motion is due to the expansion, remain at constant comoving distance from each other:

$$\boldsymbol{x} = \boldsymbol{r}/a. \tag{1.17}$$

[11] It was first introduced by Einstein himself to prevent the Universe from collapsing on itself under the effect of self gravitation, but was then abandoned as Hubble proved the expansion of the Universe. It was reintroduced in the equation long after his death.

where $\boldsymbol{x}$ is the comoving coordinate and $\boldsymbol{r}$ is the physical or proper coordinate. Thus the dynamics of the density field are governed by:

$$\frac{\partial \rho_r}{\partial t} + \boldsymbol{\nabla}_{\boldsymbol{r}} \cdot (\rho_r \boldsymbol{v}) = 0 \qquad \text{Conservation of matter,} \tag{1.18}$$

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \boldsymbol{\nabla}_{\boldsymbol{r}})\boldsymbol{v} = -\frac{\boldsymbol{\nabla}_{\boldsymbol{r}} p}{\rho_r} - \boldsymbol{\nabla}_{\boldsymbol{r}} \Phi \qquad \text{Euler's equation,} \tag{1.19}$$

$$\Delta_{\boldsymbol{r}} \Phi = 4\pi G \rho_r \qquad \text{Poisson's equation.} \tag{1.20}$$

where the subscript r refers to the physical, proper reference frame. In the comoving frame the space and time derivatives as well as the density field are modified:

$$\nabla \equiv \nabla_{[\boldsymbol{x}]} = a \nabla_{\boldsymbol{r}}, \tag{1.21}$$

$$\boldsymbol{v} \equiv \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \dot{a} r + \boldsymbol{u} \quad \text{with} \quad \boldsymbol{u} \equiv a\dot{\boldsymbol{x}}, \tag{1.22}$$

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} + \frac{\dot{a}}{a}\boldsymbol{u} + \ddot{a}\boldsymbol{x}, \tag{1.23}$$

$$\rho = a^3 \rho_r. \tag{1.24}$$

Equations 1.18 then become

$$\frac{\partial \rho}{\partial t} + \frac{1}{a}\boldsymbol{\nabla} \cdot (\rho \boldsymbol{u}) + \frac{3\dot{a}}{a}\rho = 0, \tag{1.25}$$

$$\frac{\partial \boldsymbol{u}}{\partial t} + \frac{1}{a}(\boldsymbol{u} \cdot \boldsymbol{\nabla})\boldsymbol{u} = -\boldsymbol{\nabla}\Phi, \tag{1.26}$$

$$\Delta \Phi = 4\pi G a^2 \rho. \tag{1.27}$$

Even though they can be written, these equations can not be analytically solved. The approach adopted depends on the scales on which the formation of structures is studied. For the analysis of the Universe two methods are applied. The first one is the simplification of these equations by means of a linear Taylor expansion to first order, the subject of the next section. The second is the use of computers to numerically discretize the continuous field and solve the N-body problem as described in section 1.7.

1.3.2 Linear Theory

The linear over-density field

Although the density field is a function of cosmic time t and location $\boldsymbol{x}$, for convenience we may define the over density field $\delta(\boldsymbol{x}, t)$ in terms of the mean density $\bar{\rho}(t)$, since this is just a function of time

$$\rho(\boldsymbol{x}, t) = \bar{\rho}(t)(1 + \delta(\boldsymbol{x}, t)). \tag{1.28}$$

The over-density $\delta(\boldsymbol{x}, t)$ describes the local relative variation with respect to the mean. The over-density, a function of both space and time, is the core quantity of the linear theory of structure formation. In the linear regime, deviations from the mean density field are considered small ($\delta \ll 1$) such that a first order approximation is valid. Similarly the velocity field is assumed to be slowly varying $\boldsymbol{\nabla} \cdot \boldsymbol{u} \ll 1$. These approximations are the key of the linear theory.

This separation affects the gravitational potential Φ which splits into a mean potential $\bar{\phi}$ such that $\Delta\bar{\phi} = 4\pi G a^2 \bar{\rho}$ and a perturbed potential $\Delta\phi = 4\pi G a^2 \bar{\rho}\delta$. In the linear regime the eqs. (1.25) to (1.27) become equations which govern the evolution of a small over density:

$$\frac{\partial \delta}{\partial t} + \frac{1}{a}\boldsymbol{\nabla}\boldsymbol{u} = 0, \tag{1.29}$$

$$\dot{\boldsymbol{u}} + \frac{\dot{a}}{a}\boldsymbol{u} = -\boldsymbol{\nabla}\phi/a, \tag{1.30}$$

$$\Delta\phi = \frac{3\dot{a}}{2a}\Omega_m \delta. \tag{1.31}$$

Combining these equations together, one single equation that governs the dynamics of the over-density field may be obtained (Bonnor, 1957; Strauss & Willick, 1995):

$$\ddot{\delta} + \frac{2\dot{a}}{a}\dot{\delta} = \frac{3\dot{a}}{2a}\Omega_m \delta. \tag{1.32}$$

Another simplifying assumption that may be made is that the spatial and the time components of the over-density fields may be separated:

$$\delta(\boldsymbol{x}, t) = D(a)\delta_0(\boldsymbol{x}). \tag{1.33}$$

In this model, the over-density field is frozen in space, but can change over time. Inserting eq. (1.33) into eq. (1.32), two solutions for $D(a)$ emerge (Heath, 1977; Peebles, 1980):

$$D_+(a) \propto \frac{\dot{a}}{a}\int_0^a \frac{\mathrm{d}a}{\dot{a}^3}, \qquad D_-(a) \propto \frac{\dot{a}}{a}. \tag{1.34}$$

If the value of $D_+(a)$ and $D_-(a)$ are not analytical, they are easily numerically computed. Both are smooth, but D_+ increases with a while D_- decreases with increasing a. Therefor in today's Universe, $D_-(a)$ does thus not contribute anymore, and the solution can be simplified to include just D_+, also known as the growing mode.

$$\delta(\boldsymbol{x}, t) = D_+(t)\delta_0(\boldsymbol{x}). \tag{1.35}$$

The linear velocity

In this section we examine the evolution of the linear velocity field. It follows from eq. (1.29) that

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = \frac{\partial\delta}{\partial t} = -\dot{D}\delta_0 = -\frac{\dot{D}}{D}\delta = -\frac{\mathrm{d}\log D}{\mathrm{d}a}\dot{a}\delta \tag{1.36}$$

$$= -\frac{H}{a}f\delta \tag{1.37}$$

where $f = \frac{\mathrm{d}\log D}{\mathrm{d}a}$ is the linear growth factor.

In the modern Universe, this relation simplifies to

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = -H_0 f\delta. \tag{1.38}$$

and if the universe is flat universe by a function of Ω_m only (Lahav et al., 1991):

$$f(\Omega_m) = \Omega_m^{4/7} + \frac{1}{70}(1-\Omega_m)\times(1+\Omega_m/2). \tag{1.39}$$

Note that the dependence of this parameter in Ω_Λ is implicit as $\Omega_\Lambda = 1-\Omega_m$ in these conditions. Note that this equation is the same as an equation of conservation of the matter. The flow is simply pointing towards high density regions, and away from empty regions. The linear velocity is a potential flow and thus irrotational, *i.e.* $\boldsymbol{\nabla}\wedge\boldsymbol{u} = 0$.

Limits of the linear theory

The increasing contrast predicted by eq. (1.35) is a simplistic model of gravitational collapse. Indeed, the dense regions become denser and the empty regions emptier, as one would expect. Yet, the shape of the structure itself is frozen and just grow "in place", which is not realistic.

On one hand, a key approximation of the linear theory is the small contrast of the over-density field ($\delta \ll 1$). On the other hand, this same theory predicts that the contrast increases indefinitely over time. The first hypothesis is thus eventually broken. Worse, points where $\delta(\boldsymbol{x})$ is negative eventually reach $\delta(\boldsymbol{x}, t) < -1$, thus indicating regions where the density $\rho(\boldsymbol{x}, t)$ of the universe is negative, which is clearly unphysical. The linear theory can therefore only be applied at early times.

However, a key hypothesis of cosmology is that the Universe is homogeneous on large scales. The larger the scales considered are, the smaller the amplitude of the perturbations. A second realm in which the linear theory can therefore be applied is when studying the (very) large scales of the Universe.

The limits of the linear theory are discussed further in section 1.8 from the point of view of its statistics. In chapters 2 to 4, the linear theory is applied outside of its range of validity, in clear violation of the conclusions of this discussion.

1.3.3 Galaxy bias

Because of their supposed formation process, galaxies are thought to be found in very dense regions of the Universe (*e.g.*; Kaiser, 1984; Bardeen et al., 1986; Mo & White, 1996).

The exact relation between the position of galaxies and the density of the underlying distribution of matter (dark and luminous) is, until today, an unsolved question. It is extensively discussed in the literature under the term of *galaxy bias* (Bardeen et al., 1986; Mo & White, 1996; Peacock & Smith, 2000; McBride et al., 2011). Indeed, if the distribution of galaxies is relatively easy to estimate, the distribution of matter is more challenging, as discussed throughout this work. The most widespread model is a simple linear relation between the galaxy density and the matter density:

$$\rho_g = b\rho \tag{1.40}$$

where b is the linear galaxy bias factor, and ρ_g is the density of galaxies.

The shape, color and luminosity of galaxies have been shown to be linked to their environment (Dressler, 1980; Sheth & Diaferio, 2001; van der Wel et al., 2010). Furthermore, the population of galaxies evolves with time, *i.e.* distance. The galaxy bias is thus not a constant but rather a function of other parameters (Mo & White, 1996). Although the issue of the positioning of galaxies is necessary to understand this work, we do not assume any galaxy bias model.

1.4 Observations in the ΛCDM Universe

In this section, we draw the links between the observed quantities discussed in section 1.1 and the ΛCDM model of the Universe of sections 1.2 and 1.3.

The redshift of a galaxy spawns from several physical effects. The next three sub-sections treat ofthese effects separately. The fourth next section discusses how they combine into the observed redshift. The link between the luminosity distance of section 1.1.2 and the comoving distance introduced in section 1.3 is then detailed. Next, the estimation of peculiar velocities of galaxies from measurements of their redshifts and distances is explained. Finally, the use of the redshift as a proxy to the distance is discussed.

1.4.1 The cosmological redshift

The first source of redshift to consider is thus the expansion of the Universe. This component of the redshift is called the cosmological redshift. It emanates from the dissolution of the energy of the wave in space: the light was emitted when the Universe was smaller, therefore the energy that was then comprised in a small volume is now spread over a larger volume. In terms of light, the corresponds to a increase of the wavelength, and thus a redshift. The cosmological redshift is linked to the comobile distance of the object, and has to be computed with a model of the history of the expansion of the Universe (Davis & Scrimgeour, 2014). In a flat universe, the expression of $z_{\cos}$ reads

$$d = \frac{c}{H_0}\int_0^{z_{\cos}} \frac{\mathrm{d}z}{\sqrt{(1+z)^3\Omega_m + (1-\Omega_\Lambda)}} \tag{1.41}$$

where d is the comoving distance, $z_{\cos}$ the cosmological redshift. The reader will note that eq. (1.41) become $cz_{\cos} = H_0 d$ (see eq. (1.7)) when the integral is simplified to $z_{\cos}$. Equation (1.41) can not be analytically solved. It is however trivial to integrate numerically, which is the approach used in this work.

1.4.2 The peculiar redshift

The Hubble Law described in eq. (1.7) a small scatter around $cz = H_0 d$. This scatter is due to the *peculiar* velocities of galaxies. "Peculiar" is a misnomer as this velocity is simply the gravitational velocity, namely the component of a galaxy's motion due to the gravitational forces it feels. Whereas he Hubble Law traces the *global* expansion of the Universe, the peculiar velocity follows a galaxies *local* movement relative to the inertial frame of the CMB. Even though, the peculiar velocity of a galaxy is three-dimensional, its effect on the redshift is limited to the line of sight, namely the Doppler-Fizault effect (Davis & Scrimgeour, 2014)

$$z_{\mathrm{pec}} = \sqrt{\frac{1+v_r/c}{1-v_r/c}}, \qquad v_r = \boldsymbol{v}\cdot\hat{\boldsymbol{x}} \tag{1.42}$$

where $z_{\rm pec}$ is the peculiar redshift, v_r is the *radial* peculiar velocity, $\boldsymbol{v}$ is the full peculiar velocity and $\hat{\boldsymbol{x}} = \boldsymbol{x}/|\boldsymbol{x}|$ is the direction in the sky.

In the context of the linear theory $v_r \approx \pm 300\,\mathrm{km/s}$ (see eq. (1.95)). The dispersion of velocities is slightly larger in a gravitationally evolved universe but remains of the same order of magnitude (see section 1.8). For our application then, $v_r \ll c$ and the expression of the peculiar redshift $z_{\rm pec}$ can be reasonably simplified to its non relativistic form

$$z_{\rm pec} = v_r/c. \tag{1.43}$$

1.4.3 The observer's peculiar redshift

In the same sense that the movement of the observed galaxies has an effect on their measured redshift, the motion of the observer itself has the exact same effect. Thus, the combined motions of

1. the telescope on the Earth[12],
2. the Earth around the Sun,
3. the Sun around the center of the Milky Way,
4. and the Milky Way with respect to the inertial frame of the CMB

have to be taken into account (Calcino & Davis, 2017). Since observations of the CMB show a strong dipole in the distribution of microwave temperature's the peculiar redshift of the Earth can be easily assessed. These can then be transformed to the solar frame and then, via the "Local Standard of Rest", to the Galactic frame. Planck Collaboration et al. (2016) gives a peculiar velocity of the Sun in the Universe of

$$v_{\rm obs} = 369.82 \pm 0.11\,\mathrm{km/s}, \qquad l_{\rm gal} = 264.021^\circ \pm 0.011^\circ, \qquad b_{\rm gal} = 48.253^\circ \pm 0.005^\circ \tag{1.44}$$

where $l_{\rm gal}$ and $b_{\rm gal}$ are the galactic longitude and latitude in which $v_{\rm obs}$ is pointing. This velocity can be projected on the line of sight to any galaxy and removed from the measurement of redshift or radial peculiar velocity of each and every galaxy:

$$z_{\rm obs} = \boldsymbol{v}_{\rm obs} \cdot \hat{\boldsymbol{x}}. \tag{1.45}$$

The redshifts given are then with respect to the CMB. Unless mentioned otherwise, this work makes exclusive use of such corrected redshifts.

1.4.4 Combining redshifts

The observed redshift is not the direct sum of the different redshifts but rather reads (Davis & Scrimgeour, 2014)

$$1 + z = (1 + z_{\rm cos})(1 + z_{\rm pec})(1 + z_{\rm obs}) \qquad \text{for peculiar redshifts w.r.t the Sun,} \tag{1.46}$$

$$1 + z = (1 + z_{\rm cos})(1 + z_{\rm pec}) \qquad \text{for peculiar redshifts w.r.t the CMB.} \tag{1.47}$$

For the sake of calculus, this expression can be simplified at its first order ($z_{...} \ll 1$):

$$z = z_{\rm cos} + z_{\rm pec} + z_{\rm obs} + \mathcal{O}(z^2). \tag{1.48}$$

1.4.5 Luminosity distance and comoving distance

The peculiar and cosmological redshifts are critical and needed to obtain peculiar velocity, which in turn are needed to obtain distances. As mentioned above, in the ΛCDM cosmology the concept of distance is ambiguous since the expansion of the universe affects light, angular sizes and proper motions differently. In this thesis we concentrate exclusively on luminosity distances. The distance required to measure eq. (1.50) is the comoving distance but what is discussed in section 1.1.2 is the *luminosity* distance. These differ due to the fact that the universe has expanded between when the light was emitted and when it was measured. The expansion means that the true *comoving* distance can be obtain by (Calcino & Davis, 2017):

$$d_L = (1 + z)d = (1 + z_{\rm cos})(1 + z_{\rm pec})d. \tag{1.49}$$

Note that at low redshift, $z \ll 1$, where the expansion is negligible, the comoving distance tends to the luminosity distance.

[12] The rotation of the Earth if the telescope is ground based, or its orbit around the Earth in the case of a space telescope.

1.4.6 Estimating radial peculiar velocities

As mentioned in section 1.4.2 the velocity is the deviation from the expansion due to the net gravitational force a galaxy feels. Therefore if an estimate of the galaxy's distance exists then, at low redshift, one may approximate $z \approx z_{\cos} + z_{\rm pec}$ and $d_L \approx d$, v_r can be simply written:

$$v_r = cz - H_0 d. \tag{1.50}$$

One can recognize the terms of the Hubble Law on the right hand side, leaving the radial peculiar velocity as the deviation of each point to this law on the left hand side. This approximation is only valid within a few dozen of Mpc, when $cz_{\cos} = H_0 d$ is non-relativistic.

The approximation breaks down at high redshift when the cosmological expansion velocity is no longer non-relativistic. In this case, assuming the peculiar velocity is non-relativistic, the radial peculiar velocity can be isolated from eqs. (1.43) and (1.47):

$$v_r = c\frac{z - z_{\cos}(d)}{1 + z_{\cos}(d)}. \tag{1.51}$$

where $z_{\cos}(d)$ is obtained by numerical resolution of eq. (1.41). Note that the peculiar velocity does not depend only on the redshift but rather on both the redshift and the distance of the galaxy. These two quantities thus need to be measured in order for the radial peculiar velocity to be estimated (Davis & Scrimgeour, 2014).

1.4.7 Estimating distances from observed redshifts

As detailed in section 1.1.2, estimating the distance to a galaxy is quite challenging. The result obtained can be marred by large errors and subject to biases (extensively discussed in section 1.5). As we have seen in section 1.1.3, spectroscopic redshifts are relatively easy to obtain, and have only small uncertainties.

Recall that eq. (1.41) displays a straightforward relation between the distance and the redshift. If the cosmological redshift is known, the distance can be computed (assuming values for Ω_m, Ω_Λ). This estimator is problematic because the observed redshift is not the cosmological redshift, but rather the total redhift, namely the combination of the cosmological and peculiar redshifts. In absence of knowledge of the cosmological redshift, eq. (1.41) can only be applied to the observed redshift z:

$$d_z = \frac{c}{H_0}\int_0^z \frac{1}{\sqrt{\Omega_m(1+\mathcal{Z})^3 + (1-\Omega_\Lambda)}}\mathrm{d}\mathcal{Z} \approx \frac{cz}{H_0}. \tag{1.52}$$

Note that the difference between and eq. (1.41) is the upper limit on the integral. When using section 1.4.7 to compute the distance, a bias is introduced due to the the presence of a peculiar redshift in the observed redshift. In other words, the local movement of the galaxy biases the estimation of distance when the distance is computed from the observed (total) redshift. This effect is well known under the name of Redshift Space Distortion (RSD). Equations (1.50) and (1.95) and section 1.4.7 can be used to assess the size of the bias:

$$\left.\begin{array}{l} d_z \approx cz/H_0 \approx d - v_r/H_0 \\ \\ v_r \sim \mathcal{N}(0, \sigma_v) \end{array}\right\} \implies d_z \sim \mathcal{N}\left(d, \sigma_{d_z}\right), \qquad \sigma_{d_z} = \frac{\sigma_v}{H_0}. \tag{1.53}$$

RSD manifests itself in two regimes: where galaxies are laminar flowing in the linear, non-vortical, regime towards density peaks; and where galaxies exhibit chaotic motion deep in the non-linear regime, after shell crossing and accretion into clusters. First, in the low to medium density regions where the velocity field is relatively linear the theory dictates $\sigma_v \approx 300\,\mathrm{km/s}$, the uncertainty on d_z thus reads $\sigma_{d_z} \approx 3\,\mathrm{Mpc}/h$. This effect leads to the creation of Kaiser's pancakes, structures that appear flattened transversally to the line of sight (Kaiser, 1987; Praton et al., 1997; Thomas et al., 2004). In addition, the peculiar velocities of galaxies is correlated over large distances (see section 1.6). The RSD is thus non-local and large patches of the sky can be coherently distorted.

The second regime of the RSD affect massive clusters, where the virial motion of galaxies dominates over the linear motion. It manifests under the form of so-called Fingers of God (Jackson, 1972), spurious structures elongated along the line of sight. In these regions, the radial peculiar velocities of galaxies is so high that the galaxies behind the clusters ($d_{\rm gal} > d_{\rm cluster} \implies v_{r,\rm gal} < 0$) appear in front of it in redshift space while the galaxies in front of the cluster ($d_{\rm gal} < d_{\rm cluster} \implies v_{r,\rm gal} > 0$) appear behind it. In this region, σ_{d_z} can be larger than a dozen of Mpc/h and depends on the mass of the cluster.

The correction of the RSD to recover the true distances from redshift measurements is the subject of past and ongoing research (*e.g.*; Peacock & Dodds, 1994; Tsujikawa, 2013).

1.5 Biases in the observations of distances

Any observation of distance is subject to an number of biases: physical or observational effects often counter-intuitive that lead to systematically bend the result in an undesired way when not properly accounted for.

Most of the biases that affect our work come from the estimation of the distances of the galaxies. Indeed, one would naturally think that the probability of the true distance is entirely described by the probability of observation of its distance moduli $\mu_{\rm obs}$ with an error σ_μ. This is however not the case.

Using the Bayes theorem, one can write the probability of the true distance d:

$$P(d|\mu_{\rm obs}) = \frac{P(\mu_{\rm obs}|d)P(d)}{P(\mu_{\rm obs})}, \tag{1.54}$$

$$P(d|\mu_{\rm obs}) \propto P(\mu_{\rm obs}, d)P(\mu_{\rm obs}|d)P(d) \tag{1.55}$$

where $\mu_{\rm obs}$ is the observed distance modulus. In presence of a selection function, this probability law has to be renormalized (Strauss & Willick, 1995; Hinton et al., 2017):

$$P(d|\mu_{\rm obs}, \mathcal{O}) \propto \frac{P(\mathcal{O}|\mu_{\rm obs}, d)P(\mu_{\rm obs}|d)P(d)}{P(\mathcal{O}|d)} \tag{1.56}$$

where $\mathcal{O} = \{0, 1\}$ is the event "the galaxy has been observed". Note that the normalization can not be ignored in eq. (1.56) while it was discarded in eq. (1.55). This difference of treatment is due to the fact that the normalization depends on the fitted parameter in the second case and not in the first.

A good notation convention is the key to the understanding of the these biases. We distinguish the true properties of the observed galaxies with the subscript true: $\mu_{\rm true}$, $d_{\rm true}$, etc. These are fixed qualities, there exists one and only one *true* distance for a given galaxy. Similarly, we note observed quantities with the subscript obs: $\mu_{\rm obs}$, $d_{\rm obs}$, etc. Again, these are fixed quantities, as we are dealing with one single observation per galaxy. Finally, the random variables do not have any subscript: μ, d, etc. These are variables, that are *a priori*, unknown. The purpose of observations is to have $P(d) = \mathbb{I}(d_{\rm true} - d)$, that is to say that the random variable d converges to $d_{\rm true}$. However, in presence of biases and errors, we will see that is not the case.

1.5.1 The Log-Normal bias

We begin by discussing the source of bias from $P(\mu_{\rm obs}|d)$. Note that because the errors on the redshift are relatively small compared to the errors on the estimated distances, for matter of simplifications, we neglect them in this discussion.

Let us consider a galaxy at a distance $d_{\rm true}$ having a radial peculiar velocity $v_{r,\rm true}$. We note its true distance moduli $\mu_{\rm true}$, and luminosity distance $d_{L,\rm true}$. Suppose for the sake of simplicity that this galaxy is close enough so that $d_{\rm true} \approx d_{L,\rm true} = 10^{\mu_{\rm true}/5-5}$ and $v_{r,\rm true} = cz_{\rm total} - H_0 d_{\rm true}$. Let this galaxy have a distance error σ_μ. The probability to observe a distance moduli μ reads

$$\mu \sim \mathcal{N}(\mu_{\rm true}, \sigma_\mu) = \frac{1}{\sqrt{2\pi}\sigma_\mu} \exp\left(\frac{-(\mu - \mu_{\rm true})^2}{2\sigma_\mu^2}\right) \tag{1.57}$$

which transforms into a log-normal distribution on the luminosity distance $d_L = 10^{\mu/5-5}$:

$$d \sim \mathcal{L}(d_{\rm true}, \nu_\mu) = \frac{1}{\sqrt{2\pi}\nu_\mu d} \exp\left(\frac{-\ln^2\left(d/d_{\rm true}\right)}{2\nu_\mu^2}\right), \qquad \nu_\mu = \frac{\ln(10)}{5}\sigma_\mu. \tag{1.58}$$

The log-normal distribution is skewed. One of the consequences is than its mean is skewed:

$$\langle d \rangle = d_{\rm true}\, e^{\nu_\mu^2/2} \approx d_{\rm true}\left(1 + \nu_\mu^2/2\right), \tag{1.59}$$

$$\sigma_d = d_{\rm true}\nu_\mu. \tag{1.60}$$

This equation is the core of the Log-Normal bias. It shows that the mean over an ensemble of measured distances $\langle d \rangle$ of a single galaxy is always greater than the galaxy's real true distance $d_{\rm true}$.

If we consider n galaxies all at the exact same distance $d_{\rm true}$ and measure their distance moduli $\{\mu_i\}$, the mean of the computed distances $\langle d \rangle_i$ is greater than $d_{\rm true}$. Even if the same distance moduli was

measured a great number of times independently for each galaxy, the mean distance recovered would still be over-estimated.

Equation (1.59) also highlights the proportionality of the error with the distance. Indeed, an error of $\sigma_\mu = 0.5 \implies \nu_\mu \approx 0.2$ on the distance modulus which results in a $2\,\mathrm{Mpc}/h$ uncertainty for a $10\,\mathrm{Mpc}/h$ distance measurement, and a $20\,\mathrm{Mpc}/h$ uncertainty on a $100\,\mathrm{Mpc}/h$ measurement.

The mean and standard deviation of the observed radial peculiar velocity of a single galaxy can be approximated from eq. (1.50):

$$\langle v_r \rangle = cz - H_0 \langle d \rangle = cz - H_0 \left(1 + \nu_\mu^2/2\right) d_\mathrm{true} < cz - H_0\, d_\mathrm{true} = v_{r,\mathrm{true}}, \tag{1.61}$$

$$\sigma_v^2 = \sigma_{cz}^2 + H_0^2 \sigma_d^2 = \sigma_{cz}^2 + H_0^2 \nu_\mu^2 d_\mathrm{true}^2. \tag{1.62}$$

The systematic over-estimation of the distance leads directly to an under-estimation of the radial peculiar velocity. The large error on the distance can have quite a dramatic effect on the velocity. Indeed, in the absence of measurement, we know that the radial peculiar velocity of any galaxy follows $\mathcal{N}(0, \sigma_v)$, with $\sigma_v \approx 300\,\mathrm{km/s}$ (see eqs. (1.95) and (1.96)). However, given an uncertainty on the distance modulus of say $\sigma_\mu = 0.4$ (*resp.* $\sigma_\mu = 0.05$), the observational uncertainty $\sigma_{v,\mathrm{obs}}$ exceeds $300\,\mathrm{km/s}$ for all galaxies more distant than $15\,\mathrm{Mpc}/h$ (*resp.* $120\,\mathrm{Mpc}/h$). In other words, at $150\,\mathrm{Mpc}/h$, with an uncertainty on the measured distance modulus of $\sigma_\mu = 0.4$, the implied uncertainty on the peculiar velocity is $\sigma_{v,\mathrm{obs}} = 3\,000\,\mathrm{km/s}$, *i.e.* ten times the likely ΛCDM value! This means that, for a distant galaxy, a measurement of the peculiar velocity has virtually no constraining power at all, since its error is so much greater than the expected value.

The reader will note that the error on both the distance and on the peculiar velocities are functions of the *true* distance, which is unknown. In presence of both a redshift and a direct distance measurements, the redshift should be used as proxy to the true distance to estimate the biases and the uncertainties.

The bias discussed above, implies that the under-estimation of the radial peculiar velocity of each galaxy at a given distance leads to a spurious infall. This bias motivates the designing of *estimators* of radial peculiar velocities, which try to over-come this effect and correct the statistics (see section 1.9.5 or *e.g.* Watkins & Feldman, 2015). In this thesis, we detail two other methods in chapters 2 and 3. More visual examples of this bias are to be found in chapter 3.

1.5.2 Flux bias Malmquist Biases

The second bias discussed is named after Karl Gunnar Malmquist, who first discussed it in Malmquist (1922). In its original form, it is a selection bias which describes the probability that an existing galaxy is effectively observed, which is written $P(\mathcal{O}|\mu_\mathrm{obs}, d)$ in eq. (1.56). The Malmquist bias stems from the existence of an observational flux threshold. Indeed, each telescope (associated with a light detector, *e.g.* a CCD, photographic plate, the eye) has a flux1999; Boruah et al., 2022b), and its consequences on our methods are discussed throughout this work. threshold below which light is not detected. Thus, galaxies that *appear* dimmer than this threshold are not detected and *de facto* "invisible" to this device. A galaxy that *appears* dim is either close and faint or bright but distant enough that its light has be diluted in space. If we make the naive assumptions that all celestial objects are identical and have the same absolute luminosity, the number of galaxies drops with distance (instead of increasing, see section 1.5.3) and that only bright galaxies are to be found in the far away Universe.

1.5.3 Homogeneous Malmquist bias

By abuse of notation, the term of Malmquist bias has unfortunately, in common practuce, become generalized to all observational biases. The homogeneous Malmquist is a geometrical observational bias that affects the prior distribution of distance $P(d)$. It simply models the idea that in a homogeneous universe, there are more points in the shell at $[d, d + \mathrm{d}d]$ than in the shell $[d - \mathrm{d}d, d]$. Indeed, while the number density of galaxies is constant, the volume of the shells grow with distance, and so thus should too the number of observed galaxies (Strauss & Willick, 1995):

$$P(d) \propto d^2. \tag{1.63}$$

When neglected, galaxies tend to be put to closed to the observer, which tends to over-estimate the radial velocities following eq. (1.50).

1.5.4 Inhomogeneous Malmquist bias

The inhomogeneous Malmquist bias also affects the prior distribution of the distances $P(d)$. The homogeneous Malmquist bias results for the geometry of the Universe but not from its content. The inhomogeneous Malmquist bias, is itself a consequence of the inhomogeneous distribution of the galaxies in the Universe. Indeed, the probability to find a galaxy at a distance d, given its direction $\hat{\boldsymbol{x}} = \boldsymbol{x}/d$, depends directly on the galaxy density distribution in the Universe:

$$P(d) \propto \rho_g(d\hat{\boldsymbol{x}}). \tag{1.64}$$

Modeling the inhomogeneous Malmquist bias is difficult as the true galaxy density distribution is unknown in the context of reconstruction (see section 1.9 or *e.g.* Strauss & Willick, 1995; Dekel et al., 1999; Boruah et al., 2022), and its consequences on our methods are discussed throughout this work.

Note that this bias differs slightly from the galaxy bias presented in section 1.3.3. The inhomogeneous Malmquist bias may affect the measurement of a galaxy's distance if the underlying distribution of *galaxies* is ignored. The galaxy bias arises when the matter and the galaxy distributions are confused. However, both a depend on the type of galaxy discussed, *e.g.* the distribution of elliptical galaxies is different from the distribution of dwarf galaxies.

1.5.5 Biases due to the nature of observational surveys

Lastly, there are a number of effects that can affect the apparent brightness of galaxies. For example the presence of interstellar cosmic dust or gas or even stars along the line of the sight leads to extinction effects as light is aborbed and reemitted. It is for instance notoriously difficult observe galaxies in the plane of our own galaxy since the density of Milky Way stars is so high there are no clear inter-stellar sight line. The "milky" nature of the *via lactea* creates a so called Zone of Avoidance (ZoA) (Hubble, 1934). When observing the sky from the ground, the light of galaxies at zenith travel through less atmosphere than the light of galaxies closer to the horizon. Since the atmosphere absorbs or scatters light, this makes observations on the horizon more difficult than at the zenith. Different telescopes, with different through-puts, different detection devices, and resolutions, must thus be used to compile catalogs of distance moduli. Furthermore, all telescopes have a limited frequency range in which light can be detected (*e.g.* the visible light for the eye, radio frequencies for a radio telescope, etc). Galaxies that redshifted might simply fall out of this frequency detection range. Finally, artificial limitations of the selection function, like a redshift cut (*i.e.* galaxies whose redshift is higher than a certain value are excluded from the sample because the survey is incomplete there) may lead to biases if they are not properly modeled. Indeed, any reconstruction has to be "aware" that the absence of observed galaxies is either physical, *i.e.* there are in fact no galaxies there, or artificial, *i.e.* for one of myriad of reasons the galaxies are unobservable.

The probability to observe a galaxy is thus not only function of its distance modulus $\mu_{\rm obs}$ and its distance d, but rather of many of its intrinsic properties combined with the technical specificities of telescope aiming at it. The main implications of this bias are discussed in section 1.9.2.

1.6 Statistics of the linear fields

Behind every star and galaxy, beyond the dust and the molecular gas clouds, a signal is emitted from any direction of the sky at a wavelength of about 1.063 mm: the Cosmic Microwave Background (CMB).

The CMB is close to being a perfectly uniform black body spectrum at 2.7K. Yet, variations in its temperature of about one hundred thousandth of the average value have been detected by the Smoot et al. (antenna Dicke radiometer; 1977), Smoot et al. (COBE; 1992); Fixsen et al. (COBE; 1996), Komatsu et al. (WMAP; 2011) and more recently by the Planck Collaboration et al. (2016) missions. These temperature variations are thought to be due to the inhomogenous density distribution in the early universe. Although seemingly random, these perturbations are correlated in space. This means that the knowledge the value of the field at a position constrains the expected randomness of its neighborhood. The fluctuations of the temperature field are tied to the ones of the density field. The CMB is thus used a source of information on the statistics of the early or large scale over-density field δ.

1.6.1 Gaussian process

The perturbations δ are well modeled by a Gaussian process (Kolb et al., 1990). Such a process is entirely characterized by a mean field – which we take as null here – and a two point correlation function[13] defined as (Wiener, 1930; Khintchine, 1934)

$$\xi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2) \rangle = \iint \delta(\boldsymbol{x}_1)\delta(\boldsymbol{x}_2)\ P\big(\delta(\boldsymbol{x}_1), \delta(\boldsymbol{x}_2)\big)\ \mathrm{d}\delta(\boldsymbol{x}_1)\ \mathrm{d}\delta(\boldsymbol{x}_2). \tag{1.65}$$

The probability of the Gaussian field is written on a finite set of n sampled positions $\boldsymbol{\delta} = \{\delta_i\}_{i<n} = \{\delta(\boldsymbol{x}_i)\}_{i<n}$ rather than on the continuous field itself. Such a set of values follows a Multivariate Normal Distribution (MND):

$$\begin{aligned} P : \mathbb{R}^n &\to [0, 1] \\ \boldsymbol{\delta} &\to \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|} \exp\left(\frac{-1}{2}\boldsymbol{\delta}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}\right) \end{aligned} \tag{1.66}$$

where $\boldsymbol{\Sigma} \in \mathcal{M}_{n\times n}$ is a positive definite matrix named the correlation matrix, $|\boldsymbol{\Sigma}|$ is its determinant and $\boldsymbol{\Sigma}^{-1}$ is its inverse. The coefficients of the correlation matrix are the values of the two points correlation ξ at the sampled positions

$$\boldsymbol{\Sigma}_{ij} = \xi(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \delta(\boldsymbol{x}_i)\delta(\boldsymbol{x}_j) \rangle = \langle \boldsymbol{\delta}_i \boldsymbol{\delta}_j \rangle\,, \quad 0 < i, j < n. \tag{1.67}$$

As stated above, a fundamental principle in cosmology is that the Universe is homogeneous and isotropic on large scales: there is no preferred or special direction. These assumptions simplify the form of the two point correlation functions which becomes

$$\xi(\boldsymbol{x}_i, \boldsymbol{x}_j) = \xi\left(|\boldsymbol{x}_j - \boldsymbol{x}_j|\right). \tag{1.68}$$

Namely, the two point function depends only on the *separation*. The marginal law on each value of the MND is a normal law centered on zero and with a width $\sigma_\delta^2 = \xi(0)$:

$$\delta(\boldsymbol{x}_i) \sim \mathcal{N}\big(0, \sigma_\delta\big). \tag{1.69}$$

Note that it is the same law for all values, which is a direct consequence of the hypothesis of homogeneity of the Universe.

1.6.2 The matter power spectrum

The frequencial approach is often preferred to the spatial approach. Indeed, under the assumptions of homogeneity and isotropy, the Fourier modes of the field are *independent* random variables, which simplifies greatly the use of this model. Throughout this work, the following Fourier transform convention is adopted (same convention as in Peacock, 2007):

$$\tilde{f}(\boldsymbol{k}) = \mathrm{FT}[f](\boldsymbol{k}) = \int_{\mathbb{R}^3} f(\boldsymbol{x}) e^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{x}, \tag{1.70}$$

$$f(\boldsymbol{x}) = \mathrm{FT}^{-1}[\tilde{f}](\boldsymbol{x}) = \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \tilde{f}(\boldsymbol{k}) e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{k}. \tag{1.71}$$

$$\tag{1.72}$$

The Fourier transform of the over-density field δ is thus noted $\tilde{\delta}$. The probabilty of the mode associated with the wavenumber $\boldsymbol{k}$ depends solely on the value of the *power spectrum* $\mathcal{P}(\boldsymbol{k})$, which is simply the Fourier transform of the correlation function

$$\mathcal{P} = \mathrm{FT}[\xi] \quad \Longleftrightarrow \quad \xi = \mathrm{FT}^{-1}[\mathcal{P}]. \tag{1.73}$$

Furthermore, the assumption of homogeneity and isotropy of the Universe simplifies the form of the power spectrum $\mathcal{P}(\boldsymbol{k}) = \mathcal{P}(k)$. The writing of ξ is thus also simplified:

$$\xi(r) = \int_0^\infty k^2 j_0(kx)\mathcal{P}(k)\mathrm{d}k \tag{1.74}$$

[13] Which is in fact a covariance function.

where j_0 is the spherical Bessel function of order 0.

The values of the power spectrum $\mathcal{P}(k)$ are constrained from measurements of the CMB on large wavelength by collaboration such as Planck Collaboration et al. (2016) and on shorter scale by redshift surveys (*e.g.* SDDS; York et al., 2000). The amplitude of the power spectrum is a cosmological parameters whose value is still debated. It is directly linked to the amplitude of the over-density field. The standard deviation of the over-density field can be computed from the power spectrum

$$\sigma_\delta{}^2 = \xi(0) = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3}\mathcal{P}(k)\mathrm{d}\boldsymbol{k} = \frac{1}{2\pi^2}\int_0^\infty k^2\mathcal{P}(k)\mathrm{d}k \tag{1.75}$$

but depending on the shape of $\mathcal{P}(k)$ it does not necessarily converge (*i.e.* $k^2\mathcal{P}(k)$ is not integrable). A solution is to smooth the field with a kernel $W(kR)$:

$$\delta_\mathrm{R} = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} W(kR)\tilde{\delta}(\boldsymbol{k})\mathcal{P}(k)\mathrm{d}\boldsymbol{k}, \tag{1.76}$$

$$\sigma_{\delta_\mathrm{R}}{}^2 = \frac{1}{2\pi^2}\int_0^\infty k^2W^2(kR)\mathcal{P}(k)\mathrm{d}k, \tag{1.77}$$

Two examples of commonly used kernels are:

$$W(kR) = \exp\left(-\frac{(kR)^2}{2}\right) \qquad \text{Gaussian kernel,} \tag{1.78}$$

$$W(kR) = \frac{3j_1(kR)}{kR} \qquad \text{Top-Hat kernel,} \tag{1.79}$$

where j_1 is the spherical Bessel function of order 1. Historically, a smoothing length of $8\,\mathrm{Mpc}/h$ with a Top-Hat kernel is used, as first estimations of the over-density fields gave $\sigma_8 \approx 1$ (Davis & Peebles, 1983),

$$\sigma_8{}^2 = \frac{9}{128\pi^2}\int_0^\infty {j_1}^2(8k)\mathcal{P}(k)\mathrm{d}k. \tag{1.80}$$

This measures the amplitude of fluctuations on a scale of 8 Mpc, *i.e.* $\sigma_8 = \delta T/T$. Planck Collaboration et al. (2016) gives $\sigma_8 = 0.829 \pm 0.015$. As smoothing is a linear process, the smoothed field are as Gaussian fields (Bardeen et al., 1986).

Figure 1.1 shows the power spectrum derived from Planck Collaboration et al. (2016) and the derived two points correlation function. The power spectrum peaks slightly above $k = 10^{-2}\,h/\mathrm{Mpc}$ which is equivalent to a wavelength of about $600\,\mathrm{Mpc}/h$. It tends to zero both for $k \to 0$ which is consistent with an homogeneous Universe. For large values of k, the power spectrum decreases as k^{-1}, meaning that $k^2\mathcal{P}(k)$ is not integrable and that for the non-smoothed field, $\sigma_\delta \to \infty$.

As can be seen in the third panel of Figure 1.1, the two point correlation function drops around $10\,\mathrm{Mpc}/h$. Points that are more than $50\,\mathrm{Mpc}/h$ apart are almost uncorrelated. There is however a peak of correlation at about $100\,\mathrm{Mpc}/h$ visible in the second panel. This is the signal of Baryonic Acoustic Oscillation (BAO) (Eisenstein & Hu, 1998). BAO are broadly speaking spherical shells of over-density of about $110\,\mathrm{Mpc}/h$ in radius spread across the Universe. These over densities are a result of acoustic waves in the primordial plasma which froze at the epoch of decoupling. Their detection remains until now only statistically evidenced (*i.e.* as a bump in the correlation function Beutler et al., 2011; Alam et al., 2017). The measurement of a single instance of a BAO has been however claimed by Einasto et al. (2016).

1.6.3 Statistics of the Fourier modes

The statistics of the modes of the over-density depend solely on the power spectrum (Wiener, 1930; Khintchine, 1934). Both the real and the imaginary parts of each mode follows independent normal laws (Bardeen et al., 1986):

$$P\left(\tilde{\delta}^\Re(\boldsymbol{k})\right) = \mathcal{N}\left(0, \sqrt{\mathcal{P}(k)/2}\right), \qquad P\left(\tilde{\delta}^\Im(\boldsymbol{k})\right) = \mathcal{N}\left(0, \sqrt{\mathcal{P}(k)/2}\right). \tag{1.81}$$

where $\tilde{\delta}^\Re(\boldsymbol{k})$ and $\tilde{\delta}^\Im(\boldsymbol{k})$ are respectively, the real and imaginary parts of each mode: $\tilde{\delta}(\boldsymbol{k}) = \tilde{\delta}^\Re(\boldsymbol{k}) + \mathrm{i}\tilde{\delta}^\Im(\boldsymbol{k})$. The amplitude of the mode thus follows a Rayleigh distribution

$$P\left(\tilde{\delta}(\boldsymbol{k})\right) = \frac{2\left|\tilde{\delta}(\boldsymbol{k})\right|}{\mathcal{P}(k)}\exp\left(\frac{-\left|\tilde{\delta}(\boldsymbol{k})\right|^2}{\mathcal{P}(k)}\right). \tag{1.82}$$

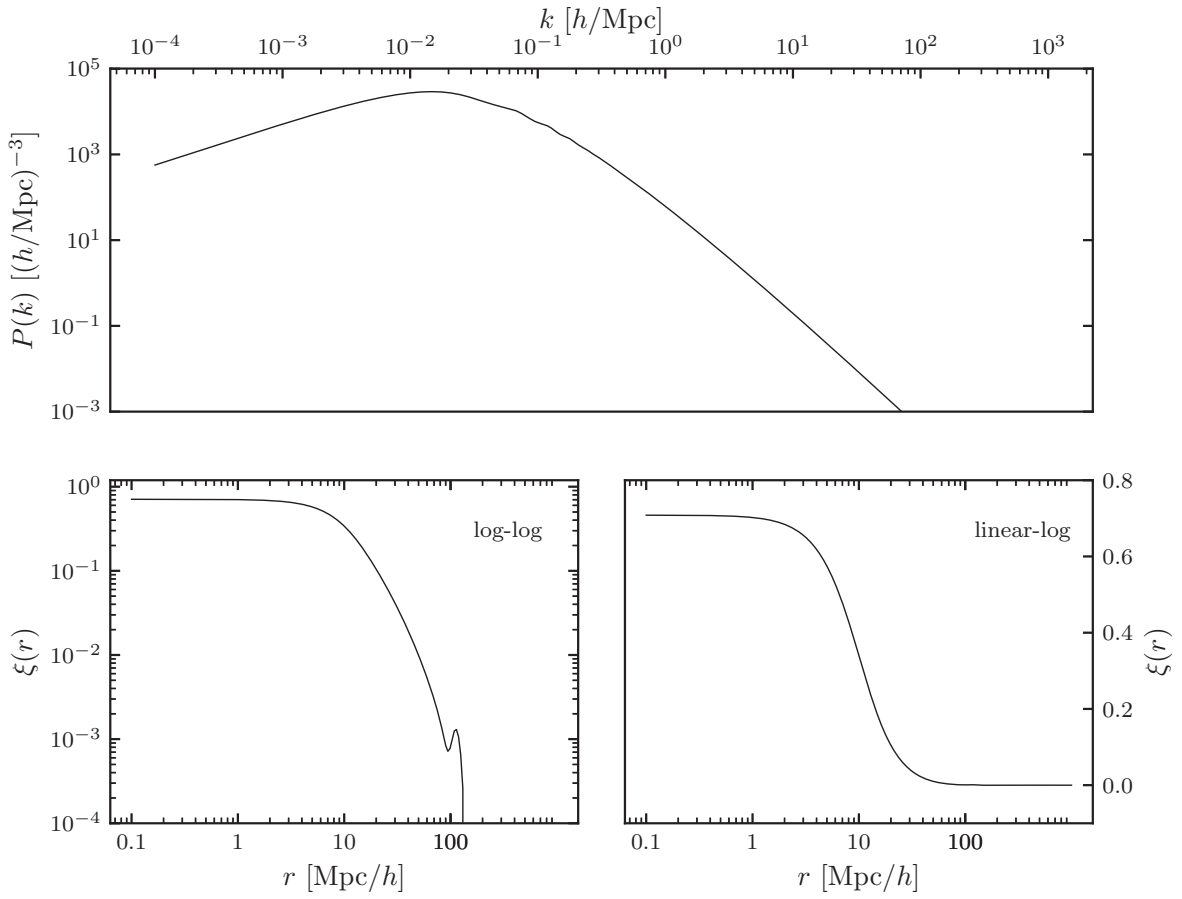

Figure 1.1: Top row: the matter-matter power spectrum of Planck Collaboration et al. (2016). Bottom row: the correlation function in log-log and in semi-log. To derive the correlation function, the power spectrum has been smoothed with a Gaussian kernel of 1 Mpc/h. Note the peak of correlation of the Baryonic Acoustic Oscillation (BAO) at $\gtrsim 110\,\mathrm{Mpc}/h$ in the correlation function.

The correlation between modes is

$$\left\langle \tilde{\delta}(\boldsymbol{k})\tilde{\delta}^{\dagger}(\boldsymbol{k}')\right\rangle(\boldsymbol{k},\boldsymbol{k}') = \mathbb{I}(\boldsymbol{k}-\boldsymbol{k}')\mathcal{P}(k) \tag{1.83}$$

where $\mathbb{I}$ is the Kronecker symbol, and $\tilde{\delta}^{\dagger}(\boldsymbol{k}')$ is the complex conjugate. These last equations highlight the independence between modes and yields the mean amplitude of each mode:

$$\left\langle |\tilde{\delta}(\boldsymbol{k})|^2 \right\rangle = \mathcal{P}(k). \tag{1.84}$$

1.6.4 The linear velocity field

The observations are made for the density field either at early times or large scales. In both case, the linear theory can be safely applied. The relationship between over-density and velocity fields eq. (1.38) reads in the Fourier space:

$$\tilde{\boldsymbol{u}}(\boldsymbol{k}) = Hf\frac{\mathrm{i}\boldsymbol{k}}{k^2}\tilde{\delta}(\boldsymbol{k}). \tag{1.85}$$

From this relationship the self- and cross-correlation functions between density and velocity fields can

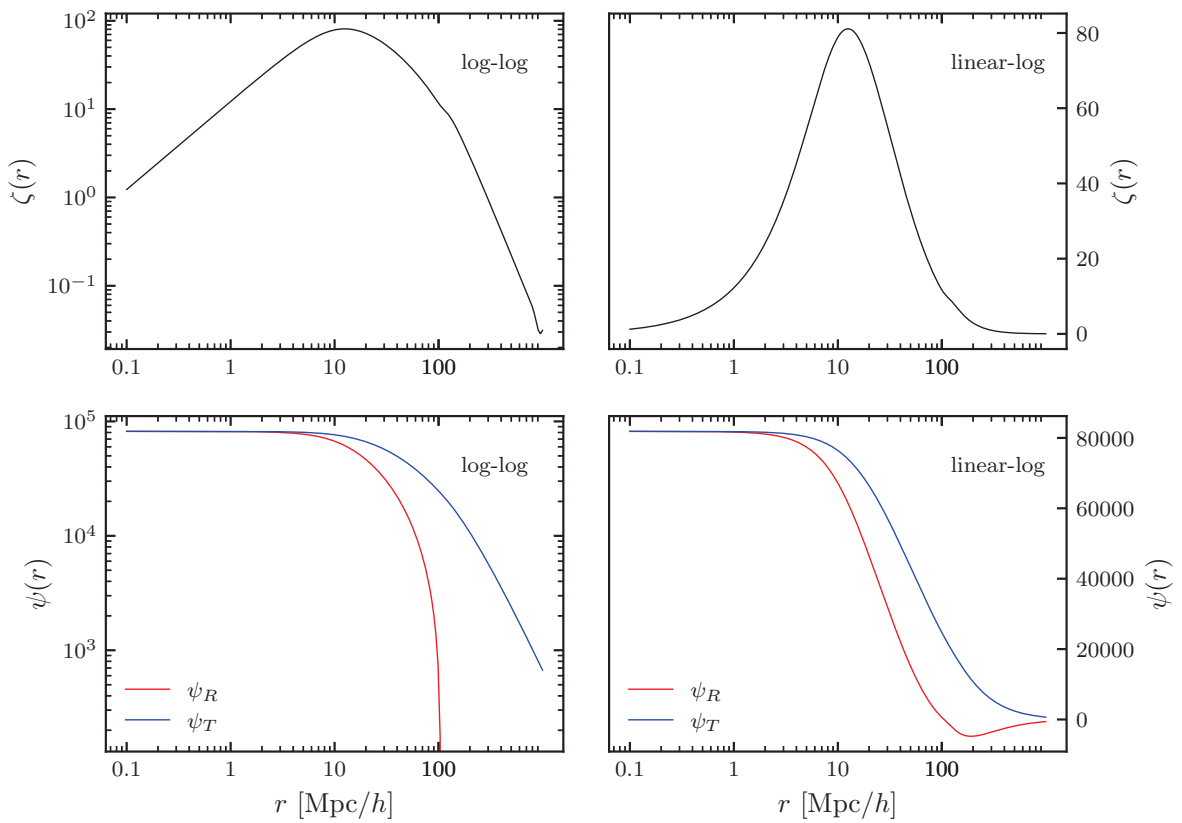

Figure 1.2: Top row: the isotropic density-velocity correlation function $\zeta(r)$ (left in log-log, right in semi-log). Bottom row: the isotropic radial and transversal velocity-velocity correlation functions $\psi_R(r)$ and $\psi_T(r)$.

be derived. The density-velocity correlation function reads

$$\begin{aligned}\zeta_a &= \langle u_a \delta \rangle & a = x, y, z \\ &= \mathrm{FT}^{-1}\left[\langle \tilde{u}_a(\boldsymbol{k})\tilde{\delta}(\boldsymbol{k})\rangle\right] & a = x, y, z \\ &= \mathrm{FT}^{-1}\left[Hf\frac{\mathrm{i}k_a}{k^2}\mathcal{P}(k)\right] & a = x, y, z\end{aligned} \tag{1.86}$$

and the velocity-velocity correlation function reads

$$\begin{aligned}\psi_{ab} &= \langle u_a u_b \rangle & a, b &= x, y, z \\ &= \mathrm{FT}^{-1}\left[\langle \tilde{u}_a(\boldsymbol{k})\tilde{u}_b(\boldsymbol{k})\rangle\right] & a, b &= x, y, z \\ &= \mathrm{FT}^{-1}\left[-(Hf)^2\frac{k_a k_b}{k^4}\mathcal{P}(k)\right] & a, b &= x, y, z.\end{aligned} \tag{1.87}$$

Simplifications can be made in a Friedman universe. The density-velocity correlation function can be written as (Monin & Yaglom, 2007; Gorski, 1988; Zaroubi et al., 1999)

$$\zeta_a(\boldsymbol{r}) = \hat{r}_a \zeta(r) \tag{1.88}$$

$$\psi_{ab}(\boldsymbol{r}) = \left[\psi_T(r)\mathbb{I}_{ab} + [\psi_R(r) - \psi_T(r)]\hat{r}_a\hat{r}_b\right], \tag{1.89}$$

where $\psi_R(r)$ and $\psi_T(r)$ are respectively the velocity-velocity radial and tangential correlation functions.

These are only a function of the separation between the points r and read

$$\zeta(r) = \frac{Hf}{2\pi^2}\int_0^\infty kj_1(kr)\mathcal{P}(k)\mathrm{d}k, \tag{1.90}$$

$$\psi_R(r) = \frac{-(Hf)^2}{2\pi^2}\int_0^\infty \left[j_0(kr) - \frac{2j_1(kr)}{kr}\right]\mathcal{P}(k)\mathrm{d}k, \tag{1.91}$$

$$\psi_T(r) = \frac{-(Hf)^2}{2\pi^2}\int_0^\infty \frac{j_1(kr)}{kr}\mathcal{P}(k)\mathrm{d}k. \tag{1.92}$$

The correlation between two velocity vectors $\boldsymbol{u}_i(\boldsymbol{x}_i)$ and $\boldsymbol{u}_j(\boldsymbol{x}_j)$, pointing in possibly different direction (*e.g.* radial velocities) can be written with the mean of the 3×3 tensor $\boldsymbol{\psi}$

$$\langle \boldsymbol{u}_i\boldsymbol{u}_j\rangle = \boldsymbol{u}_i\boldsymbol{\psi}(\boldsymbol{x}_j-\boldsymbol{x}_i)\boldsymbol{u}_j = \sum_{a,b} u_{i,a}u_{j,b}\psi_{ab}(\boldsymbol{x}_j-\boldsymbol{x}_i) \tag{1.93}$$

while the correlation between any velocity vector $\boldsymbol{u}_j(\boldsymbol{x}_j)$ and the over-density on another point in space $\delta_i(\boldsymbol{x}_i)$ reads:

$$\langle \delta_i\boldsymbol{u}_j\rangle = \boldsymbol{\zeta}(\boldsymbol{x}_j-\boldsymbol{x}_i)\cdot\boldsymbol{u}_j = \sum_a u_{j,a}\zeta_a(\boldsymbol{x}_j-\boldsymbol{x}_i) \tag{1.94}$$

Note that, even though the correlations functions $\zeta(r)$, $\psi_R(r)$ and $\psi_T(r)$ are isotropic, the tensor $\Psi(\boldsymbol{r})$ and the vector $\boldsymbol{\zeta}(\boldsymbol{r})$ are not: the isotropy is broken by the $\hat{r}_a$ and $\hat{r}_a\hat{r}_b$ terms of eqs. (1.88) and (1.89). For instance, the correlation of the x component of the velocity field along the x axis $\langle \boldsymbol{u}_x(0,0,0)\boldsymbol{u}_x(x,0,0)\rangle = \psi_R(x)$ is different from the correlation function of the same quantity along the y axis $\langle \boldsymbol{u}_x(0,0,0)\boldsymbol{u}_x(0,y,0)\rangle = \psi_T(y)$.

In an isotropic Universe, the standard deviation of each component of the velocity field can be computed as:

$$\sigma_v^2 = \Psi_R(0) = \frac{(Hf)^2}{2\pi^2}\int_0^\infty \mathcal{P}(k)\mathrm{d}k \tag{1.95}$$

and the marginal law of the components of the velocity can thus be written

$$v_a(\boldsymbol{x}_i) \sim \mathcal{N}(0,\sigma_v). \tag{1.96}$$

As opposed to σ_δ, the value of σ_v converges (*i.e.* $\mathcal{P}(k)$ is integrable). Its value is about $275-300\,\mathrm{km/s}$ at $z=0$ for the power spectra of Komatsu et al. (WMAP; 2011) and Planck Collaboration et al. (Planck; 2016). As each component of the velocity field is normal, the distribution of its magnitude follow a Maxwell distribution.

Figure 1.2 displays correlation functions related to the velocity field. The top rows shows the density-velocity correlation ζ of eq. (1.86) in both log-log and log-linear scales. The peak of the correlation occurs around $10\,\mathrm{Mpc}/h$. Below $\sim 1\,\mathrm{Mpc}/h$ and above $\sim 100\,\mathrm{Mpc}/h$, no correlation between the two fields is expected. The form of this function is interesting as it does not peak at $r=0$ as the other correlation functions presented here. This means that the knowledge of the over-density field at a point in space does give any information about the velocity field at that same point, but it does constrain its neighborhood. This can be understood by looking at eq. (1.38): the over-density field is proportional to the divergence of the velocity field. Knowing the value of the over-density field does thus not constrain the value of the velocity but its gradient. The velocity field tends to converge (locally) to over-densities and diverge (locally) from under-densities. However, a (locally) constant offset in the velocity field does not affect the over-density field.

The bottom row of fig. 1.2 shows the radial and transversal correlation functions of the components of the velocity field ψ_R and ψ_T introduced in eqs. (1.91) and (1.92). The correlation length of the velocity field is about an order of magnitude bigger than the one of the over-density field: the radial correlation drops at $100\,\mathrm{Mpc}/h$ while the transversal correlation slowly decreases between 50 – $500\,\mathrm{Mpc}/h$. The longer correlation length of the velocity field with respect to the over-density field could be predicted from eq. (1.87). Indeed, the high frequencies are damped by the $\boldsymbol{k}_a\boldsymbol{k}_b/k^4 = \mathcal{O}(k^{-2})$ term. In absence of these frequencies, the velocity field is expected to vary slower than the over-density field.

1.6.5 Random realizations of periodic universes

The independence of the Fourier modes makes the creation of *random realizations* (RRs) of the density field trivial. In practice, these RRs are always evaluated on a grid. Let that grid be a cubic lattice and have n nodes in each direction and have a side of L. The Discrete Fourier Transform (DFT) is thus:

$$\tilde{f}_j = \sum_i f_i e^{-\mathrm{i}\boldsymbol{k}_j\cdot\boldsymbol{x}_i}, \quad f_i = \frac{1}{(2\pi)^3}\sum_j \tilde{f}_j e^{\mathrm{i}\boldsymbol{k}_j\cdot\boldsymbol{x}_i}, \tag{1.97}$$

where $1 \leq i \leq n^3$ is the ith node of the grid and $1 \leq j \leq n^3$ is the jth mode of the grid. Because of the intrinsic functioning of the DFT, modes whose wavelength are greater than $2L$ – so called tidal modes – are lost, as well as modes whose wavelength are shorter than $2L/n$.

To account for the discretization, the power spectrum needs to be renormalized to the resolution of the grid[14]:

$$\mathcal{P}(k) \to \mathcal{P}_{\mathrm{norm}}(k) = \mathcal{P}(k)\left(\frac{n}{L}\right)^3. \tag{1.98}$$

The fastest method to create a RR is in two steps:

1. the real and imaginary parts of each mode are independently drawn from eq. (1.81);
2. the over-density field is evaluated by inverse Fourier transform of these modes.

A second method is often used, even though is it computationally more expensive:

1. a random normal signal $\mathcal{N}(0,1)$ is drawn on the grid;
2. the signal is transformed in the Fourier space;
3. the modes of the signal are multiplied by $\mathcal{P}(k)$;
4. the over-density field is evaluated by inverse Fourier transform of these modes.

The velocity field can be easily computed using the inverse Fourier transform of eq. (1.85)

$$\boldsymbol{u} = \mathrm{FT}^{-1}\left[Hf\frac{\mathrm{i}\boldsymbol{k}}{k^2}\tilde{\delta}(\boldsymbol{k})\right] \tag{1.99}$$

which is in fact much easier than solving the Poisson problem of eq. (1.38). Using the DFT leads to the creation of (computational) universes that are periodic on the size of the box and thus anisotropic[15]. This modifies the correlation functions. Indeed, points that are on opposite corners of the cube are distant in an infinite universe but are in fact close in a periodic universe. This must be reflected in the correlation functions.

The function ζ has to computed from eq. (1.86). However, only one component needs to be computed, *e.g.* ζ_x, while the others can be evaluated thanks to the symmetries of the periodic box: $\zeta_x(x,y,z) = \zeta_y(y,x,z) = \zeta_z(z,x,y)$. ψ_R and ψ_T can be respectively computed from ψ_{xx} and ψ_{xy} using eq. (1.87). Again, the periodicity of the box allows for permutations and other functions ψ_{ab} for all combinations of $a, b = x, y, z$ can be evaluated from ψ_{xx} and ψ_{xy}.

The use of the DFT also limits the power found in the grid because of the lacking modes: the standard deviation of the over-density field on a grid is finite and always smaller than the σ_δ computed from eq. (1.75). The finer and the bigger the grid, the more power is to be expected.

[14] This factor comes from the implicit discretized $\mathrm{d}k \to \Delta_k$ of the DFT: $f_i = \frac{1}{(2\pi)^3}\sum_j \tilde{f}_j e^{\mathrm{i}\boldsymbol{x}_i\boldsymbol{k}_j}\Delta_k^3$. To insure convergence of $f_i \to f(\boldsymbol{x}_i) = \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3}\tilde{f}(\boldsymbol{k})e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{k}$ when $L/n \to 0$, one has to set $\Delta_k = n/L$. We simply report this factor to the power spectrum.

[15] For instance, the direction towards the face of the cube is different from the direction towards a corner of the cube. The first is periodic of length L, the second is periodic of length $\sqrt{3}L$. Any other direction is more complex.

1.7 Simulating the Universe

Even though linear theory describes quite accurately the formation of structure on very large scales or the early times of the Universe, it cannot render properly the modern Universe. Indeed, these equations are obtained through the simplification and the linearization of the full equations of the hydro-dynamics of matter. Among other things, non-linear motion, gas dynamics, or even electro magnetic interactions are omitted. The full equations can however not be solved analytically in the general case. This is a major issue in astrophysics and cosmology: the physical models of formation of different structures (cosmic web, galaxies, stars, etc) can be quite easily written but solving them can only be achieved numerically.

Following the increasing capabilities of computers and super-computers over the past few decades, astrophysicist have taken the path of cosmological simulations (Klypin & Shandarin, 1983; Springel, 2005). A simulation is like a lab experiment, but for a Universe and in a computer. Cosmological simulations start from an initial state – a very early universe – which is evolved, step by step following a set of rules – the aforementioned hydro-dynamics equations. This way, models can be run and tested against the observations of the modern Universe.

1.7.1 Initial conditions for cosmological simulations

Random initial conditions

A philosophical limitation of this approach is that the initial state of our Universe is only known statistically (ie inferred from the power spectrum of temperature fluctuations in the CMB) and not specifically.

A metaphor could maybe bring some light on the meaning and the implications of this statement. Imagine an artist who has never seen a mirror. They can draw very well any human face, and they can even make up human faces on command. They can however not possibly draw their own face: they can assume that they have a nose, a mouth, eyes, etc, like other humans they have seen, yet they do not know exactly what the exact shape their nose or their mouth is nor what the color of their eyes is. We are in the same situation: we have observed the early Universe over a very large volume very far away (the CMB), which enables us to create initial conditions statistically indistinguishable from these observations, but does not tell us anything about the exact features of our own local volume - since this is of course a subset of the entire Universe.

Cosmological simulations thus all start from statistically similar random initial conditions, and thus produce statistically similar universes, but not our exact Universe. These universes can be statistically compared to ours, *e.g.* to study the the mass function of Milky Way haloes and the merging history of Milky Way like galaxies, but cannot answer certain specific questions, *e.g.* the merging history of the actual Milky Way. In other words they can constrain the Milky Ways merger history statistically insofar as they can constrain the merger history of *all* haloes similar to the Milky Way, yet they can not predict the merging history of the real Milky Way.

The initial density and velocity fields of simulations have to describe the early Universe: the linear theory is then to be used here. For technical reasons, simulated universes are always periodic, thus the methodology of section 1.6.5 can be employed to create them.

Constrained initial conditions

Even though constrained initial conditions are slightly outside of the scope of this work, a short description could help understand the long term goal and the stakes of the reconstruction methods introduced in section 1.9 and discussed throughout this thesis.

It is possible to *constrain* the early density and velocity fields from an estimation of the evolved density and velocity fields (Dekel et al., 1999). The same simulation codes applied to these constrained initial conditions lead to the creation of constrained simulations of the Universe, where both the statistics and the positions of observed objects are recovered.

The fidelity of constrained simulation is limited at two levels. First, the complex equations that model the evolution of a universe are *chaotic*. This means that a small perturbation to the initial state can lead to a large difference at later times. A direct implication of this characteristic is the impossibility to properly go back in time: a tiny difference in the estimation the modern field leads to a very inaccurate set of initial conditions. This is a mathematical property due to the chaotic nature of the solutions to the coupled differential equations, and cannot be solved with higher computational capabilities. The machinery employed to reconstruct initial conditions from evolved field is thus bound to a certain degree of inaccuracy.

Secondly, and this is the main discussion of this work, estimating the modern density and velocity fields is itself very challenging. This introduce another source of uncertainty and possible biases (see section 1.5) that limits the quality of the constraints on the initial conditions. Thus part of the motivation of improving the reconstruction of today's universe (and of this thesis) is to improve future constrained simulations of the local universe (Yepes et al., 2009).

1.7.2 Dark matter only simulations

Cosmological simulations may model just the formation of structure due to gravity, namely solve the problem of N gravitating bodies. When considering pressure-less dark Matter, this is assumed to be sufficient - as opposed to simulations which also include baryons and must thus also follow magento-hydrodynamic forces. Dark matter only codes thus integrate over time the N-body problem of often millions of particles solely interacting through gravitation (Barnes & Hut, 1986). These particles are assumed to sample the mass distribution. Their mass is fixed depending on the size and the resolution of the simulated box, and they do not collide with each other in accordance with the assumption that dark matter particles are collision-less. The integration in time also takes into account the global expansion of the Universe. This type of simulation is suitable for studies of structure formation on scales where gravity dominates the dynamics. The approximation made by such simulations becomes less and less valid where galaxies are expected to form (*e.g.* Vogelsberger et al., 2019).

On smaller scales, the simulated dark matter particles collapse into very dense groups. These can be identified as *halos* by different methods (Friend of Friend, etc; Cole & Lacey, 1996; Behroozi et al., 2013). Galaxies are thought to form in and only in these halos of dark matter (Peacock & Smith, 2000). Most halo finding algorithms impose a threshold defining haloes as regions above a specific over density (Knebe et al., 2011).

Particle based approaches to the N-body problem are in essence Lagrangian, *i.e.* they trace the movement of particles not of the fields. To recover the density and velocity fields, a Cloud in Cell (CiC) algorithm can be used (Birdsall & Fuss, 1969). At the 0th order, a CiC method consists simply in counting the number of particles in a cell of a given grid. A slight improvement can be seen when the mass of the particle is spread on the nodes of the grids with respect to their spatial separation to the particle. The velocity field is computing in the same manner: at the 0th order, the velocity of a cell is the mean velocity of the particles in this cell. Again, more subtle approaches can refine the result. In general, the minimal size of a cell is governed by the number of particles of the simulation in the sense that if the CIC is too fine, the cells are too small, and do not smooth over the particle distribution. As such, a rule of thumb dictates that the number of cells in each dimension has to be at most 4 times the number of particles in each dimension (*i.e.* a simulation with 512^3 particles should be projected at most on a 128^3 nodes grid; Libeskind et al., 2014).

The reader will note that simulations are only used in this thesis in section 1.8 for a discussion on the limits of the linear theory and in chapter 3 where our method is tested against mock galaxy catalogs drawn from cosmological simulations.

1.8 Statistics of the fully evolved density and velocity fields

The statistics of the density and velocity evolved by dark matter simulations is discussed, with a focus on the comparison with the statistics of the linear fields. We use the New MultiDark Planck[16] (Riebe et al., 2013), a N-body run of $N = 3840^3$ particles in a periodic box of side length $L = 1\,\mathrm{Gpc}/h$. The cosmological parameters of the simulation are from Planck Collaboration et al. (2016), *i.e.* a flat ΛCDM Universe $\Omega_\mathrm{m} = 0.307$, $\Omega_\mathrm{b} = 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 = 0.8228$, $n_s = 0.96$ and a dimensionless Hubble parameter $h = 0.678$ where $H_0 = 100 \times h$ km/s/ Mpc. The fields are taken on a regular grid of 512^3 nodes, leading to a cell resolution of $1.9^3\,[\mathrm{Mpc}/h]^3$. Similar analyses are presented in Sheth & Diaferio (2001); Hamana et al. (2003, 2005); Doumler (2012).

In this section we compare two definitions of the over-density:

$$\delta_\rho = \rho/\bar{\rho} - 1 \qquad \text{obtained from the CiC,} \tag{1.100}$$

$$\delta_{\boldsymbol{v}} = -H_0 f \boldsymbol{\nabla} \cdot \boldsymbol{v} \qquad \text{the "linear" density of eq. (1.38).} \tag{1.101}$$

[16] https://www.cosmosim.org/metadata/mdpl2/

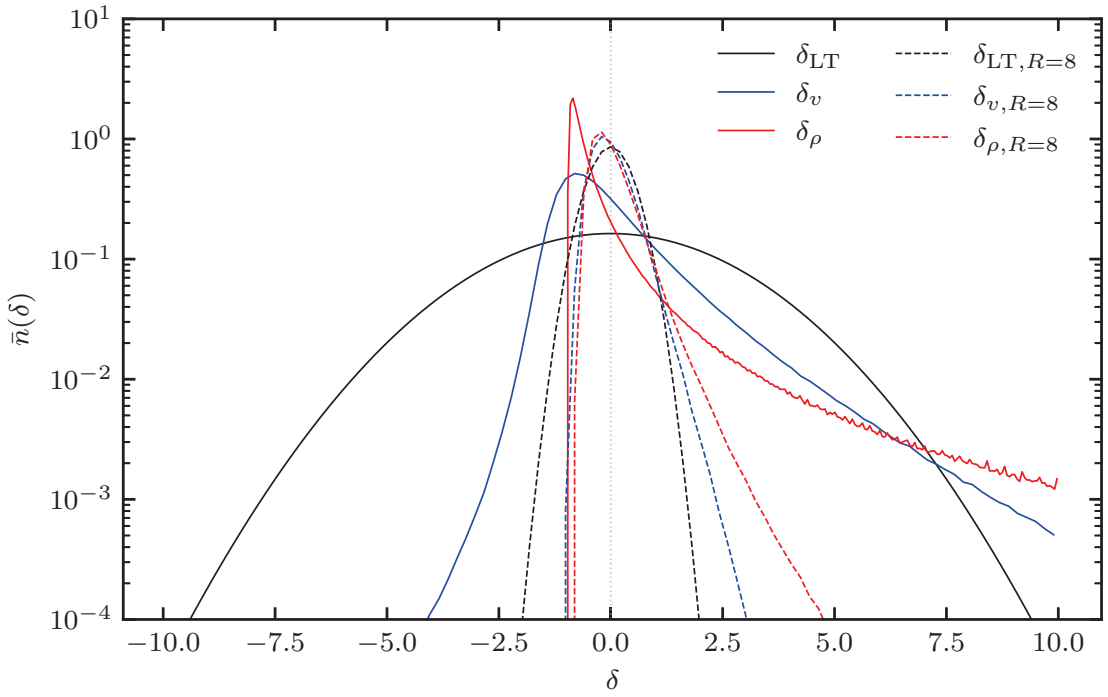

Figure 1.3: Distributions of several definitions of the over-density: predicted by the linear theory (black), derived from the evolved density field of a simulation (red) and by applying eq. (1.38) to the evolved velocity field of the simulation. The statistics of the fields smoothed at $R = 8\,\mathrm{Mpc}/h$ are dashed.

In the context of the linear theory, δ_ρ and $\delta_{\boldsymbol{v}}$ represent the same field. This is however not the case for non-linear evolved fields: $\delta_\rho \neq \delta_{\boldsymbol{v}}$. The quantity commonly understood as density for an evolved field is δ_ρ whereas $\delta_{\boldsymbol{v}}$ is just the divergence of velocity field.

These definitions are motivated by the use of the linear theory throughout this thesis to sometimes model non-linear fields. Some insight on the fundamental differences between these two statistics are thus necessary. In order to compare the statistics of non-linear fields to the ones of the linear theory, we also introduce $\delta_{\rm LT}$, a linear over-density field drawn from a random realization of the (Planck Collaboration et al., 2016) power spectrum on a grid similar to the one on which the CiC is computed (see section 1.6.5). Note that this field describes an universe unrelated to the one of the simulation, and that it is just used to compute the statistics predicted by the linear theory in a periodic universe evaluated on a finite grid.

1.8.1 Comparison with the linear theory

This section is articulated in two parts: the non-smoothed (over-density and velocity) fields are first compared and discussed, then the influence of smoothing on this comparison is studied.

The density field

Let us start this discussion with the statistics of the fields in the linear regime. As shown in Figure 1.3, the distribution of the density field $\delta_{\rm LT}$, is normal: it is symmetrical, with a null mean equal to its median. Half of the volume is over-dense, the other half is under-dense. Its 95%, 99% and 99.9% quantiles are respectively 4.0, 5.7 and 7.55 (these values are the same for the 5%, 1% and 0.1% percentiles). The two point correlations of the linear over-density field shown in fig. 1.4 (top panel), has been described in section 1.6.

The distribution of δ_ρ appears in fig. 1.3 to be log-normal. The most striking consequence is the skewness of this distribution. δ_ρ has (by definition) a null mean, has a lower boundary of -1 (due to the fact that ρ cant be negative) but reaches very high values. On one hand, more than 80% of the volume has a negative density and half has a density below -0.65, on the other hand the 95%, 99% and 99.9% percentiles respectively read 2.2, 9.8 and 45.1. The maximum value for this particular simulation and

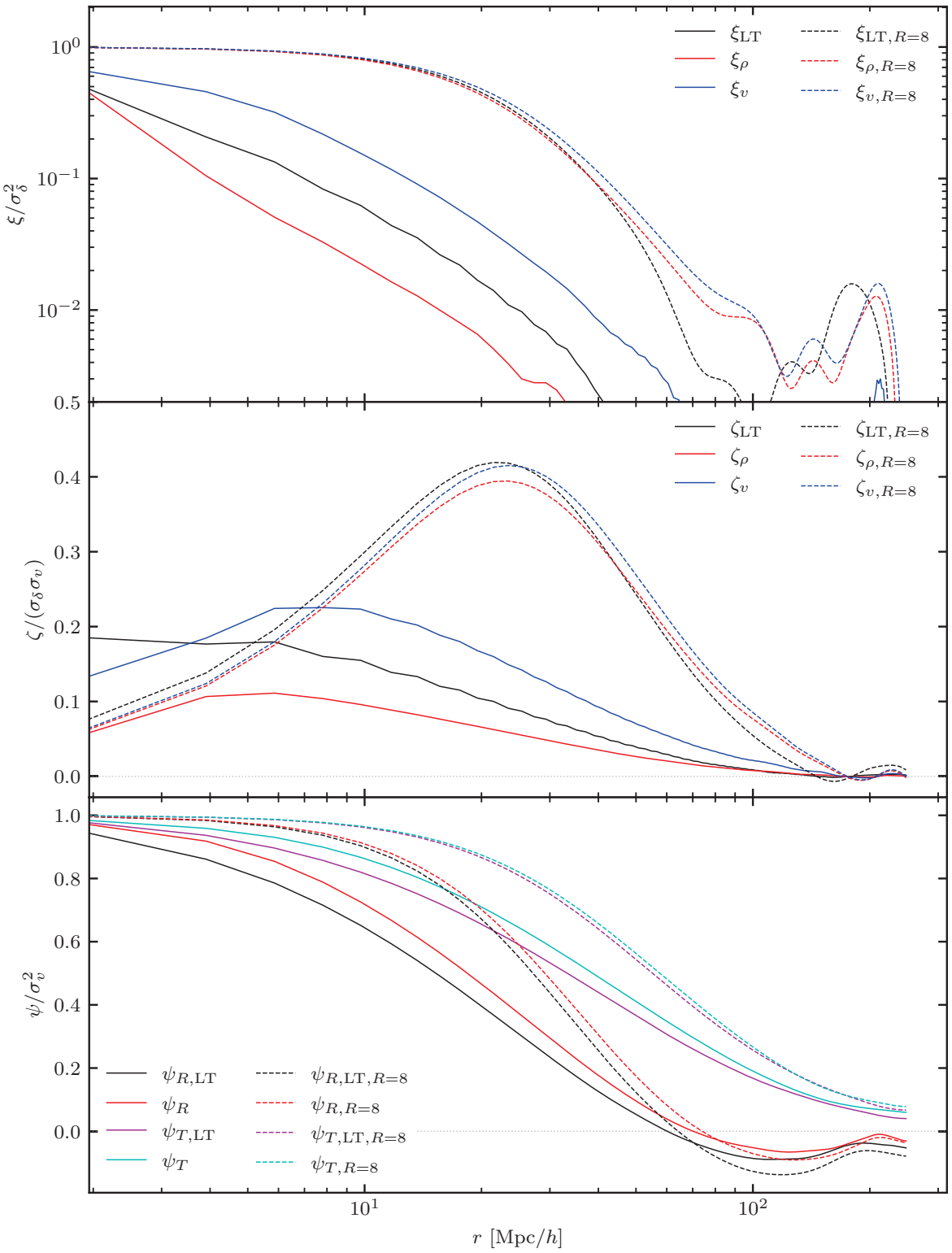

Figure 1.4: Two-point correlation functions for different fields and quantities. The functions are normalized by the standard devations of the considered field, they do thus represent the *correlation* rather than the *covariance*. **Top:** Two-point correlation functions of several definitions of the over-density: predicted by the linear theory (black), derived from the evolved density field of a simulation (red) and by applying eq. (1.38) to the evolved velocity field of the simulation. **Middle:** Two-point correlation functions with the velocity field of several definitions of the over-density, with the same color convention as the top panel. **Bottom:** Two-point correlation functions of the velocity field predicted by the linear theory (black) and of an evolved velocity field (in red). The statistics of the fields smoothed at $R = 8\,{\rm Mpc}/h$ are dashed.

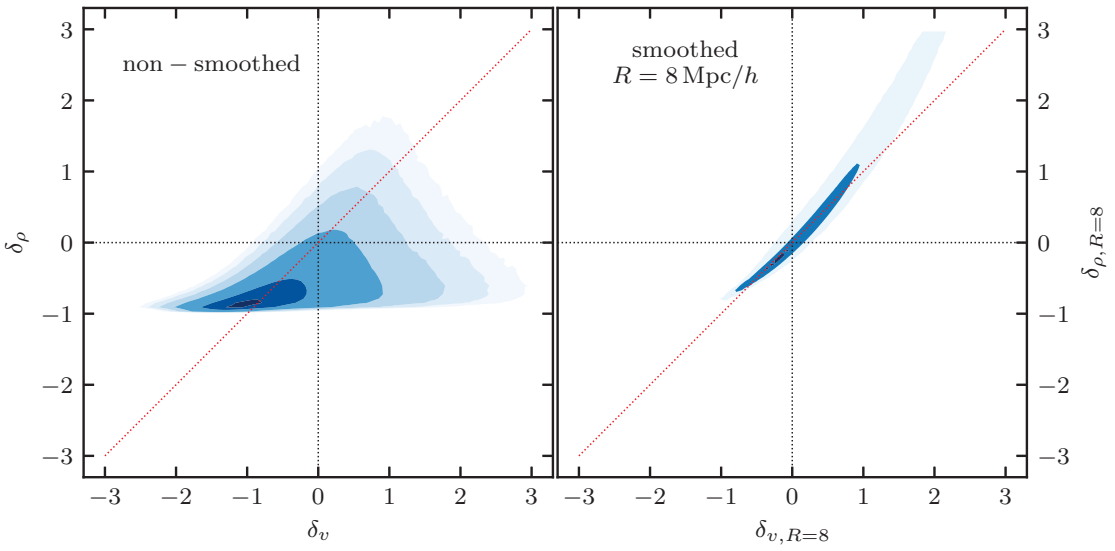

Figure 1.5: Two dimensional distribution of δ_ρ again $\delta_{\boldsymbol{v}}$. **Left:** for the non-smoothed field. **Right:** for the fields smoothed at $R = 8\,\mathrm{Mpc}/h$.

grid size reaches 1095 (not plotted), although this value is subject to large fluctuations depending on the size of the sample[17] and the resolution of the grid. The two-point correlation presented in fig. 1.4 (top panel) decreases even faster than the one predicted by the linear theory. The matter in an evolved universe is thus extremely clumped, rapidly changing in space, and leaves most of the volume empty.

The three main differences with the linear theory are thus the positivity of the density field ($\delta_\rho+1 > 0$) and its skewness. Describing the universe with the linear theory thus overestimates the depth of the void regions and underestimates the extreme compactness of the dense regions. Moreover, a linear field has as many over-dense cells as under-dense cells which does not capture the extreme high contrast and predominant void of an evolved universe. Finally, the size of the cosmic structures is smaller in an non-linear universe than in the linear theory. This does not indicate that they are less complex or rich, but rather that they are more compact.

In the face of this discrepancy, one might argue that these two quantities are not comparable: the linear over-density is the divergence of the linear velocity field and should thus be compared to the divergence of the evolved velocity field, which we note $\delta_{\boldsymbol{v}}$. Figure 1.3 shows that this quantity also does not follow a normal distribution. It is indeed less skewed than δ_ρ and not bound to -1: its 0.1% percentile reaches -2.6. However, it keeps some properties of δ_ρ: 65% of the volume is "under-dense" and half displays a $\delta_{\boldsymbol{v}} < -0.37$. The 95%, 99% and 99.9% percentiles are respectively 2.6, 5.4 and 10.1. The correlation function of $\delta_{\boldsymbol{v}}$ presented in fig. 1.4 (top panel) reveals a field that varies slower than its equivalent of the linear theory. Structures are thus less compact and more extended.

Finally, fig. 1.5 (left panel) shows the distribution of δ_ρ against $\delta_{\boldsymbol{v}}$ on a cell by cell basis. It demonstrates that the divergence of the evolved velocity field $\delta_{\boldsymbol{v}}$ is not a good proxy for the true over-density δ_ρ. Indeed, the bulge of the distribution is wide, and even though these two quantities are not completely independent, a very large scatter is seen. There is thus no simple function such that $\delta_\rho = f(\delta_{\boldsymbol{v}})$, and the knowledge of $\delta_{\boldsymbol{v}}$ only partially constrains the range of the possible values of δ_ρ at the same position. Interestingly, the correlation between $\langle \delta_{\boldsymbol{v}}(\boldsymbol{x})\delta_\rho(\boldsymbol{x}+\boldsymbol{r})\rangle = \langle \delta_{\boldsymbol{v}}\delta_\rho\rangle(r)$ is very similar to $\zeta_\rho(r)$ for $r > 0$, as displayed in fig. 1.4. The knowledge of $\delta_{\boldsymbol{v}}(\boldsymbol{x})$ or the knowledge of $\delta_\rho(\boldsymbol{x})$ are thus put the same amount of constraint on the value of $\delta_\rho(\boldsymbol{x}+\boldsymbol{r})$.

The velocity field

Let us again start with a description of the velocity field as predicted by the linear theory. Like the over-density field, the components of the velocity field are normal, as can be seen in fig. 1.6. Their 95%, 99% and 99.9% percentiles read 460 km/s, 643 km/s and 843 km/s (these values are the same for the

[17] Statistically speaking, it theoretically diverges to $+\infty$ with the size of the sample.

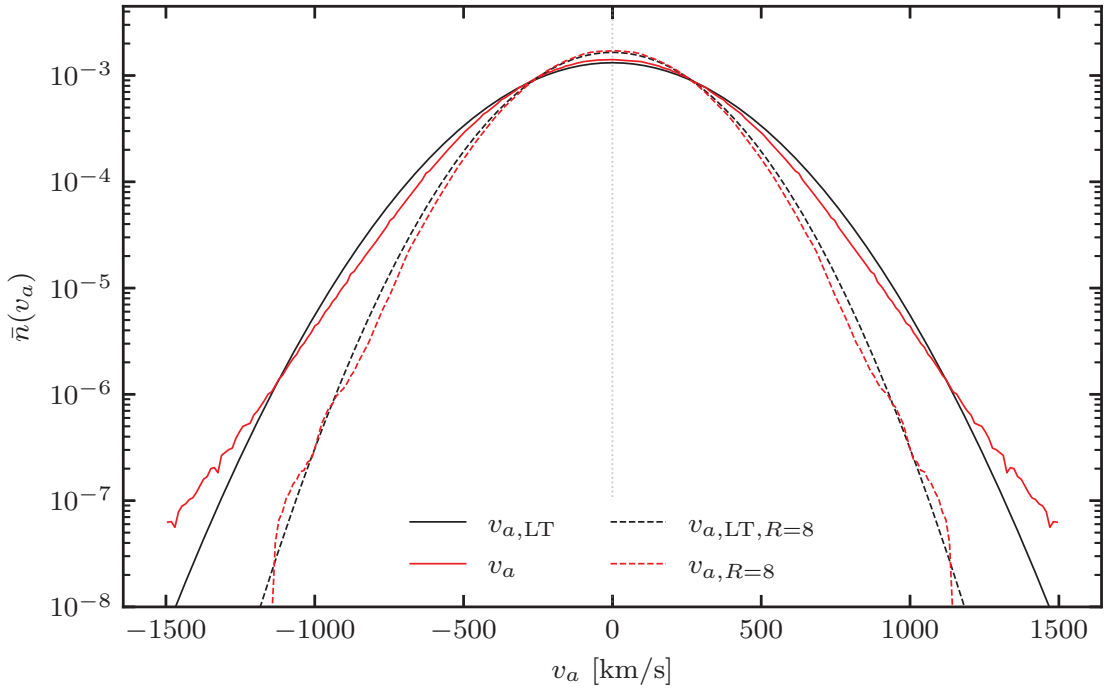

Figure 1.6: Distributions of the velocity field predicted by the linear theory (black) and of an evolved velocity field (in red). The statistics of the fields smoothed at $R = 8\,\mathrm{Mpc}/h$ are dashed.

5%, 1% and 0.1% percentiles). A detailed description of the correlation functions ψ_R and ψ_T plotted in fig. 1.4 (lower panel) is found in section 1.6.

The linear and evolved velocity field exhibit very similar distributions in fig. 1.6. The evolved velocity field is also symmetrical, which is a consequence of the hypothesis of isotropy. Only the tails slightly deviate from one another: the evolved velocity field has more high velocity cells than its linear equivalent. This is due to the gravitational collapse and the formation of compact clusters. The 95%, 99% and 99.9% percentiles of the evolved velocity field read 465 km/s, 672 km/s and 931 km/s. The deviation from the linear theory is thus minimal.

The correlations functions of both fields are also very similar. While the evolved over-density field is less correlated than the one of the linear theory, it is the opposite for the velocity fields. Indeed, for almost all sampled separations r, $\psi_R(r) > \psi_{R,\mathrm{LT}}(r)$ and $\psi_T(r) > \psi_{T,\mathrm{LT}}(r)$. This interesting and unexpected property comes from the large voids of evolved universes in which the velocity field is self-coherent.

The density-velocity correlation

The last question that can be answered with the one- and two-point statistics is the correlation between density and velocity fields. Figure 1.4 (lower panel) shows the cross-correlation between the over-density and the velocity fields.

Very similar observations to those made in section 1.8.1 can be drawn. Indeed, the same behavior as for the density-density correlations is observed: δ_ρ and $\boldsymbol{v}$ are the less correlated, and over the shortest distance while $\delta_{\boldsymbol{v}}$ and $\boldsymbol{v}$ shows a higher correlation and over larger distance than the linear theory itself. At $r = 1.9\,\mathrm{Mpc}/h$, the correlation between $\delta_{\boldsymbol{v}}$ and $\boldsymbol{v}$ suddenly drops below the predictions of the linear theory. It has to be noted that all these functions converge to $\zeta \to 0$ when $r \to 0$. The resolution of the grid does however not allow us to explore the very local behavior below separations $r < 1.9\,\mathrm{Mpc}/h$.

1.8.2 Smoothing the fields

The linear theory is deemed fit for the description of the early times of the Universe or at (very) large scales. Smoothing the fields should thus increase the agreement between the different statistics. We propose here a smoothing of $8\,\mathrm{Mpc}/h \approx 12\,\mathrm{Mpc}$.

The density field

The smoothing procedure brings all the distributions of fig. 1.3 closer. The linear theory predicts a normal law of standard deviation 0.46 with a 99.9% percentile at 1.4 for the smoothed over-density field. The upper tail of δ_ρ is cut and its 99.9% now read 2.9 (against 45.1 for the non-smoothed field). Only 60% of the volume is now under-dense (against 80% for the non-smoothed field). As for $\delta_{\boldsymbol{v}}$, both its upper and lower tails disappear: it now ranges from -0.8 to 2.0, its 0.1% and 99.9% percentiles (against -2.6 and 10.2 for the non-smoothed field). It is interesting to note that, like for δ_ρ, -1 becomez a lower bound. Almost exactly half of the volume is "under-dense".

The two point correlations of fig. 1.4 (top panel) are also very similar for the smoothed fields. All the functions are flat up to $10\,\mathrm{Mpc}/h$ and decrease slowly until $100\,\mathrm{Mpc}/h$. Interestingly, the linear theory predicts a drop of the two point correlation function of the density field between $70\,\mathrm{Mpc}/h$ and $100\,\mathrm{Mpc}/h$ which is not seen in the other fields. The peak at $110\,\mathrm{Mpc}/h$ is the BAO, stronger in the evolved field as in the linear theory.

As demonstrated in fig. 1.5 (right panel), the two definitions of the over-density δ_ρ and $\delta_{\boldsymbol{v}}$ converge when the fields are smoothed on large enough scales. Indeed, although there is some scatter left in the distribution of δ_ρ against $\delta_{\boldsymbol{v}}$, it is much smaller than in the non smoothed case, and for most points $\delta_\rho \approx \delta_{\boldsymbol{v}}$ is a good approximation. Moreover, the deviation between the two fields is continuous and smooth, and there exists a simple function f such that $\delta_\rho = f(\delta_{\boldsymbol{v}})$ gives a better fit than $\delta_\rho \approx \delta_{\boldsymbol{v}}$.

The velocity field

Smoothing the velocity field has a similar effect. As shown in fig. 1.6, it removes the tails of distribution of the evolved velocity field, which becomes nearly identical to that predicted by the linear theory. The 95%, 99% and 99.9% percentiles of the linear velocity field read 351 km/s, 483 km/s and 626 km/s while these same percentiles are 381 km/s, 548 km/s and 737 km/s for the evolved velocity field. The fields are not exactly statistically equivalent, but 90% of the cells (from 5% to 95% percentile) cover the same range of values.

The two-point correlations functions shown in fig. 1.4 (bottom panel) are also affected by the smoothing. There again, the convergence to similar statistics is striking, the curves sitting almost on top of each other.

The density-velocity correlation

As for the over-density field, smoothing brings the statistics together to a single behavior as demonstrated in fig. 1.4. This behavior very much resembles the one of the linear theory (see fig. 1.2 top right panel). Although close, the curves are not identical. The linear theory shows the most correlation, followed by the pair $\delta_{\boldsymbol{v}}$-$\boldsymbol{v}$ while δ_ρ-$\boldsymbol{v}$ remains less correlated. The peak of correlation remains somewhat shifted between the different pairs. It is interesting to see that the order between non-smoothed and smoothed has changed.

1.8.3 Conclusion of the comparison between evolved and linear fields

This discussion demonstrates very clearly using the one- and the two-points statistics that the evolved over-density field strongly differs from the linear theory. Indeed, low z universes are much clumpier and emptier than the linear theory predicts. Shifting the focus to the divergence of the evolved velocity field can somewhat reduce the tension between the two, but the physical meaning of this quantity is of lesser interest. Finally, the divergence of the evolved velocity and the over-density are not similar and only poorly correlated, the former thus cannot be used as a proxy for the latter.

It is very interesting to note that this deviation is much milder for the velocity field. Indeed, the one- and two-points statistics are very similar, with a slightly increased amplitude for the evolved field. Another remarkable conclusion is that the non-linear velocity field correlates over longer distances than its linear equivalent.

Smoothing appears as the best way to solve these tensions between the two. After a smoothing of $8\,\mathrm{Mpc}/h$, all the fields display rather similar statistics. Moreover, the divergence of the velocity field and the over-density converge. There is no saying at what scale the linear theory describes the (smoothed) evolved fields properly because such a scale itself would depend on the environment. The convergence with the smoothing is slow and a metric should be defined to quantify an quality threshold, however this discussion suggests that a value of $8\,\mathrm{Mpc}/h$ for the smoothing gives good results.

1.9 Reconstruction methods

1.9.1 An introduction to the problem of reconstruction

The problem of reconstruction as we understand it in this work, can be summed up as the following question: "How the distribution of matter in the Universe and its associated velocity field be inferred from observations of galaxies?"

In other words, a reconstruction is a map of the matter in the Local Universe. In theory, other sources of information than galaxies could be used, but in practice they are the only ones employed in this work. Although cosmography - making maps of the universe - is a goal in itself, it is not the only goal reconstructions can achieve: associated with an evolution model, they allow us to constrain physical models and cosmological parameters (see section 1.7.1 for more details). Reconstruction is thereby a important tool of cosmology.

Several methods and approaches have been developed since the 1990s, adapting to the quantity and the quality of the data and growing more and more capable as the power of computers increased. As of today, two main families of reconstruction methods are in use [18]. The first reconstructs the density field using only the redshift positions of galaxies. The second, discussed extensively in this work, reconstructs the velocity field using both the redshifts and the distances of galaxies, *i.e.* the peculiar velocities of galaxies. A similar, more detailed, discussion can be found in Strauss & Willick (1995).

1.9.2 Reconstruction of the density from redshifts surveys

A few words about the reconstruction methods from redshifts surveys are due to understand the motivations of using peculiar velocities. The redshift of a galaxy is quite easy to measure with a satisfactory precision, and many large redshifts surveys have been carried over the last decades (*e.g.* de Lapparent et al. (CfA; 1986) , York et al. (SDSS; 2000), Jones et al. (6dF; 2009), DESI Collaboration et al. (DESI; 2016)) while more are to come (Euclid; Laureijs et al., 2011). Some surveys are pencil like, going very deep in a restricted area of the sky; some display a flat selection function, covering a large angle in one direction but with a very restricted thickness; and some explore a larger solid angle (Strauss & Willick, 1995). Only the 2 Micron All Sky Survey (2MASS; Skrutskie et al., 2006) covers the whole sky up to about $160\,\mathrm{Mpc}/h$. As it appears here, the issue of the coverage of the sky by the data is the first limit of all reconstruction methods.

The two key ideas to create a reconstructions from redshift surveys are: (1) the redshift is a good proxy to the distance and (2) the distribution of galaxies is tightly correlated to the distribution of matter. The first point can be even corrected for: provided that a velocity field is constructed in the process, the exact distance can be estimated from the measured redshift (see eqs. (1.41), (1.42) and (1.47)). However, the second point may be the Achilles heel of these methods. This correlation, the galaxy bias, is subject to discussion (see section 1.3.3). When using this approach, one has no other solution than assuming a model – be it parameterized – to estimate the matter density field.

A second difficulty that these methods encounter is the incompleteness of redshift surveys. Indeed, because of technical limitations of the observations not all galaxies at all distance can be observed. The main bias is the absence in the catalogues of distant faint galaxies, the flux Malmquist Bias discussed in section 1.5.2. Extinction through dust or galaxies morphology and color evolution amongst other effects crossed with the complex selection function are many unknowns that can plague a survey, which is thus called incomplete. The galaxy density ρ_g cannot be well retrieved from it. A careful treatment of any survey is therefore needed before the application of these methods. The need for completeness also makes the merging of surveys difficult, as they have inherently different selection functions.

Very powerful methods have been developed recently in the AQUILA collaboration with the BORG code[19], an HMC sampler very close from the one discussed in this work (Jasche & Wandelt, 2013). Applied to an improved 2MASS catalogue, they reconstructed the non linear velocity and density field of the Local Universe within about $160\,\mathrm{Mpc}/h$ (Jasche & Lavaux, 2019). A similar approach is taken by the ELUCID collaboration (Wang et al., 2014, 2016).

[18] We ignore here the methods based on the weak gravitational lensing of images of galaxies, (*e.g.* Bartelmann & Schneider, 2001).

[19] https://www.aquila-consortium.org/method/borgpm.html

1.9.3 Reconstruction of the velocity field from peculiar velocities surveys

An alternative is the reconstruction of the velocity field from measurements of peculiar velocities. Instead of asking the question "where are galaxies?", assuming that there is matter where they are, it asks the question "where do galaxies go?", assuming there is matter where they converge. Galaxies – or any luminous object – is thus taken as a probe of the velocity field rather than of the density field.

There is no such thing as a peculiar velocity survey strictly speaking. Peculiar velocity catalogues are derived from pre-existing redshifts catalogues, from which all or part of the galaxies have seen their distance estimated, resulting in a catalogue where each entry has (at least) two values: a redshift and a distance (*e.g.* Willick et al., 1997; Tully et al., 2008, 2013; Tully et al., 2016; Tully et al., 2023). These can be converted into a catalogue of pairs distance – peculiar velocity (see section 1.4.6).

Such catalogues are not directly sensitive to the very complex aforementioned galaxy bias. Indeed, it does not matter if the galaxy is old or young, elliptic or spiral, in a dense or an empty region, as long as its velocity is measured. The completeness of the catalogue is thus much less of an issue. Furthermore, the velocity field is correlated over much larger distances than the density field, and therefore easier to constrain. This is even more true for the non-linear fields (see section 1.8).

Although the aforementioned reasons may be considered quite convincing, reconstructions from peculiar velocities also have their limitations. The main one is the scarcity and the poor quality of estimations of galaxies distances. Indeed, as discussed in section 1.1.2, indicators of distance are difficult to get and have errors ranging from 5% to 20%. This leads to a truly small signal to noise ratio for distant galaxies. Furthermore the error is log-normal, which complicates the interpretation and the use of the estimated peculiar velocities (see section 1.5.1).

As demonstrated throughout this work, the placing of the constraints is actually the core issue of these methods, even though the form that it takes depends on the method. The galaxy bias and the selection effects are thus not completely absent from the problem (Strauss & Willick, 1995).

Finally, the density field cannot be directly estimated from the velocity field. Measuring the later is thus intrinsically different from measuring the former. Note that this shortcoming is also present in the methods estimating the density field from redshifts surveys, which do not directly estimate the velocity field. Up to now, methods of reconstruction from peculiar velocities have been limited to the linear theory to link density and velocity field (see section 1.6). However, this model is a quite a strong simplification of the problem, as highlighted in section 1.8, which curbs the quality of the produced maps.

Reconstruction methods from peculiar velocities include the POTENT method (Bertschinger & Dekel, 1989; Dekel et al., 1999), the methods based on the Wiener Filter method detailed in the next section (Hoffman & Ribak, 1992; Ganon & Hoffman, 1993; Zaroubi et al., 1995; Zaroubi et al., 1999), while other methods include forward modeling (Lavaux, 2016; Graziani et al., 2019) and this work, chapters 2 to 4.

1.9.4 The Wiener Filter

Since a series of papers in the early 1990s, the Wiener Filter has become the tool of choice for the reconstruction of the velocity and density fields from peculiar velocities. This method has been (very) extensively described in the literature, in or outside of the context of reconstruction. (Ganon & Hoffman, 1993; Zaroubi et al., 1995; Zaroubi et al., 1999; Doumler, 2012; Doumler et al., 2013b; Graziani, 2018).

In this section, I would like to propose a simple mathematical approach ensuing from the discussion of section 1.6.

If two random variables X and Y are correlated, the knowledge of X influences directly the probability distribution of Y: $P(Y|X) \neq P(Y)$. Amongst other summary statistics, its mean might be shifted $\langle Y|X\rangle \neq \langle Y\rangle$ and its standard deviation affected $\sigma_{Y|X} \neq \sigma_Y$. This law can be computed in the context of a Multivariate Normal Distribution (MND) as described in section 1.6: this is what the Wiener Filter is about (Wiener, 1930). When fixing some of the variables of the MND, the others are affected, and it is possible to derivate their new conditional probability.

Wiener Filter and Gaussian process

Let's consider a spacial Gaussian process, *i.e.* a random field whose values sampled in space follow a MND depending on the positions where these values are sampled. First, let's fix n values of this field (*e.g.* we have measured them). These constraints are noted $\{c_i\}$ and situated at the positions $\{\boldsymbol{x}_i\}$. Note that $\boldsymbol{c} = \{c_i\}$ is a generic description of the constraints. In our case it may be the constrained density or peculiar velocity field. Now, we would like to recover the conditional probability of the m a

priori unknown values of the field $\{s_j\}$ (for sample) at the positions $\{\boldsymbol{y}_j\}$. Note that $\boldsymbol{s} = \{s_j\}$ represents the value of say peculiar velocity or density that is sampled in the reconstruction.

The probability law resulting from that process is written $P(\boldsymbol{s}|\boldsymbol{c}) = P(\{s_j\}|\{c_i\})$. In the next section, we will see that both the constraints and the samples can be values of the over-density or of the velocity field.

Remember that the value of the field at every point of space is correlated with the value at every other point through the two points correlation function, it is therefore possible to write the probability of the all $n+m$ values:

$$P(\boldsymbol{s},\boldsymbol{c}) = P(\boldsymbol{a}) = \frac{1}{(2\pi)^{(n+m)/2}|\boldsymbol{\Sigma}|}\exp\left(\frac{-1}{2}\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\right) \tag{1.102}$$

where $\boldsymbol{a}$ (for all) is the concatenation of n constraints $\{c_i\}$ and m samples $\{s_j\}$:

$$\boldsymbol{a} = (\boldsymbol{c},\boldsymbol{s}) = \left(c_1,\ldots,c_n,s_1,\ldots,s_m\right) \tag{1.103}$$

and $\boldsymbol{\Sigma}$ is the self-correlation matrix that can be split as such:

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Gamma} & \boldsymbol{\Delta}^T \\ \hline \boldsymbol{\Delta} & \boldsymbol{\Pi} \end{array}\right) \tag{1.104}$$

where

— $\boldsymbol{\Gamma} \in \mathcal{M}(n\times n)$ is the constraints-constraints correlation matrix: $\Gamma_{ij} = \langle c_i c_j\rangle$;

— $\boldsymbol{\Pi} \in \mathcal{M}(m\times m)$ is the samples-samples correlation matrix: $\Pi_{ij} = \langle s_i s_j\rangle$;

— $\boldsymbol{\Delta} \in \mathcal{M}(m\times n)$ is the constraints-samples correlation matrix: $\Delta_{ij} = \langle s_i c_j\rangle$.

The distribution of the sampled values alone ($\{s_j\}$), prior to the measurement of the constraints($\{c_i\}$), is also straightforward to write:

$$P(\boldsymbol{s}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Pi}|}\exp\left(\frac{-1}{2}\boldsymbol{s}^T\boldsymbol{\Pi}^{-1}\boldsymbol{s}\right). \tag{1.105}$$

It is an MND whose correlation matrix is the bottom right block of $\boldsymbol{\Sigma}$. The same goes for the constraints, prior to there own measurement:

$$P(\boldsymbol{c}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Gamma}|}\exp\left(\frac{-1}{2}\boldsymbol{c}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{c}\right). \tag{1.106}$$

The distribution $P(\boldsymbol{s})$ is however modified by the injection of the knowledge of the constraint. The demonstration is out of the scope of this work, but is rather simple. The distribution of the samples, posterior to the constraining the field reads

$$P(\boldsymbol{s}|\boldsymbol{c}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Lambda}|}\exp\left(\frac{-1}{2}(\boldsymbol{s}-\boldsymbol{s}_{\rm WF})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{s}-\boldsymbol{s}_{\rm WF})\right), \tag{1.107}$$

$$\boldsymbol{s}_{\rm WF} = \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{c}, \tag{1.108}$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Pi} - \boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}^T. \tag{1.109}$$

The constraints shift the mean of the samples to a possibly non-null value and modifies their correlations. From a more physical point of view, the new mean $\boldsymbol{s}_{\rm WF}$ represents the expected values of the samples given the constraints. The residuals $\boldsymbol{s}_{\rm R} = (\boldsymbol{s}-\boldsymbol{s}_{\rm WF})$ hold the remaining randomness and follow an MND, covering less volume than the prior (*i.e.* they are constrained). It is very interesting to note that the correlation matrix of the residuals $\boldsymbol{\Pi}$ *does not* depend on the values taken by the constraints, but rather on their positions and the ones of the samples. Similarly, the expected values of the samples $\boldsymbol{s}_{\rm WF}$ and the constraints $\boldsymbol{c}$ are linked through is a simple matrix $\boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}$ which depends only on the positions of both constraints and samples.

Application to the linear theory

The application of the previous theorem to the fields of linear theory is straightforward. Indeed, the only needed ingredients are the self- and cross-correlations functions between constraints and samples.

Let us take a simple example. Suppose the over-density field has been measured at $\boldsymbol{x}$, $\delta(\boldsymbol{x}) = \delta_c$, where the subscript c refers to "constraint". We would like to know the value of the over density field at another position $\boldsymbol{y}$, $\delta(\boldsymbol{y}) = \delta_s$. We note $r = |\boldsymbol{y} - \boldsymbol{x}| = |\boldsymbol{y}|$. A priori, the probability of δ_s is

$$P(\delta_s) = \frac{1}{\sqrt{2\pi}\sigma_\delta} \exp\left(\frac{-\delta_s^2}{2\sigma_\delta^2}\right) \tag{1.110}$$

where σ_δ is the standard deviation of the over-density field. Following eqs. (1.107) to (1.109), the distribution of the sample given the constraint reads:

$$P(\delta_s|\delta_c) = \frac{1}{\sqrt{2\pi}\lambda_\delta} \exp\left(\frac{-(\delta_s - \delta_{\text{WF}})^2}{2\lambda_\delta^2}\right) \tag{1.111}$$

where

$$\delta_{\text{WF}}(\boldsymbol{y}) = \frac{\xi(r)}{\sigma_\delta^2}\delta_c, \qquad \lambda_\delta^2(\boldsymbol{y}) = \sigma_\delta^2 - \frac{\xi^2(r)}{\sigma_\delta^2}. \tag{1.112}$$

It appears that the value taken by $\delta_s(\boldsymbol{y})$ solely depends on the density-density two point correlation function $\xi(r)$ plotted in fig. 1.1. As the Universe is isotropic, only the distance between the sample ($\{s_j\}$) to the constraint ($\{c_i\}$) matters, not its direction.

When $r \to 0$, that is to say when the sampled value is taken very close from the constraint, $\xi(r) \to \sigma_\delta^2$ and thus $\delta_{\text{WF}} \to \delta_c$, the sampled value converges to the constraint. The standard deviation of the residual tends to zero: the result is certain. The posterior distribution becomes $P(\delta_s|\delta_c) \to \mathbb{I}(\delta_s - \delta_c)$.

Reciprocally, when $r \to \infty$, $\xi(r) \to 0$ and thus $\delta_{\text{WF}} \to 0$, while $\lambda_\delta \to \sigma_\delta$. The volume affected by the constrained is limited, and the further away the unknown value is sampled, the more its posterior distribution converges to its prior behavior: $P(\delta_s|\delta_c) \to P(\delta_s)$. In other words far from the region where constraints exist, the Wiener Filter returns the mean field.

Note that there is no configuration where the value of δ_{WF} exceeds the one of δ_c. Indeed, $\xi(r) \leq \sigma_\delta^2$ for all $r \geq 0$. This shows a very fundamental property of The Wiener Filter: it is conservative. It does not add any power to the signal but only discards it. It can thus not create a false positive, which is very important for the rest of this work. When several constraints are set, the fields evolves from one value to the next "as smoothly as possible", going to the null field in the absence of constraint. Similarly, the constrained standard deviation λ_δ is bound between 0 (of course) and σ_δ. There is thereby only a gain of information, or worse case no gain at all ($\lambda_\delta \to \sigma_\delta$).

Errors in the measurements

In certain aspects of cosmology, errors may dominate the signal. The methodology of the Wiener Filter is needed to formalize the possible uncertainties on the measurements.

Errors can be added as an additional MND field. The errors $\boldsymbol{e}$ on the constraints $\boldsymbol{c}$ are thus:

$$P(\boldsymbol{e}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Upsilon}|} \exp\left(\frac{-1}{2}\boldsymbol{e}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{e}\right), \tag{1.113}$$

such that $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{e}$ where $\boldsymbol{e} = \{e_i\}$ is the "error field" drawn at the location of each constraint. The distribution of the constraints with the now added errors $\boldsymbol{c}'$ is easy to recover: the sum of two MNDs is another MND with a correlation matrix:

$$\langle c_i' c_j' \rangle = \langle (c_i + e_i)(c_j + e_j) \rangle = \langle c_i c_j \rangle + \langle e_i e_j \rangle + \langle e_i c_j \rangle + \langle e_j c_i \rangle \tag{1.114}$$

In practice the errors are considered uncorrelated between measurements, *i.e.* $\boldsymbol{\Upsilon}$ is diagonal, and uncorrelated with the signal, *i.e.* $\langle e_i c_j \rangle = 0$. This simplifies the calculus. Recall from equation 1.106 that $\boldsymbol{\Gamma}$ is the constraint auto-correlation function. The standard deviation of the error MND is $\varepsilon_i^2 = \langle e_i e_i \rangle$. This implies

$$\boldsymbol{\Gamma}' = \boldsymbol{\Gamma} + \boldsymbol{\Upsilon} \tag{1.115}$$

$$= \boldsymbol{\Gamma} + \text{diag}(\varepsilon_1^2, \ldots, \varepsilon_n^2). \tag{1.116}$$

The Wiener Filter and the residuals correlation matrix of eqs. (1.108) and (1.109) are modified in the presence of these errors

$$\boldsymbol{s}_{\text{WF}} = \boldsymbol{\Delta}\left[\boldsymbol{\Gamma}'\right]^{-1}\boldsymbol{c}', \tag{1.117}$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Pi} - \boldsymbol{\Delta}\left[\boldsymbol{\Gamma}'\right]^{-1}\boldsymbol{\Delta}^T. \tag{1.118}$$

Let us continue the simplified problem proposed in the previous section and add an error drawn from $e \sim \mathcal{N}(0, \varepsilon)$ on the constraint $\delta_c' = \delta_c + e$. How is the probability of the sample affected? Equation (1.111) holds but both δ_{WF} and λ_δ change:

$$\delta_{\text{WF}}(\boldsymbol{y}) = \frac{\xi(r)}{\sigma_\delta^2 + \varepsilon^2}\delta_c', \qquad \lambda_\delta^2(\boldsymbol{y}) = \sigma_\delta^2 - \frac{\xi^2(r)}{\sigma_\delta^2 + \varepsilon^2}. \tag{1.119}$$

The behavior the sampled value of the field remains unchanged when $r \to \infty$. The convergence for $r \to 0$ is however different:

$$\delta_{\text{WF}}(\boldsymbol{x}) = \frac{\delta_c'}{1 + \varepsilon^2/\sigma_\delta^2}, \qquad \lambda_\delta^2(\boldsymbol{x}) = \frac{\varepsilon^2}{1 + \varepsilon^2/\sigma_\delta^2}. \tag{1.120}$$

On the contrary to the previous case, the sampled value does not converge to the constraint δ_c' when $r \to 0$, but rather to smaller value $\delta_{\text{WF}}(\boldsymbol{x}) \leq \delta_c'$. In parallel, the standard deviation does not fully cancel $\lambda_\delta(\boldsymbol{x}) \geq 0$. Indeed, if the constraint is tainted by error, the true value of the field at this point is not necessarily the one measured. In presence of many measurements with well modeled errors, the Wiener Filter is a better estimator of the real value of the field δ_c than δ_c' is: $|\delta_s(\boldsymbol{x}) - \delta_c| < |\delta_c - \delta_c'|$. Being conservative, the Wiener Filter proposes a value that is less that the one measured, but also leaves room for variation. Yet, $\lambda_\delta(\boldsymbol{x})$ cannot exceed σ_δ, even for very large uncertainties. Indeed, if the constraint is very weak (or virtually absent) the prior behavior takes over, and again $P(\delta_s|\delta_c') \to P(\delta_s)$.

Application of the Wiener Filter to the problem of reconstruction from peculiar velocities

Now that all the theoretical foundations of the Wiener Filter methodology have been laid, its application to the problem of this work can be explained in a few words.

In practice, the constraints are observations of peculiar velocities $c_i' = v_{r,i}^{\text{obs}}(\boldsymbol{x}_i)$ of which there are n_{obs} observations. As previously discussed, each measurement is a position - radial peculiar velocity pair. The coefficients of the correlation matrix $\boldsymbol{\Gamma}$ are computed with the velocity-velocity correlation tensor $\boldsymbol{\psi}$ whose expression is given in eq. (1.93). To build the modified matrix $\boldsymbol{\Gamma}'$, the observation dependent uncertainties $\varepsilon_i = \sigma_{v,i}^{\text{obs}}$ are added to the diagonal of $\boldsymbol{\Gamma}$ following eq. (1.116). The expression of $\sigma_{v,i}^{\text{obs}}$ is given in eq. (1.62).

The reconstruction is the value of the over-density field sampled on a regular grid $s_j = \delta(\boldsymbol{y}_j)$ of which there are $m = n_{\text{grid}}^3$ points. The self-correlation matrix $\boldsymbol{\Pi}$ is thus computed using the density-density correlation function ξ of eq. (1.74). Finally, the constraint-sample correlation matrix $\boldsymbol{\Delta}$ is in this case built with the density-velocity cross-correlation function ζ presented in eq. (1.88).

Extending the $\langle\ \rangle$ notation to Matrices, the mean field and the residual correlation matrix can be simply summarized as

$$\delta_{\text{WF}} = \langle \delta v_r \rangle \left[\langle v_r v_r \rangle + \langle \varepsilon\varepsilon \rangle\right]^{-1} \boldsymbol{v}_r', \tag{1.121}$$

$$\Lambda_\delta = \langle \delta\delta \rangle - \langle \delta v_r \rangle \left[\langle v_r v_r \rangle + \langle \varepsilon\varepsilon \rangle\right]^{-1} \langle v_r \delta \rangle\,. \tag{1.122}$$

The reader should bear in mind that the values of $\boldsymbol{v}_r' = \{v_{r,i}^{\text{obs}}\}$ are not the values of the radial velocity fields at $\{\boldsymbol{x}_i\}$ but rather estimations of peculiar velocities of galaxies set at their estimated positions. The fact that these estimations are inherently tainted with errors and biases makes a conceptual difference. Moreover, this is what truly prevents the Wiener Filter to perfectly reconstruct the fields, as detailed in section 1.9.5.

The velocity field can be recovered from the over-density field using eq. (1.38) or eq. (1.99). Reciprocally, the 3 components of the velocity field can be separately estimated directly from the constraints

$$v_{a,\text{WF}} = \langle v_a v_r \rangle \left[\langle v_r v_r \rangle + \langle \varepsilon\varepsilon \rangle\right]^{-1} \boldsymbol{v}_r, \qquad a = x, y, z, \tag{1.123}$$

and the over-density derived later with eq. (1.38) or eq. (1.99). This approach allows for the reconstruction of the tidal modes, but its computational cost is about three times greater.

Computing the Wiener Filter and the Constrained Realizations

In the context of reconstructions, two objects are of interest. First, the field recovered from the Wiener Filter, which is the mean expectation of the density distribution in the Universe given our constraints. However, this very conservative estimation of the over-density field is not consistent with the statistics of the linear theory: its power spectrum (or its Fourier Transform, the spatial correlation function) is not the $\mathcal{P}(k)$ (respectively $\xi(r)$) presented in fig. 1.1. Because of the scarcity of the data and the uncertainties, the fields has less power, mostly at high frequencies.

Thus, the *constrained* realizations of the density field ($\delta_{\rm CR}$) can be written as the sum of the mean field ($\delta_{\rm WF}$) and the residual field $\delta_{\rm R}$ *i.e.* $\delta_{\rm CR} = \delta_{\rm WF} + \delta_{\rm R}$, where the residual field follows the statistics of eq. (1.107). Indeed, these so called constrained realizations (CR) are fields that are compatible with the constraints and that have the intrinsic statistics dictated by the linear theory. In other words, they are plausible Universes whereas the Wiener Filter only gives an average picture. CRs are notably used for the generation of initial conditions for constrained cosmological simulations (see section 1.7.1). Generating these CRs is however non trivial.

If the mathematics are fairly easy to write in the general case, applying them to a practical case can be indeed more challenging. Modern applications of these methods deal with tens of thousands of constraints $n_{\rm obs} \gtrsim 10\,000$, and reconstruct the over-density fields on grids of up to $n_{\rm grid}^3 = 1024^3$.

Analytical calculus is prevented by the inversion of the $\mathbf{\Gamma}'$ matrix, which is virtually not tractable for $n_{\rm obs} > 2$. Furthermore, the reconstruction of the field in finite boxes (*i.e.* periodic Universes) prevents the use of formula of the correlation functions as written in eqs. (1.74), (1.88), (1.91) and (1.92). Numerical methods and computers thus have to be employed. The inversion of the $\mathbf{\Gamma}'$ matrix remains an obstacle, as it is grows in $\mathcal{O}(n_{\rm obs}^3)$. To avoid the inversion, a Cholesky decomposition can be performed, and then the problem $\boldsymbol{c}' = \mathbf{\Gamma}'\boldsymbol{\eta}$ is solved for $\boldsymbol{\eta}$, which is also $\mathcal{O}(n_{\rm obs}^3)$. This obstacle will not cease to grow as the size of the data increases.

Another bottleneck of the numerical application is the computation of the $\mathbf{\Delta}$ matrix and the product of $\mathbf{\Delta}$ with the intermediary n-vector $\boldsymbol{\eta} = \left[\mathbf{\Gamma}'\right]^{-1}\boldsymbol{c}'$. Indeed, in practical applications, the constaints-samples correlation matrix $\mathbf{\Delta}$ can have as many as $n_{\rm grid} \times n_{\rm obs} \approx 10^9 \times \cdot 10^4 = \cdot 10^{13}$ coefficients. All of these coefficients have to be computed one by one using the proper correlation function evaluated on each pair constraint-sample. Then the vector-matrix product $\mathbf{\Delta}\boldsymbol{\eta}$ has to be performed, which means another 10^{13} multiplications and additions. Fortunately this can be easily parallelized and the matrix $\mathbf{\Delta}$ does not need to be stored as a whole[20].

However, the computation and the storage of the residual correlation matrix $\mathbf{\Lambda}$ is out of reach: it can reach 10^{18} coefficients. Even it were to be computed, it would be impossible to use in order to produce realizations of the residual field[21]. This makes the direct construction of a CR on this scale impossible.

Hoffman & Ribak (1992) designed the so-call Hoffman-Ribak algorithm that allows the generation of CRs while avoiding the construction and the use of this very cumbersome matrix:

1. A random realization of the over-density field $\delta_{\rm RR}$ is generated;
2. The velocity field $\boldsymbol{v}_{\rm RR}$ is derived from $\delta_{\rm RR}$ using eq. (1.99);
3. The random field to constrain is evaluated at the positions of the observational constraints yielding the $\boldsymbol{c}_{\rm RR}$;
4. An error consistent with the observational error is added on the constraints $\boldsymbol{c}'_{\rm RR} = \boldsymbol{c}_{\rm RR} + \mathcal{N}(0, \boldsymbol{\varepsilon})$;
5. The CR is evaluated using

$$\delta_{\rm CR} = \delta_{\rm RR} + \Delta\left[\mathbf{\Gamma}'\right]^{-1}(\boldsymbol{c}' - \boldsymbol{c}'_{\rm RR}). \tag{1.124}$$

In other words, the Hoffman-Ribak algorithm creates a CR from a RR by replacing the Wiener Filter field of the random constraints by the one of the observational constraints. This algorithm truly opened the doors to the field of constrained simulations described in section 1.7.1.

[20] The vector-matrix product can done block by block or line by line.

[21] Which would notably involve a Cholesky decomposition, whose cost grows as $\mathcal{O}(n_{\rm grid}^9)$!

1.9.5 The limits of the Wiener Filter

The methodology of the Wiener Filter is directly derived from the properties of the Gaussian fields and the associated MNDs. From the point of view of the linear theory, it thus is a perfectly adapted tool. Moreover, it can be easily proved that it yields the optimal field in this context.

Yet, when applied to measurements of peculiar velocities, the Wiener Filter fails to capture the complexity of the observations. First, the errors on the velocity is closer to log-normal than normal. Consequently, in absence of a correction, it produces fields plagued by the log-normal bias. Furthermore, in order to write all the correlation matrices necessary for its application, the positions of the constraints need to be known. This, again, does not correspond to the reality of the data: the positions of the galaxies, and thus the constraints, are subject to large errors and possible biases.

There are then two possibilities: setting the constraints at their redshift positions knowing that these are biased by the redshift space distortion; or setting them at their directly observed position even though these can have dramatic errors and positional biases. In the very early applications of these methods (up to CF1, (Tully et al., 2008)), the data was limited to the close neighborhood and a direct measurements of distance with small errors could still be used. Starting with CF2 (Tully et al., 2013) and later releases, as data extended deeper in the sky, using direct distances became impossible due to the larger errors and biases. While some groups chose to use redshifts positions, other developed methods that prepare the data before their use by the Wiener Filter (Sorce, 2015; Hoffman et al., 2021).

These methods suffer two inherent problems. First, there is not one unique way to "unbias" the data. Simplifications have to be made depending on exactly which bias the method aims to correct, or how it interprets the data – which is again not as trivial as it sounds. These differences from a method to another make the comparison and the objective assessment of their quality difficult and subjective to some extent. This is the topic of chapter 3, where we compared one of these unbiasung methods (The Bias Gaussianization correction; Hoffman et al., 2021) to the algorithm we developed and propose in chapter 2.

The second issue is that these methods ultimately have to pick one distance per point, which then is fed to the Wiener Filter. Of course galaxies have only one true position however since this is cant be recovered from such error prone data, it is more sensible to assign galaxies a positional probability. That is the guiding philosophy employed here.

These profound flaws of the Wiener Filter methodology motivated the development of a novel approach that bypasses both the splitting of the unbiasing and Wiener Filter and encloses both the treatment of the data and the reconstruction of the fields in a single self-consistent probabilistic algorithm. That is the main content of this thesis and explained in detail in the following chapter.

Chapter 2

Hamiltonian Monte Carlo Reconstruction of the Local Environment (HAMLET)

The aim of the present chapter is to present a new numerical approach to the MCMC algorithm of Lavaux (2016) and by Graziani et al. (2019). The Gibbs sampling algorithm suffers from a slow convergence and is very CPU inefficient in comparison with the WF/CRs formalism. A very considerable improvement is presented here by a numerical implementation of the Hamiltonian Monte Carlo (HMC) sampling technique. The HMC sampling technique was applied before to reconstruct the Large Scale Structure (LSS) from a galaxy redshifts survey by Jasche & Wandelt (2013); Jasche & Lavaux (2019) and it is applied here for the first time to a galaxy velocities survey. Both the Gibbs sampling based MCMC reconstruction implementations and the present HMC one are constructed within the same theoretical framework and to the extent that they are applied to the same data aiming at the same resolution they should yield very similar results. Our motivation here is to considerably improve the numerical efficiency of MCMC algorithm, aiming in particular to achieve the numerical resolution needed for setting up constrained initial conditions for numerical simulations of the local universe (*e.g.* Sorce et al., 2014; Sorce et al., 2017; Sorce & Tempel, 2017; Libeskind et al., 2020). The HAmiltonian Monte carlo reconstruction of the Local EnvironmenT (HAMLET) code is presented here and is tested against a mock catalog.

While using the HMC sampling over the Gibbs sampling accelerates greatly the exploration, it remains computationally very heavy. In order to reach convergence in a reasonable time, HAMLET is employs GPUs, which enables a computational speed up of several orders of magnitude in time. The code takes advantage of the very powerful python library `Tensorflow`, which permits python code to run on CPU(s), GPU(s) and even be compiled[1] with a Just In Time compilation library.

The chapter starts with a few discussion on Bayesian inference (section 2.1.1), followed by a description of its application to the reconstruction of the LSS (section 2.2). The HMC method is presented in details after an introduction of the more "classical" Monte Carlo method and a the key tools and concept to understand their functioning (section 2.3). Finally, a the application of the HAMLET code to a mock velocity survey and its analysis conclude this chapter (section 2.5).

This is work is published as Valade et al. (2022). It has been augmented by section 2.1.1, added during the redaction of this thesis to improve the readability. A few editorial modifications have been made to insure the global consistency of the thesis.

2.1 Bayesian inference and Monte Carlo exploration

2.1.1 Bayesian inference

The Bayesian probabilistic approach is adopted here, according to which the PDF represents one's confidence in the certainty of the knowledge of the values of some parameters/variables. These can be either observable quantities that can be measured or theoretical parameters whose values are to be inferred. A very complete description of the use Monte Carlo methods in the context of Bayesian

[1] Python is an interpreted language, which makes it orders of magnitude slower at run-time than compiled languages, like C, C++ or Fortran.

inference and forward modeling can be found in Betancourt (2017). This section is limited to the key ideas necessary to the understanding of this work.

Bayesian inference in the context of forward modeling consists in recovering the probability density function of these parameters, given fixed observations or constraints, and a model. We denote the parameters $\boldsymbol{q} = \{q_a\}$, such that $\boldsymbol{q} \in \mathcal{Q}$, where $\mathcal{Q} \subset \mathbb{R}^n$ is the so-called parameters space, and n is the number of parameters (*i.e.* the dimensionality of the parameter space). Using Bayes' theorem, one can write the posterior distribution of the parameters given the data $P(\boldsymbol{q}|\text{data})$ as a product of the likelihood $P(\text{data}|\boldsymbol{q})$ and the prior $P(\boldsymbol{q})$:

$$P(\boldsymbol{q}|\text{data}) \propto P(\text{data}|\boldsymbol{q})P(\boldsymbol{q}). \tag{2.1}$$

The likelihood function can be understood as the probability that the data stem from these parameters, while the prior is the probability of these parameters in the model, prior to any measurement. The posterior is in itself not of great interest, but it is the core function necessary to the derivation of any observable or summary statistics:

$$\langle F(\boldsymbol{q})|\text{data}\rangle = \int_{\mathcal{Q}} F(\boldsymbol{q})P(\boldsymbol{q}|\text{data})\mathrm{d}^n\boldsymbol{q} \tag{2.2}$$

where F represents any function of the parameters[2]. For instance, taking $F(\boldsymbol{q}) = \boldsymbol{q}$ simply yields the mean of the posterior of the posterior distribution; taking $F(\boldsymbol{q}) = (\boldsymbol{q} - \langle\boldsymbol{q}|\text{data}\rangle)^2$ yields the variance. More complex functions can be used (*e.g.* the Fourier Transform as below). Any other observable quantity of the model can be reproduced, and potentially be compared against other sets of data or to produce predictions.

2.1.2 Monte Carlo exploration of the parameters space

However, and this is the core issue of forward modeling, eq. (2.2) cannot be simply evaluated, even for trivial functions of the parameters $F(\boldsymbol{q})$. First, unless the posterior has a very simple form, analytical calculus cannot be done. Secondly, as the number of parameters grows, numerical evaluation by discretization of the parameters space becomes quickly too expensive. For instance, if the parameters space is regularly sampled on a grid of m cells in each dimension, the number of evaluations $F(\boldsymbol{q})P(\boldsymbol{q}|\text{data})$ to perform grows with the dimensionality of the parameters space as $\mathcal{O}(m^n)$.

It is possible to proceed with a maximization of the posterior. Indeed, very efficient algorithms have been designed even for highly dimensional parameters spaces, and it is apriori reasonable to approximate a distribution to its most probable value. This procedure is known under the name of "likelihood optimization" as priors are often constant, *i.e.* the only source of information comes from the data and not the model. Yet, this approach gives only limited results: the higher the dimensionality of the parameters space is, the less the peak itself weights in the integral of eq. (2.2), *i.e.* $P(\boldsymbol{q}_{\max}|\text{data})\mathrm{d}^n\boldsymbol{q} \to 0$ when $n \to \infty$. In parallel, the higher the dimensionality of the parameters space is, the higher the weight of the volume *around* the peak is. This volume, called the "typical set", contains the parameters such that $P(\boldsymbol{q}|\text{data})\mathrm{d}^n\boldsymbol{q}$ is the most significant (Betancourt, 2017).

The only way forward is to perform a Monte Carlo exploration of the parameters space, and more specifically of the typical set (Metropolis et al., 1953; Betancourt, 2017). The purpose of a Monte Carlo method is to iteratively build a series of n_s sets of parameters $\{\boldsymbol{q}^s\}$ which follow the posterior probability law

$$\boldsymbol{q}^s \sim P(\boldsymbol{q}^s|\text{data}). \tag{2.3}$$

The evaluation of the eq. (2.2) then simplifies to

$$\langle F(\boldsymbol{q})|\text{data}\rangle = \lim_{n_s\to\infty} \frac{1}{n_s}\sum_{s=1}^{n_s} F(\boldsymbol{q}_s) \tag{2.4}$$

The first Monte Carlo method that have been designed is the so-called Metropolis-Hasting sampling (Metropolis et al., 1953). The Hamiltonian Monte Carlo approach presented here (as almost any Monte Carlo method) is a variation of it. The Metropolis Hasting method is an iterative process constructed

[2]Note that $F(\boldsymbol{q})P(\boldsymbol{q}|\text{data})$ has to be integrable on the parameters' space.

on two building blocks. First, it it creates a Markov Chain: a series of random states, which has the particularity that the probability distribution of a state *only* depends on the previous state:

$$P(\boldsymbol{q}^{s+1}|\boldsymbol{q}^s, \dots, \boldsymbol{s}^0) = P(\boldsymbol{q}^{s+1}|\boldsymbol{q}^s). \tag{2.5}$$

Secondly, it relies on the so-called *fine balance* that insures that eq. (2.3) is respected: the probability of moving to a new step has to follow

$$P(\boldsymbol{q}^{s+1}|\boldsymbol{q}^s)P(\boldsymbol{q}^s) = P(\boldsymbol{q}^{s+1}, \boldsymbol{q}^s) = P(\boldsymbol{q}^s|\boldsymbol{q}^{s+1})P(\boldsymbol{q}^{s+1}) \quad \Longleftrightarrow \tag{2.6}$$

$$\frac{P(\boldsymbol{q}^{s+1}|\boldsymbol{q}^s)}{P(\boldsymbol{q}^s|\boldsymbol{q}^{s+1})} = \frac{P(\boldsymbol{q}^{s+1})}{P(\boldsymbol{q}^s)}. \tag{2.7}$$

The Metropolis Hasting algorithm reads:

Initialization

— an initial state $\boldsymbol{q}^0$ is set or drawn at random;

Loop over s until n_s states are kept

1. a candidate state is drawn at random from a probability law $\boldsymbol{q} \sim C(\boldsymbol{q}|\boldsymbol{q}^s)$;
2. a random number $u \sim \text{Uniform}[0, 1]$ is drawn;
3. if $u < P(\boldsymbol{q})/P(\boldsymbol{q}^s)$, the step is accepted, $\boldsymbol{q}$ becomes the new current state $\boldsymbol{q}^{s+1}$;
4. else, the step is rejected, the current state remains $\boldsymbol{q}^s$.

By construction, this approach creates a Monte Carlo Markov Chain: step (1) insures that the $\boldsymbol{q}^{s+1}$ only depends on $\boldsymbol{q}^s$ while step (3) insures that the fine balance is respected. As the initial guess may not necessarily be "likely", the first steps are often discarded, until the chain reaches the typical set. This number is fixed by the operator and is regarded as a technical parameter in the rest of this work.

A key issue of Monte Carlo methods is the proposition of a "good" candidate $\boldsymbol{q} \sim C(\boldsymbol{q}|\boldsymbol{q}^s)$. If $C(\boldsymbol{q}|\boldsymbol{q}^s)$ tends to yield unlikely candidates, they are systematically rejected. This is quantified by the acceptance rate

$$\alpha = \frac{\text{number of accepted steps}}{\text{number of proposed steps}}, \tag{2.8}$$

which takes low values in this peculiar case. The loop converges very slowly and a large amount of computational power is wasted.

To avoid this pitfall, $\boldsymbol{q}$ can be draw in the close neighborhood of $\boldsymbol{q}^s$, which has been already accepted and is thus likely. This may however lead to a second issue: if the successive states of the Monte Carlo chain are too close, a very large number of steps is needed to cover the entire typical set, which also may result in a very inefficient use of the computational power. This effect is strongly exacerbated by the increasing of the dimensionality of the parameters space, this is the "curse of dimensionality". Moreover, an incomplete exploration of the typical set may lead to a dramatically biased estimation of eq. (2.4).

A balance thus has to be found between the acceptance rate and the length of the chain. The variations between Monte Carlo algorithms often concern the mechanism of proposition of a candidate state $\boldsymbol{q} \sim C(\boldsymbol{q}|\boldsymbol{q}^s)$, so as to maximize the distance between successive states while also insuring a high acceptance rate. The more a Monte Carlo maximizes these two quantities, the more *efficient* it is said to be.

2.2 Application to the large scale structure

The aim of the Bayesian analysis, in the present context, is the construction of the posterior PDF of the distances of the data points, the set of the Fourier modes that define the density and velocity fields, and σ_{NL} (to be defined later), given Cosmicflows-like data and under the assumption of the ΛCDM model. More specifically, the posterior PDF is calculated within the linear approximation of the ΛCDM model, with the Planck parameters (Planck Collaboration et al., 2016). A mild non-linear correction is added to the linearly calculated velocity field whose amplitude is controlled by σ_{NL}, whose magnitude is to be estimated as well.

The Bayesian posterior PDF is numerically evaluated by means of the MCMC sampling. Given the posterior PDF the mean and variance of the desired density and velocity fields are readily calculated, much in the same way as in the WF/CRs formalism (Zaroubi et al., 1995). The main difference between the WF/CRs and the MCMC cases is that in the former the posterior PDF is assumed to be know analytically and in the later it is evaluated numerically.

2.2.1 Bayesian posterior PDF and likelihood function

Bayes' theorem states that the posterior PDF, of the model given the observed data, is the product of the conditional probability of the data given the model, hence the likelihood function, times the prior probability of the model, normalized by the evidence. In the language of Bayes' theorem the distribution of the true distances of the data points, the ensemble of the Fourier modes and $\sigma_{\rm NL}$ consist of the multi-parameter model whose parameters are to be estimated given the data. The ΛCDM cosmological model provides the framework within which that multi-parameter model is constructed. Neglecting here the evidence the posterior PDF is:

$$P\big(\Delta_k, \mathcal{D}, \sigma_{\rm NL}|\mathcal{M}, \mathcal{Z}\big) \propto L\big(\mathcal{M}, \mathcal{Z}|\Delta_k, \mathcal{D}, \sigma_{\rm NL}\big) P\big(\Delta_k, \mathcal{D}, \sigma_{\rm NL}\big) \tag{2.9}$$

Here $\mathcal{D} = \{d_i\}$ (proper distances of the data points, $i = 1, ..., n$ where n is the number of data points), $\Delta_k = \{\delta_{\boldsymbol{k}}\}$ (the ensemble of the Fourier modes, where $\delta_{\boldsymbol{k}} = FT\left(\delta(\boldsymbol{r})\right)$ is the Fourier transform of the fractional over-density field $\delta(\boldsymbol{r})$) and $\sigma_{\rm NL}$ are the output variables/parameters to be estimated. We denote $m = |\Delta_k|$ the number of complex Fourier modes. $\mathcal{Z} = \{z_i\}$ (observed redshifts of the data points) and $\mathcal{M} = \{\mu_i\}$ (observed distance moduli) are the input observed data. $L\big(\mathcal{M}, \mathcal{Z}|\Delta_k, \mathcal{D}, \sigma_{\rm NL}\big)$ is the likelihood function and $P\big(\Delta_k, \mathcal{D}, \sigma_{\rm NL}\big)$ is the prior.

The errors on the observed data points are assumed to be independent, and so are the errors on the redshift and distance modulus of a given galaxy. All observational errors are assumed to be normally distributed. The likelihood function is thus the product of n likelihood functions, one for each of the i constraints used. Furthermore, given that the redshifts and distance moduli are independent, the i-th likelihood function is a product of two independent likelihood functions denoted here by $L_i^{v_r}$, associated eventually with the velocity of the data point (eq. (2.15) below) and L_i^{μ}. The likelihood function is:

$$L\big(\mathcal{M}, \mathcal{Z}|\Delta_k, \mathcal{D}, \sigma_{\rm NL}\big) = \prod_i^n L_i^{\mu}(\mu_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL}) L_i^{v_r}(z_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL}) \tag{2.10}$$

2.2.2 The Likelihood function: distances

Given a Gaussian error $\sigma_{\mu,i}$ on the measurement, L_i^{μ} is written:

$$L_i^{\mu}(\mu_i|d_i) = \frac{1}{\sqrt{2\pi\sigma_{\mu,i}^2}} \exp\left(\frac{-\big(\mu_i - \mu(d_i)\big)^2}{2\sigma_{\mu,i}^2}\right) \tag{2.11}$$

where

$$\mu(d) = 5\log_{10}\left(\frac{d_L(d)}{1\,\mathrm{Mpc}}\right) - 25, \tag{2.12}$$

where $d_L(d)$ is the luminosity distance associated with the proper distance d

$$d_L(d) = d\big(1 + z_{\cos}(d)\big) \tag{2.13}$$

and to 2nd order the cosmological redshift $z_{\cos}$ corresponding the proper distance d is given by:

$$z_{\cos}(d) = \frac{2}{3\Omega_m}\left(1 - \sqrt{1 - \frac{3\Omega_m H_0}{c} d}\right). \tag{2.14}$$

Here Ω_0 is the matter density parameter, H_0 is Hubble's constant evaluated at the present epoch and c is the speed of light.

Note that this writing of the luminosity distance ignores the peculiar redshift of the observed constraint as it appears in eq. (1.49). This simplification has been brought in Graziani et al. (2019), whose model we try to match in this chapter. Neglecting the peculiar redshift simplifies the correlation between Δ_k and $\mathcal{D}$, making the posterior possibly smoother and its exploration faster.

2.2.3 The Likelihood function: velocities

The velocity of the i-th data point is related to its observed redshift via

$$v_{r,i}^{\rm obs} = c\frac{z_i - z_{\cos}(d_i)}{1 + z_{\cos}(d_i)}. \tag{2.15}$$

Given the ensemble of Fourier modes, Δ_k, the assumed cosmological 3D velocity field is given by the following inverse Fourier transform,

$$\boldsymbol{v}(\boldsymbol{r}|\Delta_k) = FT^{-1}\left(-\mathrm{i}H_0 f(\Omega_m)\frac{\boldsymbol{k}}{k^2}\delta_{\boldsymbol{k}}\right), \tag{2.16}$$

where $f(\Omega_m)$ is the linear growth factor. One should note here that $\boldsymbol{v}(\boldsymbol{r}|\Delta_k)$ is the velocity field predicted by the linear theory from a given over-density field $\delta(\boldsymbol{r})$.

The evaluation of eq. (2.16) is in practice not direct. Indeed, when employing the FFT algorithm, the values of the field can only be computed on the nodes of the regular grid. Yet, the constraints are irregularly placed in space, and an interpolation from the grid is thus necessary.

In this work we limit ourselves to a (tri-)linear interpolation of the components of the fields $v_{x,y,z}$. This apparently technical detail has some physical consequences. First, for a points that is not on a node of grid, the interpolated value is different from the true value of the discrete Fourier transform. This error is complex to described, as it is depends on the positions of the constraint in a cell and is different for each Fourier mode. Still, a general trend can be drawn: the higher the frequency of the modes, the less smooth the field between two nodes is, the more error is injected by the linear interpolation. This error also leads the correlation functions to be violated within a cell and between neighbor cells – which is not a direct concern for our algorithm, but is worth being noted. As the velocity field has a correlation length much bigger than the effective resolution of our grids, this error is minimal. It would however not be the case when interpolating the over-density field, which shows much more small-scale variations.

Other interpolation methods can be used. For instance, Lavaux (2016) proposes a Taylor-Fourier interpolation. The value of a point in a cell is computed from the Taylor approximation of field on the closest grid node. This method is equivalent to the linear approximation when the expansion is carried out to the first order. It surpasses the linear approximation at the second order, in which case ten FFTs need to be computed (instead of one), which significantly affects the computational cost of the interpolation.

Our aim here is to reconstruct the LSS from the grouped version of a Cosmicflows-like catalog of galaxy velocities. A grouped catalog means here a catalog in which all galaxies belonging to a group or cluster of galaxies are collapsed onto one data point. The grouping acts as a smoothing process of the internal virial velocities and thereby serves as a filter of non-linear velocities. The following crude approximation is introduced here to account for the residual non-linear component of the observed data. The full velocity field is assumed to include a non-linear component:

$$\boldsymbol{v}^{\rm full}(\boldsymbol{r}) = \boldsymbol{v}(\boldsymbol{r}|\Delta_k) + \boldsymbol{v}^{\rm NL}(\boldsymbol{r}). \tag{2.17}$$

The residual component, $\boldsymbol{v}^{\rm NL}(\boldsymbol{r})$, is assumed to constitute a white noise with a variance given by:

$$\sigma_{\rm NL}^2 = \left\langle (\boldsymbol{v}^{\rm NL})^2 \right\rangle. \tag{2.18}$$

The likelihood function for the observed redshift is readily written here in terms of the observed velocity:

$$L_i^{v_r}(z_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL}) = \frac{1}{\sqrt{2\pi}\kappa_i}\exp\left(\frac{-\left(v_{r,i}^{\rm obs} - \boldsymbol{v}(d_i\hat{\boldsymbol{r}}_i|\Delta_k)\cdot\hat{\boldsymbol{r}}_i\right)^2}{2\kappa_i^2}\right) \tag{2.19}$$

where $\hat{\boldsymbol{r}}_i$ is the unit vector in the direction of the i-th data point and

$$\kappa_i^2 = \sigma_{\rm NL}^2 + \frac{\sigma_{cz,i}^2}{(1 + z_{\cos}(d_i))^2}. \tag{2.20}$$

2.2.4 Priors

The elements of the model under consideration are the Fourier modes, the distribution of the distances of the data points and $\sigma_{\rm NL}$. Following the steps of Graziani et al. (2019), we split the joint prior into marginal priors:

$$P(\Delta_k, \mathcal{D}, \sigma_{\rm NL}) = P(\Delta_k)P(\mathcal{D})P(\sigma_{\rm NL}). \tag{2.21}$$

The Fourier modes are evaluated on a discrete grid which is written here symbolically as $\{\boldsymbol{k}_j\}$, where $j = 1, ..., m$ where m is the number of the Fourier modes. The prior of the ensemble of the Fourier modes is:

$$P(\Delta_k) = \prod_{j=1}^{m} \frac{1}{2\mathcal{P}(k_j)} \exp\left(\frac{-|\delta_{\boldsymbol{k}_j}|^2}{\mathcal{P}(k_j)}\right) \tag{2.22}$$

Here $\mathcal{P}(k)$ is the ΛCDM power spectrum at wave number k. Note that this probability law is the same as eq. (1.82), but it applies to two amplitude *squared*. Writing it under this form simplifies and thus accelerates its computation.

Writing the prior on the distances is complicated and somewhat ad hock. As discussed in section 1.5, biases can arise from a wrong derivation of this probability, namely the homogeneous in in-homogenous Malqmist biases. In the mean time, this prior function has to take into account the selection effects, that are impossible to properly describe for the Cosmicflows catalogues, as each one gathers and merges different survey, each one having their own specificities.

We opt for a simple description based on the fact that the redshift distance, d_z is a good proxy to the actual distance for all but the very nearby data points:

$$d_z = \frac{c}{H_0} \int_0^z \frac{\mathrm{d}z'}{\sqrt{\Omega_m(1+z')^3 + (1-\Omega_m)}}, \tag{2.23}$$

which is an application of eq. (1.41) to the observed redshift. This distance estimator is thus subject to the redshift space distortion (RSD).

A histogram of the distribution of the redshift distances, $N_{d_z}^{\rm obs}$ is constructed from the distribution of the observed redshifts.

$$P(\mathcal{D}) = \prod_{i=0}^{n} N_{d_z}^{\rm obs}(d_i). \tag{2.24}$$

To account for the RSD, this histogram is smoothed with a Gaussian kernel of width $\sigma_v/H_{100} \approx 3\,\mathrm{Mpc}/h$. The binning and the smoothing lead to what seems to be a good proxy to the histogram of the distribution of the true distances. Note that using the measurements of redshift as an estimator has the upside of partially modeling selection effects and partially accounting for the in-homogeneous Malmquist bias by pulling distances to fall in higher-density shells. However, the exact limitations and biases created by this approach have yet to be studied and understood.

2.2.5 Shape of the posterior

The exact shape of the posterior distribution is, needless to say, unknown. Yet, a rough estimation of its properties can be done before exploration with Monte Carlo methods, which is of great help for the choice of the sampling method and for the estimation of meta-parameters (see section 2.3.3).

Fourier modes of the over-density field

The formalism adopted for the creation of linear over-density fields yields an prior distribution for the real and the imaginary components of each Fourier mode, written in eq. (1.81). We also know from previous works with Wiener-Filter / Constrained Realization method (*e.g.* Doumler et al., 2013b) that low-frequency modes are more constrained than high frequency modes. We thus expect the posterior to be roughly normal for the Fourier modes of the field, with a standard deviation dictated by the power spectrum for high-frequency modes and a reduced standard deviation for low-frequency modes. Correlation between the low-frequency modes is also expected.

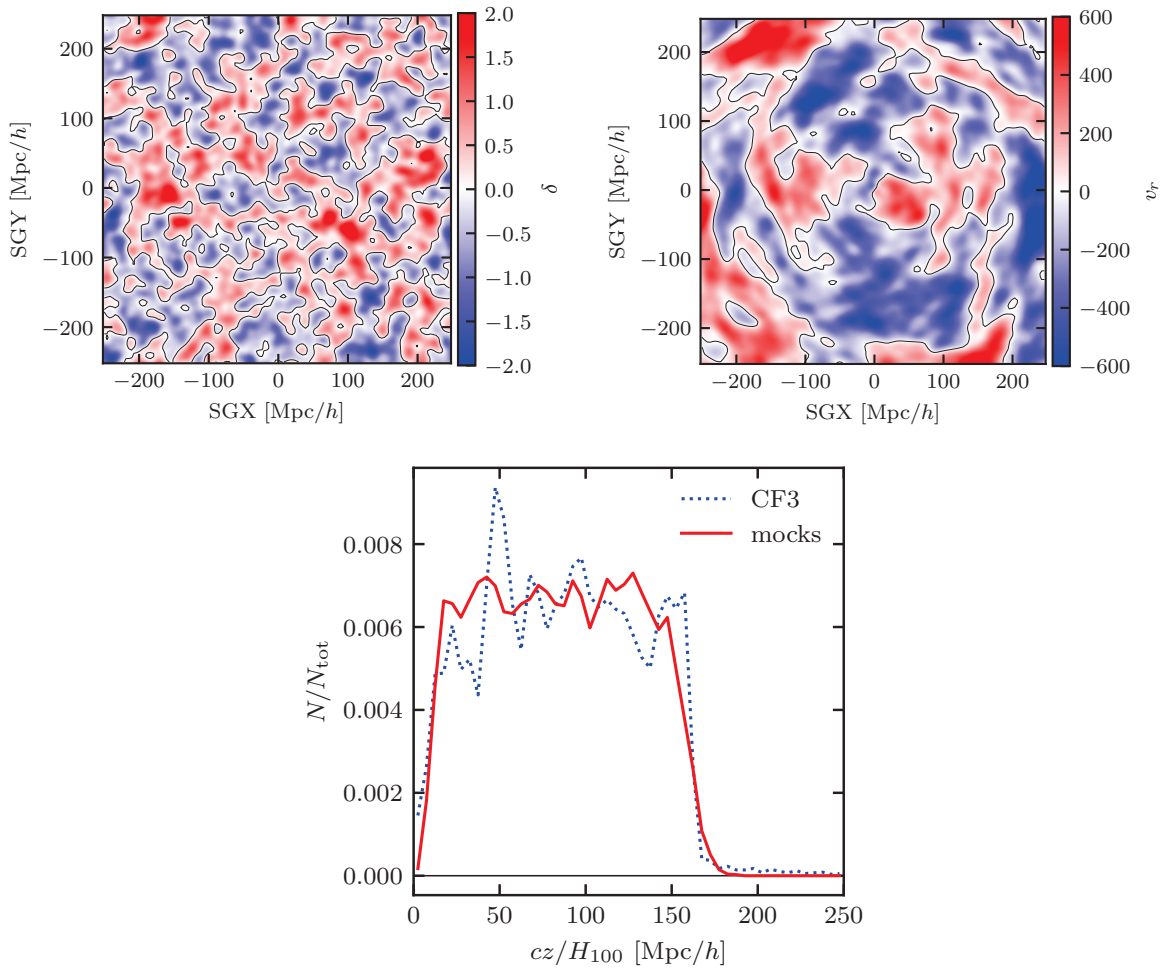

Figure 2.1: **Top left:** a slice through the over-density (δ) of the target field. **Top right:** the radial component of the velocity fields (v_r) of the target field. The mock observer is located at the origin of the coordinate system. The color bars present the color coding of the presented fields. Velocities are in units of km/s. The contour lines corrrespond to the zero values of the two fields. The target field is Gaussian smoothed with a kernel of $5\,\mathrm{Mpc}/h$. **Bottom:** the "selection function" *i.e.* the distribution of the distances for the sample of points used as constraints. For comparison we also show CF3.

Distances of the probes

Here again, the form of the posterior can be easily outline prior to its exploration. The informative value of the observation of the distance moduli is close from null, it is thus better to start from the redshifts. Section 1.4.7 highlights that the true distance of each constraint is found in the neighborhood of its redshift distance, with a scatter due to its radial peculiar velocity. Marginalizing over the later – a priori unknown – we expect the distances to follow a normal law centered on the redshift with a standard deviation lower or equal to $\sigma_v/H_{100} \lesssim 3\,\mathrm{Mpc}/h$. As the velocity field is expected to be well constrained at large scales, a constant shift from the redshift is predictable.

General properties

The posterior is general expected to be smooth, virtually mono-modal and roughly gaussian in any direction. Correlations between parameters are expected, but nothing indicates complex structures or stiffness in the posterior. The only challenge brought by this posterior is its extremely high dimensionality: $n_p = 2m + 2n + 1$. In this work, n_p reaches several millions (from $2 \cdot 10^6$ to $16 \cdot 10^6$) parameters.

The HAMLET algorithm is tested against a CF3-like survey drawn from a linear random realization of the density and velocity fields, constructed within the framework of the ΛCDM model. The selection of the data points and the assignment of the observational errors replicate the selection and the errors

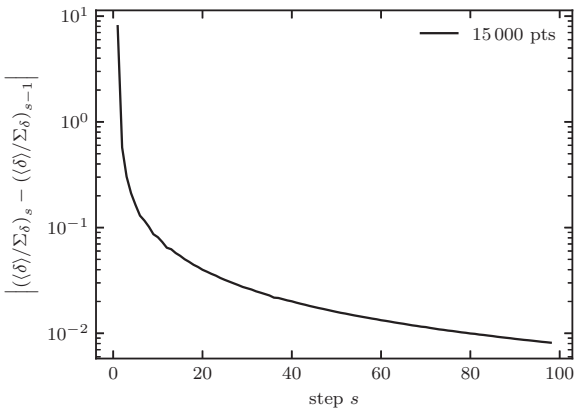

Figure 2.2: Differences between successive partial estimations of the signal over noise ratio field as a function of the number of steps used to estimate it. All mocks converge in a very identical way.

of the actual CF3 data. Our aim here is to test the the HAMLET algorithm and its performance in the ideal case where in the limit of perfect data - densely, homogeneously and and isotropically sampled and with negligible errors - the HAMLET should accurately recover the input density and velocity fields (cf. Graziani et al., 2019).

2.3 Hamiltonian trajectories in phase space

The detail of the Hamiltonian Monte Carlo (HMC; Hoffman & Gelman, 2011; Neal, 2011; Betancourt, 2017) is complex and out of the scope of this thesis. Only the key general elements of the HMC are given here. In this section we will describe the HMC method through its application to our model, for the reconstruction of the large scale structure.

The parameters of the model are denoted here by:

$$\boldsymbol{q} = \{q_a\} = \left(\delta^R_{k,1}, \ldots, \delta^R_{k,m}, \delta^I_{k,1}, \ldots, \delta^I_{k,m}, d_0, \ldots, d_n, \sigma_{\rm NL}\right). \tag{2.25}$$

Here, the real and imaginary components of the complex $\delta_k = \delta^R_k + i\delta^I_k$ are denoted as separate parameters. For the sake of the clarity of the presentation we define here the posterior function as $\Pi(\boldsymbol{q}) = P\left(\Delta_k, \mathcal{D}, \sigma_{\rm NL} | \mathcal{M}, \mathcal{Z}\right)$.

A Hamiltonian system is defined by means of associating $\Pi(\boldsymbol{q})$ with a potential of classical particles, $\Psi(\boldsymbol{q})$,

$$\Psi(\boldsymbol{q}) = -\ln \Pi(\boldsymbol{q}). \tag{2.26}$$

The parameters $\{q_a\}$ are assumed to consist a set of canonical coordinates. These are supplemented by auxiliary quantities, referred to as their associated momenta, $\boldsymbol{p} = \{p_a\}$ and a "mass matrix" M. The dynamics of this Hamiltonian system is governed by the Hamiltonian,

$$H\left(\boldsymbol{p}, \boldsymbol{q}\right) = \frac{1}{2}\boldsymbol{p}^T M^{-1} \boldsymbol{p} + \Psi(\boldsymbol{q}) \tag{2.27}$$

The equations of motions of the $\boldsymbol{q}$ and $\boldsymbol{p}$ are given by Hamilton equations:

$$\begin{aligned} \frac{\mathrm{d}p_a}{\mathrm{d}t} &= -\frac{\partial H}{\partial q_a} = -\frac{\partial \Psi(\boldsymbol{q})}{\partial q_a}, \\ \frac{\mathrm{d}q_a}{\mathrm{d}t} &= \frac{\partial H}{\partial p_a} \end{aligned} \tag{2.28}$$

Considering these Hamiltonian system as a many body system, probability distribution function of $\boldsymbol{q}$ and $\boldsymbol{p}$, $\Pi(\boldsymbol{q}, \boldsymbol{p})$ is related the the Hamiltonian via:

$$\Pi(\boldsymbol{q}, \boldsymbol{p}) \propto \exp\left(-H(\boldsymbol{q}, \boldsymbol{p})\right) = \Pi(\boldsymbol{q}) \exp\left(-\frac{1}{2}\boldsymbol{p}^T M^{-1} \boldsymbol{p}\right). \tag{2.29}$$

This is a key result. There is no cross-correlation between the distributions of the coordinates and the momenta in the joint PDF $\Pi(\boldsymbol{q}, \boldsymbol{p})$. It follows that the Hamiltonian trajectories properly sample the desired posterior PDF of the parameters of the model under consideration.

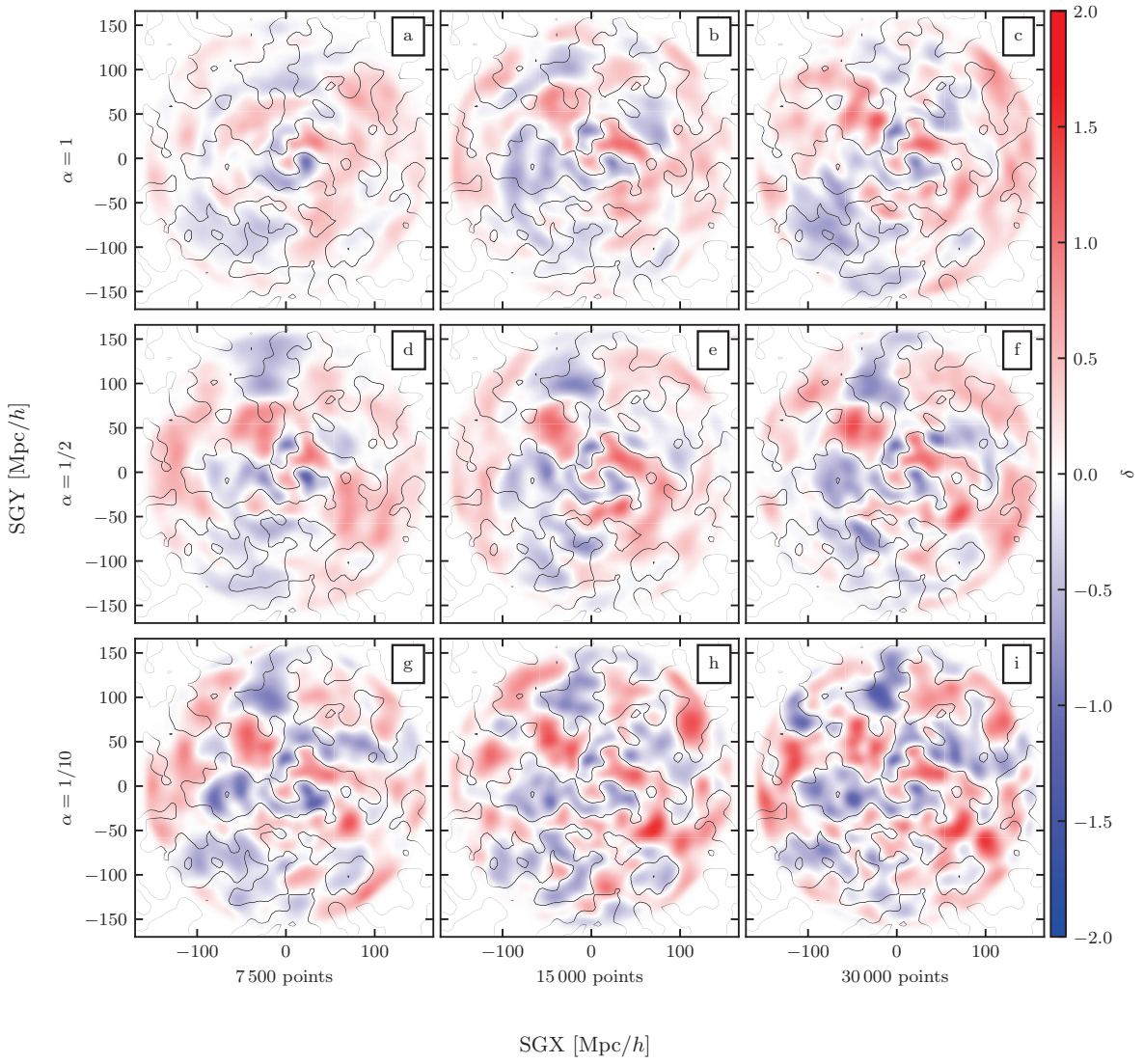

Figure 2.3: The conditional mean δ field given CF3-like mock data and the ΛCDM prior model. The field is Gaussian smoothed with a kernel of $5\,\mathrm{Mpc}/h$. The plots show color maps of a slice of the δ field. The color bar indicates the color coding of δ and the contour lines correspond to the zero level of the target field. The frames correspond to the different mock catalogs dented by $(N/10^4, \alpha)$, where N is the number of data points and α controls the level of the observational errors, $\sigma_\mu = \alpha\sigma_\mu^{\mathrm{CF3}}$. The nine mock data are: a. (0.75, 1.0), b. (1.5, 1.0), c. (3., 1.0); d. (0.75, 0.5), e. (1.5, 0.5), f. (3., 0.5); g. (0.75, 0.1), h. (1.5, 0.1), i. (3., 0.1). Frame **b** corresponds to the actual CF3 data in terms of the number of data points and the magnitude of the errors.

2.3.1 Construction of HMC chains

A chain starts with the coordinates and momenta randomly drawn and serve as the initial conditions for the integration of equations of motion (eq. (2.27)). These are integrated over a pseudo time τ. The final position of that trajectory in the $\{\boldsymbol{q}, \boldsymbol{p}\}$ phase space, is the candidate state. If the endpoint failed the modified Metropolis-Hastings acceptance rule (step (3), see section 2.1.2)

$$u < \frac{P(-\boldsymbol{p}, \boldsymbol{q})}{P(\boldsymbol{p}, \boldsymbol{q})}, \qquad u \sim \mathrm{Uniform}[0, 1], \tag{2.30}$$

the integration starts all over with the same initial coordinates but with new randomly drawn momenta. If the endpoint passed the acceptance rule, that trajectory becomes a step along the chain. The next step starts with the coordinates $(\boldsymbol{q})$ accepting the final coordinates from the last step and the momenta

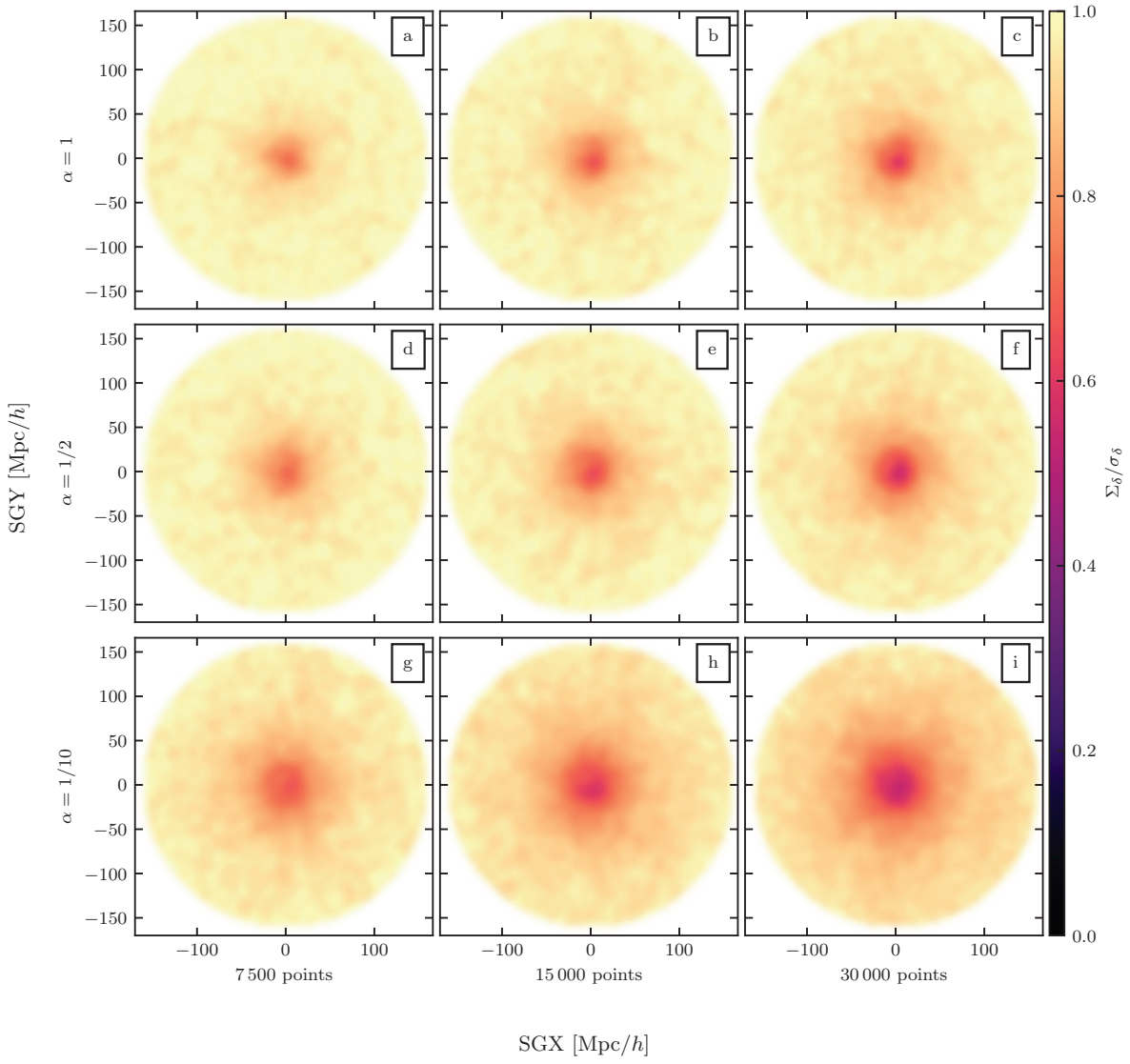

Figure 2.4: Color maps of the local constrained variance normalized by the cosmic variance of the δ field, $\Sigma_\delta(\boldsymbol{r})/\sigma_\delta$. The conventions and structure of fig. 2.3 are followed here.

($\boldsymbol{p}$), on the other hand, are again randomly drawn. The final positions of the successive trajectories form the chain.

2.3.2 Integrating the HMC trajectories

If the integration can be done analytically, the Hamiltonian framework ensures an acceptance rate of 1, namely all candidate states are accepted, even when candidates are far from the current position. However, for most problems, analytical integration is impossible and numerical solvers must be used. The integrator of choice is the Leapfrog algorithm, which ensures ergodicity, namely the conservation of the Hamiltonian. This ensures that the error introduced when computing the trajectory depends only on the integration step size and not on the number of integration steps. In other words, the use of the Leapfrog algorithm yields stable trajectories. In practice, this stability is limited by the numerical precision of the derivatives $\left(\frac{\partial \ln\ \Pi(\boldsymbol{q})}{\partial q_a}\right)$. Thus, the HMC is only applied to models where these derivatives can be computed analytically, which is the case for this work.

Note that trajectory length is simply the product of the step size and the number of steps used in the (leapfrog) integration. In this context, the acceptance rate for a given state is only a function of the step size. Therefore the step size can be tuned in order to obtain a given acceptance rate. Studies

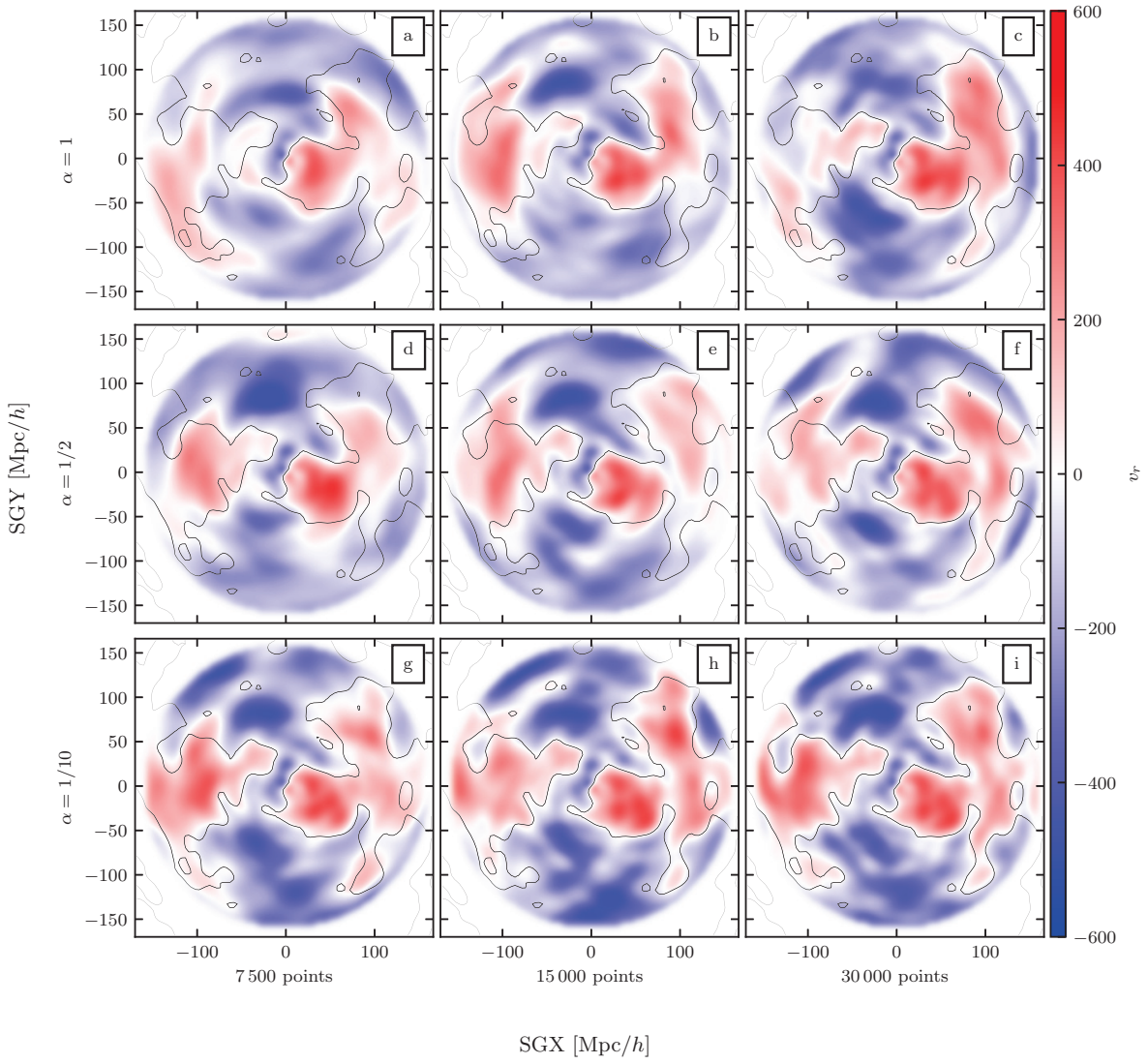

Figure 2.5: Same as fig. 2.3 but for the radial component of the velocity field.

in the literature (*e.g.* Hoffman & Gelman, 2011) advocate that an acceptance rate of 0.65 (used here) provides an optimal balance between computational resource usage and exploration. Hoffman & Gelman (2011) introduced the "Dual Averaging" method that dynamically tunes the step size to reach any given acceptance rate. The step size set to achieve this acceptance rate depends on the complexity of the problem: the more complex the problem and the more correlated the variables, the smaller the step size must be.

Secondly, since the integration is ergodic, every trajectory ultimately returns to the initial state (to within integration error). This is problematic since such closed orbits return a final state identical to the initial state resulting in no additional knowledge of the parameter space (and a waste of computational resources in the process). Therefore it is absolutely critical that the integration is halted after a designated number of steps, specifically chosen such that the candidate state is the furthest away from the initial position. At this maximum, the trajectory starts turning back towards the initial state. Depending strongly on the initial momentum, this value is different for each trajectory. Several methods have been developed to automatically tune this parameter, the most well known being the No U-Turns (NUTS) algorithm proposed in Hoffman & Gelman (2011), whose detail is out of the scope of this paper. NUTS and the "Dual averaging" technique can be used together.

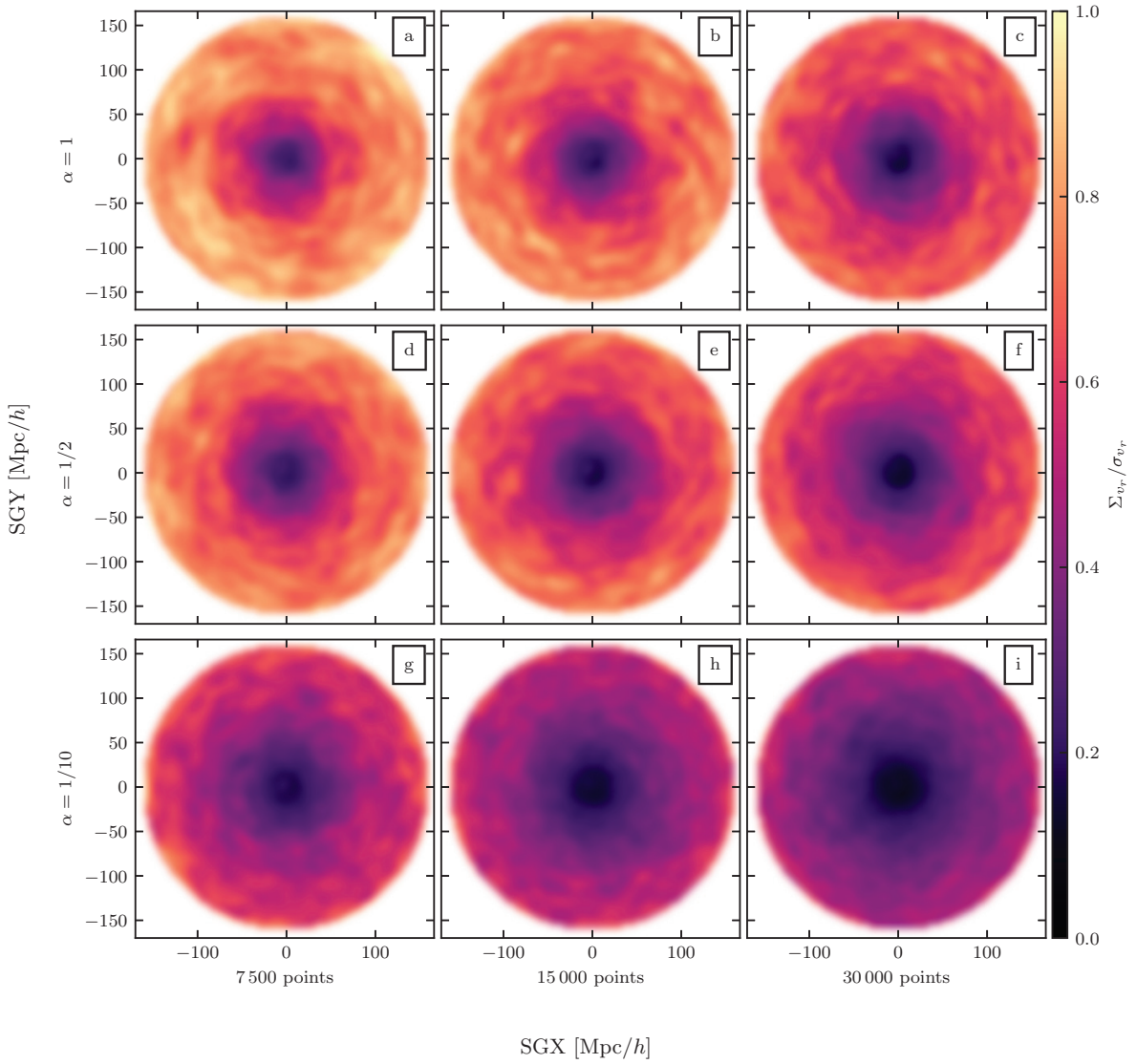

Figure 2.6: Same as fig. 2.4 but for the radial component of velocity field ($v_{\rm r}(\boldsymbol{r})$). The constrained and cosmic variances are evaluated for $v_{\rm r}(\boldsymbol{r})$.

2.3.3 The Mass matrix

Consider the classical dynamics described by eq. (2.28) - a trajectory evolves from a random position in the multi-dimensional phase space towards a local minimum of the potential $\Psi(\boldsymbol{q})$. In the presence of dissipative forces, the trajectory would reach a local minimum of the potential and stay there. For a Hamiltonian system whose energy is conserved, the trajectory oscillates around the local minimum with an amplitude dictated by the energy of the system and its mass. As the energy of each trajectory is set by the random choice of its initial momentum. The statistics of these initial momenta are encoded in the mass matrix. The selection of a mass matrix influences heavily the efficiency of the exploration and the rapidity of the convergence. Asymptotically however, it does not bias nor modify the result. Even though there is theoretically no optimal choice of mass matrix, using the covariance of the parameters is the canonical approach. This covariance matrix is however a priori not know and has to be estimated.

We base the coefficients of the mass matrix on our estimation of the shape of the posterior PDF described in section 2.2.5. In the absence of knowledge about the cross-correlation between parameters, we set the non diagonal coefficients of the matrix to zero. Furthermore, the posterior being close from normal in all direction, using a correlation matrix as mass matrix is almost optimal. The choice of the size of the other parameters (limited to $\sigma_{\rm NL}$ in this work) has to be estimated more freely.

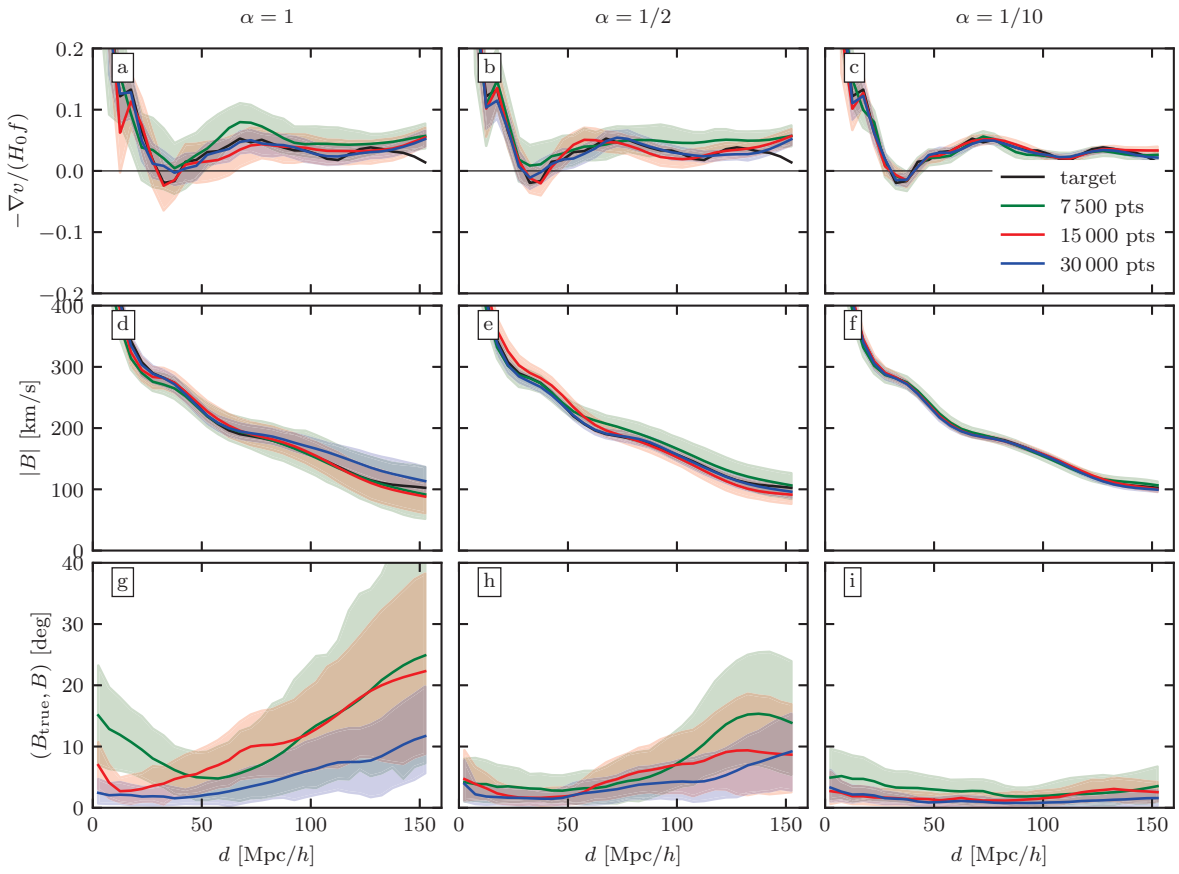

Figure 2.7: Top row: The monopole moment (namely the mean fractional over-density; upper panel), the amplitude (middle panel) and alignment (with the target; lower panel) of the dipole moment (namely the bulk velocity) are shown shown for $\alpha = 1, 1/2$ and $1/10$ (columns left to right) and for $N = \big(0.75$ (green), 1.5 (red) and 3.0 (blue)$\big) \times 10^4$, where N is the number of data points. The mean (solid lines) and the scatter are calculated over the ensemble of HMC steps. The moments of the target field are shown in black.

$$M^{-1}_{ab} = I_{ab} \times \begin{cases} \mathcal{P}(k_a)/2 & \text{if } a \leq m, \\ \mathcal{P}(k_{a-m})/2 & \text{if } m < a \leq 2m, \\ \sigma_v^2/H_0^2 & \text{if } 2m < a \leq 2m+n, \\ 1 & \text{if } a = 2m+n+1, \end{cases} \tag{2.31}$$

where I is the identity matrix and k_a is the wavenumber of the Fourier mode $\delta_{k,a}$. The mass matrix of eq. (2.31) is used here in the general case, where the estimation is done given the data and the prior.

More complex mass matrices could be used in the future, where some cross-correlation coefficients. It is however impossible to date to encore the full correlation matrix of the parameters space, as it contains $(2m+n+1)(2m+n)/2 \approx 10^{14}$ coefficients for our application, which exceeds by orders of magnitude the amount of memory available in modern GPUs[3].

2.4 Technical Implementation

An HMC algorithm computationally outperforms the more traditional Metropolis - Hastings and Gibbs sampling algorithms. This is also the case with the problem of the reconstruction of the LSS from a CF3-like catalogue, addressed here and by Graziani et al. (2019) which used the Gibbs sampling

[3]Encoded on 4-bytes float, this matrix would occupy close to 400 000 GB of memory.

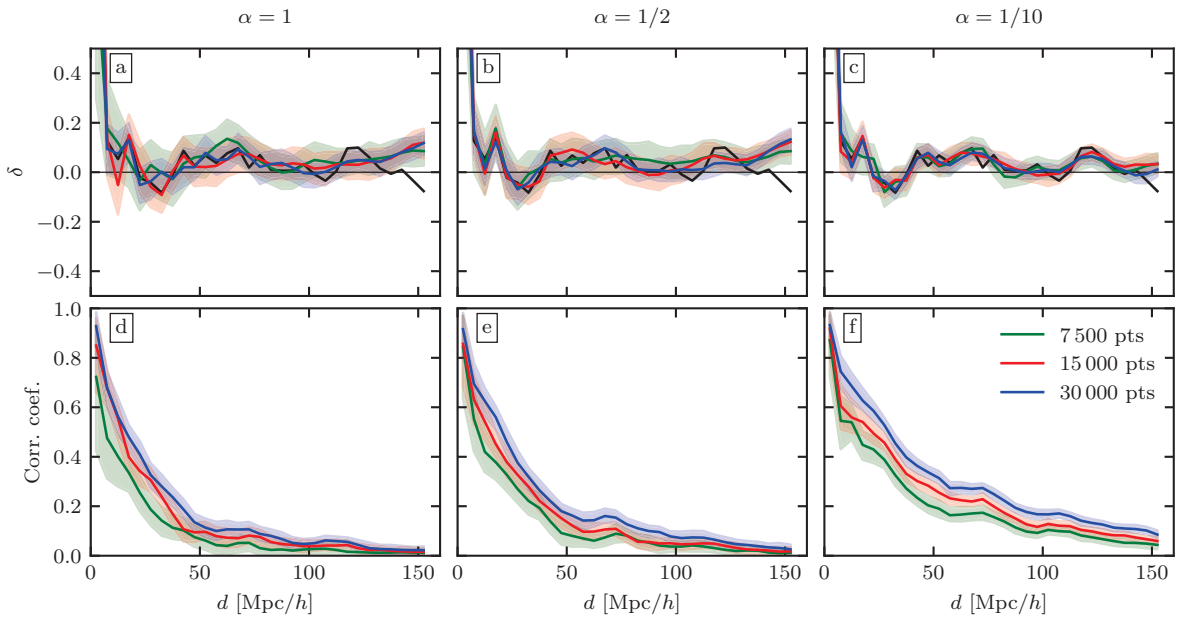

Figure 2.8: Top row: The mean plus-minus standard deviation of the density field per distance shell is shown for $\alpha = 1, 1/2$ and $1/10$ (left to right). In black we show the target results while in green, red and blue we show the curves reconstructed from 7 500, 15 000 and 30 000 points respectively. Beyond this region there are no constraints. Bottom row: same plots but for the correlation between the reconstructed field and the target field.

approach. Quantifying the speed-up of the HAMLET method compared to that of Graziani et al. (2019) is complicated because there are differences to both the implementation (compiled on GPUs versus interpreted on a single CPU) and the algorithm (HMC versus Gibbs sampling). Its thus not straight forward to identify exactly which aspect of the HAMLET method is mostly responsible for the increase in efficiency. A quick comparison shows that the HAMLET code outperforms the code of Graziani et al. (2019) by several orders of magnitude (between 3 to 4) in speed while fitting orders of magnitudes more parameters ($\sim 2 \times 10^6$ versus $\sim 4.5 \times 10^4$). For example, in the case of reconstructing the LSS from the $\sim 1.5 \times 10^4$ constraints provided by the grouped CF3 catalogue, the MCMC method of Graziani et al. (2019) takes more than a month compared with the HAMLET which takes on the order of 10 minutes. The increase in speed is a necessary condition for future applications of the HAMLET code. One such application is the setup of constrained initial conditions for high resolution cosmological simulations (cf. Libeskind et al., 2020), for which the number of the needed Fourier modes is much larger that what is used here. Also, the upcoming 4th Cosmicflows data release (CF4) is expected to roughly triple the size of the CF3 data. Preliminary analysis suggests that HAMLET will be capable in exploiting the CF4 data.

A brief review of the computational implementation of the HAMLET code follows. It takes advantage of a number of highly-abstract layers as implemented by open-source `Python` libraries `tensorflow` and `tensorflow-probabilities`. The `tensorflow` library provides a framework that enables a python code to transparently scale on multiple CPUs and/or GPUs and to be compiled at run time, while the `tensorflow-probabilities` provides a plug-and-play implementation of the HMC, NUTS and other tools to run and analyse MCMC chain. While the gradient of the the posterior PDF, can be extremely tedious to write by hand, `tensorflow` is capable of transparently computing it, by constructing a complex derivation graph that can be very efficiently evaluated. Only the gradient of the inverse of the Fourier transform had to be constructed.

2.5 Testing Hamlet against a linear mock Cosmicflows-3 survey

2.5.1 Mock Catalogue construction

A linear realization of a Gaussian random field, defined by the ΛCDM power spectrum and cosmological parameters (Planck Collaboration et al., 2016) is used here as a base for our mocks. The field is

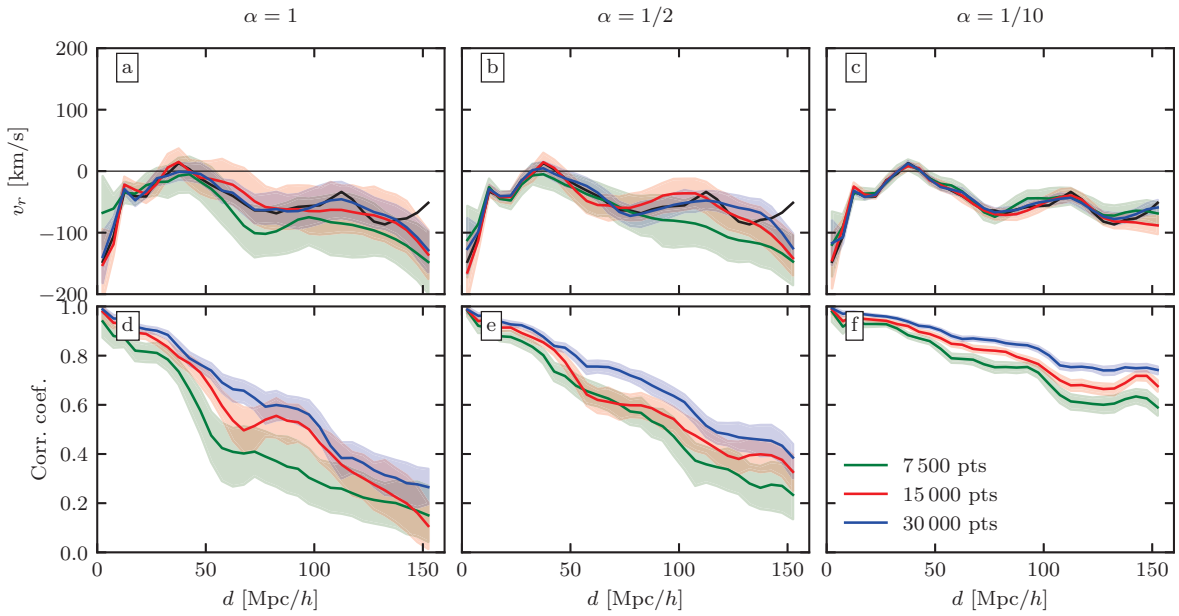

Figure 2.9: Same as fig. 2.8 but for the radial velocity field instead of the density field.

constructed on a 128^3 Cartesian grid within a box with side length $L = 500\,\mathrm{Mpc}/h$. Periodic boundary conditions are assumed. A random observer is selected to reside at the center of the computational box and a mock Supergalactic coordinate system is assigned centered on the observer and aligned with the principal directions of the grid. A mock catalog consists of Supergalactic latitude (SGB) and longitude(SGL), distance modulus (μ), its error (σ_μ), the redshift (z) and its error (σ_{cz}).

fig. 2.1 presents the "target" density and velocity field, from which the mock data has been drawn and which the HAMLET algorithm is designed to recover. The linear over-desnity (δ) and the radial component of the velocity field (v_r) are depicted. The target field is smoothed with a Gaussian kernel of a radius of $5\,\mathrm{Mpc}/h$.

The constraints are isotropically selected within a sphere of radius of $160\,\mathrm{Mpc}/h$. A radial selection function is imposed so as to have a uniform distribution per radial distance bins, $P(d) = \text{constant}$, where distances are measured with respect to the mock observer at the relative centre of the box. This choice is motivated by the relative flatness of the redshift distance distribution of the CF3 grouped data points (fig. 2.1 bottom). The 160 Mpc/h cut corresponds to the effective distance cut of the CF3 data. The errors assigned to the mock data points follow the redshift distribution of the errors of the actual CF3 data. The following procedure is used. Given a mock point it inherits the error of the actual CF3 data point that is closest to it in redshift.

The two main factors that affects the quality of the Bayesian reconstruction in general and the HMC in particular are the numbers of the data points and their associated errors. In the limit of very dense sampling of the data points and negligible errors the target field should be reconstructed with high fidelity. In the other extreme case of very sparse sampling and large observational uncertainties, the null field predicted by the prior PDF is recovered. An ensemble of 9 different mock databases has been constructed so as to investigate how these two factors affect the outcome of the HAMLET reconstruction. Three different numbers of data points are selected, $(7.5,\ 15,\ 30) \times 10^4$. The distance moduli errors are gauged by an α parameter, $\sigma_\mu = \alpha\sigma_\mu^{\mathrm{CF3}}$, where $\sigma_\mu^{\mathrm{CF3}}$ is the value for the actual CF3 survey error[4]. In otherwords $\alpha = 1$ corresponds to the typical errors associated with the inherent uncertainty in scaling relations (eg. Tully-Fisher) distance measures, while $\alpha = 1/10$ is meant to mimic a catalogue constructed entirely with more accurate distance measures, like TRGB or SuperNovae. The 9 mock databases are assigned the 3 different numbers of data points and 3 different α values (table 2.1). The case of 1.5×10^4 data points and $\alpha = 1$ corresponds most closely to the grouped CF3 data.

A note of caution on the expected effect of the sharp drop has on the Bayesian reconstruction is

[4]Varying σ_μ is meant to mimic the different precision of different standard candles since *e.g.* the error on a distance obtained from SN method is around 3-5% while from scaling relations are closer to 20-30%. Thus a catalogue of just scaling relations like TF would correspond to $\alpha = 1$ while a catalogue of just SN would correspond to $\alpha \approx 0.1$.

Table 2.1: Mock catalogues and their characteristics

Name	N of points	α, Factor on σ_μ
CF3+ like	15 000	1
Better measurements	15 000	1/2
Very good measurements	15 000	1/10
More measurements	30 000	1
More, better measurements	30 000	1/2
More, very good measurements	30 000	1/10
Fewer measurements	7 500	1
Fewer, better measurements	7 500	1/2
Fewer, very good measurements	7 500	1/10

due here. Hinton et al. (2017) have investigated this exact problem: the effect that a sample selection function with a sharp cutoff has on the likelihood function and thereby on the Bayesian posterior PDF. Their conclusion is that the inferred variables close to the edge of the data, namely close to the cutoff, are biased. Our analysis and findings support the finding of Hinton et al. (2017). Consequently we limit our analysis and present results only within a sphere of a radius of 150 Mpc/h.

2.5.2 Convergence

A critical issue that MCMC methods in general and the HMC in particular face is that of convergence, namely how long should an MCMC chain or HMC trajectory be to meet some given criteria of confidence in the estimated parameters. (The discussion that follows focuses on HMC trajectories but it implies to MCMC chains in general.) The HMC trajectories never "rest" and keep on "moving" in the parameters space. Three obvious issues to consider are: a. Does the HMC trajectory oscillate around the "true" values of the parameters, namely the issue of bias; b. What is the scatter exhibited by the trajectory, *i.e.* the variance around the mean (defined by eq. (2.4)); c. What is the rate of convergence. The convergence rate is discussed here and the issues of bias and variance are addressed in subsections 2.5.3 and 2.5.5 below.

The rate of convergence is examined here by monitoring the change of the mean of the density field, $\langle\delta\rangle$, along the HMC trajectory. fig. 2.2 presents the differential change in $\langle\delta\rangle$ normalized by the (square root of the) cosmic variance between successive steps, $\left|\left(\langle\delta\rangle/\Sigma_\delta\right)_s - \left(\langle\delta\rangle/\Sigma_\delta\right)_{s-1}\right|$. Here Σ_δ^2 is the variance of the δ field evaluated at a given step of the HMC. fig. 2.2 indicates that to get to percent level convergence requires on the order of 100 chains.

2.5.3 Reconstruction of the large scale structure

The HAMLET method's main mission is to perform a Bayesian estimation of the LSS, namely the density and velocity fields, from Cosmicflows-like databases. To meet that end and examine it we focus here on a subset of the $\boldsymbol{q}$ parameters, $\boldsymbol{q}_{\rm LSS} = \left(\delta^R_{k,1}, \ldots, \delta^R_{k,m}, \delta^I_{k,1}, \ldots, \delta^I_{k,m}\right)$, namely the ones that determine the LSS. The conditional mean field, given the data and the prior model is readily written (as a particular case of the general eq. (2.4)) as:

$$\langle\boldsymbol{q}_{\rm LSS}\rangle = \langle\boldsymbol{q}_{\rm LSS} \mid \text{data},\ \text{prior}\rangle \approx \frac{1}{n_s}\sum_{i=1}^{n_s}\boldsymbol{q}_{\rm LSS,i} \tag{2.32}$$

The conditional mean field and the ensemble of states are the HMC equivalent of the WF estimator and the ensemble of CRs of the WF/CRs algorithm.

Figure 2.3 shows a slice of the δ field for the 9 different mock data sets (see table 2.1). The presented density field is the conditional mean field given the data, namely $\langle\delta\rangle = \delta\left(\langle\boldsymbol{q}_{\rm LSS}\rangle\right)$. The grid of density maps reflects the change of quality of the data - more data points and smaller errors corresponding to better data. Degradation of the data leads to an attenuation of $\langle\delta\rangle$. This a manifestation of the well known property of the Bayesian estimation - the worse the data the more biased the results are towards the null field predicted by the prior model (cf. Zaroubi et al., 1995). The grey lines represent the $\delta = 0$ contour of the target. The reader will note how similar these are to the the limit of small errors and large data set (*i.e.* fig. 2.3(i)) Indeed, in general, the target and reconstructions $\delta = 0$ contours match

for $\alpha = 1/10$, fig. 2.3(g,h,i). In the case of the CF3 mock ($\alpha = 1$, 15,000 points, figure fig. 2.3(b)), the target density contour is fairly accurately recovered in the inner regions, ie within around $50-60$ Mpc/h. Examining the density field reconstructions "vertically" indicates that the single most important factor for obtaining an accurate density field reconstruction is the data quality. Namely: *Better data is more important than larger data sets*. For a given catalogue size, better data allows the reconstruction to be accurate at greater distances. For a given error, more data improves the reconstructions at fixed distances, instead of extending the improvement of the reconstruction.

The linear over-density field constitutes a random Gaussian field whose variance is determined by the power spectrum of the field and the resolution of the given realizations of the field. This is the cosmic variance of the δ field, denoted by σ_δ^2. The variance of different states along the HMC chain, Σ_δ^2, varies according to the "strength" of the data and the properties of the prior model. Furthermore, it varies with the location at which it is evaluate, *i.e.* $\Sigma_\delta = \Sigma_\delta(\boldsymbol{r})$. The spatial variation of $\Sigma_\delta(\boldsymbol{r})$ reveals the constraining power of the data, given the prior model. Where $\Sigma_\delta(\boldsymbol{r})/\sigma_\delta \ll 1$ the density field is strongly constrained by the data and the prior model and only very small scatter is expected to be found around the mean field. Where $\Sigma_\delta(\boldsymbol{r})/\sigma_\delta \sim 1$ the field is essentially unconstrained by the data and the prior model. fig. 2.4 presents the variation of normalized constrained variance, $\Sigma_\delta(\boldsymbol{r})/\sigma_\delta$, for the 9 mock databases presented in fig. 2.3. Inspection of fig. 2.4 reveals that for all the mock data considered the quality of the reconstruction degrades with the distances from the (mock) observer. This is a reflection of the degradation of the data with the distance - the magnitude of the errors increases and the density of data point decreases with distance. The picture of how the reconstruction degrades with distance as a function of date set size and error reinforces the conclusions drawn from fig. 2.3 namely smaller errors on the data improves the reconstructions more than larger data sets. Small data sets with small errors are worth more than large data sets with large errors when examining the reconstructed density field.

Next, the radial component of the velocity field is investigated (figs. 2.5 and 2.6). The velocity field power spectrum is "redder" than that of the δ field, namely it has more power on long wavelengths than on short ones, hence the velocities' effective correlation length is larger than that of the densities. Hence one expects the velocities to be more constrained by the data than the densities. This is clearly manifested by fig. 2.5. A visual of how well the target's $v_r = 0$ contour matches the reconstructed radial velocity field indicates that even in the case of CF3 like mock (ie fig. 2.5b) the reconstructed velocity field is doing a good job at greater distances (as compared with the δ field). The velocity field around large distance concentrations of matter and voids is accurately reconstructed with HAMLET . We can qualitatively asses the superiority of the velocity field reconstruction as compared to the density field by examining fig. 2.6 (and comparing to fig. 2.4) which shows just how well similar the reconstructed velocity field is to the target.

2.5.4 Monopole and dipole

Global measures of the velocity field are given by the volume-weighted mean monopole and dipole moments of the velocity field in spheres of radius R (Hoffman et al., 2021, for details). The monopole moment is the mean of $-\nabla \cdot \boldsymbol{v}/H_0$, where the scaling by H_0 is introduced so as to make the expression dimensionless and proportional to the mean (linear) over-density within R. The minus sign is introduced so as to make the monopole within a sphere of R to be proportional to the mean over-density within that volume. The dipole moment is the (volume weighted) mean value of the velocity, namely it is the bulk velocity of a sphere of radius R. The variation of the monopole and dipole moments with depth provides a global measure of the underlying LSS of the universe, and as such they serve as good monitors of the quality of the HAMLET reconstruction.

fig. 2.7 presents the variation with depth of the monopole (upper row) and the dipole (middle row) of the velocity field, namely the mean fractional over-density (δ) and the bulk velocity of a sphere of radius R. The lower row shows the alignment of the bulk velocity of the reconstructed and the target velocity fields. The dependence of the estimation of the radial profiles of the moments on the quality of the data is investigated. The profiles are shown as a function of the magnitude of the errors ($\alpha = (1.0,\ 0.5,$ and $0.1)$ and of the number of data points, $N = (0.75,\ 1.5$ and $3.0) \times 10^4$. The profiles are presented by their mean and standard deviation taken over the ensemble of steps.

The plots of fig. 2.7 are informative. Reconstructions behave as expected, with the exception of the monopole moment close to the edge of the data, $R \lessapprox 150$ Mpc/h, where the reconstructed monopole exceeds that of the target one. The edge-of-the-data discrepancy is in line with the findings of Hinton et al. (2017). As for all the other cases they behave as expected. The mean HAMLET profiles' deviation from the target profile and the scatter around the mean profiles grow with R and get smaller with

the increase of the number of steps. A note is due here on the large scatter in the amplitude and alignment of the bulk velocity at $R \gtrapprox 100$ Mpc/h, say. The bulk velocity of a sphere of radius R is induced by structures outside that radius. The target field is constructed within a box of side length of $L = 500$ Mpc/h with periodic boundary condition, which renders the power within the box and outside these radii, and its constraining power to be rather small. This is manifested by the large scatter around the mean. The lesson to be learnt here is that the reconstruction of the bulk velocity on a given scale R needs to be done within boxes of L much larger than R.

2.5.5 Correlation Coefficients

Next, the fidelity of the HAMLET reconstruction is monitored by means of the scatter of the density and radial velocity fields evaluated in radial shells. The upper panel of fig. 2.8 shows the mean and scatter of the density in spherical shells of width of $\Delta R = 10$ Mpc/h for the 3 values of α and the 3 assumed sizes of the mock catalogue. The mean density of the target field is presented for reference. The disagreement at $R \lessapprox 150$ Mpc/h is again clearly manifested. Next the density field within the shells is examined by studying the correlation between the densities of individual voxels (grid cells) within the shells. The lower panel of fig. 2.8 shows the (Pearson) correlation coefficients of the HAMLET reconstructed and the target densities. The correlation coefficient profiles are again plotted for the 3 levels of errors and the 3 ensembles of the HMC chain. In general, the HAMLET reconstruction follows the target quite well. In the limit of much data and small errors, the reconstruction is very close to the target.

fig. 2.9 applies the analysis of fig. 2.8 to the case of the radial velocity field. The mean and scatter of v_r within spherical shells (upper row) and the correlation coefficient with the target field (lower row) are presented in fig. 2.9. As expected, the comparison of the 2 figures clearly shows that the (radial) velocity field is much more correlated than the density field. Smoothing the target and the reconstructed density fields would make them much more correlated. Again, in the limit of much data and small errors the reconstructed velocity field is so accurate that a correlation coefficient of nearly 0.8 is obtained at 150 Mpc/h. Obtaining a CF like catalogue with such large numbers of data points is already a reality in CF4. On the other hand, having small errors corresponding to $\alpha = 1/10$ is still slightly unrealistic, but not unimaginable with future purpose built telescopes designed specifically to monitor variable stars at cosmological distances.

2.6 Partial conclusion

The problem of the reconstruction of the large scale density and velocity fields from peculiar velocities surveys is addressed here within a Bayesian framework. In particular, the reconstruction aims at Cosmicflows-like data where observational uncertainties are on the distance moduli, which results in a Log-normal bias on the estimated distances and velocities (see section 1.5.1). The HAmiltonian Monte carlo reconstruction of the Local EnvironmenT (HAMLET) algorithm performs the reconstruction within the framework of the linear theory of the ΛCDM standard cosmological model, which is taken here as the Bayesian prior, using the Hamiltonian Monte Carlo (HMC) method to sample the posterior probability distribution function (PDF) given the ΛCDM model and the Cosmicflows-like data (*e.g.* Tully et al., 2016). Like previous MCMC treatments of the problem (Lavaux, 2016; Graziani et al., 2019) the HAMLET samples the posterior PDF of true distance of the data points coupled with the underlying linear density field. This differs from the Wiener filters and constrained realizations (WF/CRs) approach where the correction of the Log-normal bias is done independently of the Bayesian reconstruction of the LSS (see section 1.9.5 or Sorce, 2015; Hoffman et al., 2021).

The current HAMLET HMC algorithm and the Lavaux (2016); Graziani et al. (2019) MCMC ones are formulated within the same mathematical Bayesian framework, making similar assumptions on the prior PDF and deriving the same posterior PDF from the same input data. The main difference between the standard MCMC algorithm and the HMC is in the sampling of the posterior PDF. The extremely low rejection rate of the HMC steps and its ability to be run on GPUs, makes the procedure very efficient. A comparison of the performance of the HAMLET algorithm with the one presented in Graziani et al. (2019) finds an efficiency gain factor of two to four orders of magnitude in favor of HAMLET . This gain in efficiency will enable a very significant increase of resolution in future applications of the HAMLET algorithm compared to to the resolutions used in the MCMC cases (Lavaux, 2016; Graziani et al., 2019).

Chapter 3

Tests on realistic mocks and comparison with the BGc/WF

3.1 Introduction

As such, compilations of peculiar velocities are difficult to analyze (for a comprehensive review Strauss & Willick, 1995) and are usually a patchwork of various surveys and methods observed with different telescopes in different locations on earth (or in space). The POTENT method was the first attempt to produce continuous maps of the density and velocity fields based on peculiar velocities surveys (Bertschinger & Dekel, 1989). The main underlying assumption of the POTENT method is that galaxy velocities are drawn from an irrotational, potential flow. No further assumptions, were made on the statistical nature of the flow field beyond the existence of a galaxy bias (Kaiser, 1987). Therefore, its ability to handle the shortcomings of such peculiar velocity surveys was limited. Subsequent approaches to the reconstruction of the LSS from peculiar velocities have been formulated within Bayesian frameworks - these include the Wiener filter (WF) and constrained realizations (CRs) methodology (Ganon & Hoffman, 1993; Zaroubi et al., 1999; Tully et al., 2019) as well as Markov Chain Monte Carlo algorithms (MCMC Lavaux, 2016; Graziani et al., 2019; Boruah et al., 2022; Prideaux-Ghee et al., 2022). They have been remarkably successful in "mapping the invisible" and recovering the underlying cosmic fields.

Beyond the issues of noisy, sparse data, plagued with inhomogenous errors, there is one additional inherent conceptual problem common to all surveys of peculiar velocities and that is that peculiar velocity itself is not observed but is a *derived* quantity. Given the redshift of and a distance to a galaxy, it is the radial component of its peculiar velocity that can be computed. But only the redshift is observed; distances themselves are not directly observed. What is measured is the distance modulus of a galaxy (cf. Tully et al., 2016). Because the error of the measured distance modulus is assumed to be normally distributed, the errors on the observed distances are thus log-normally distributed. This leads to a biased estimate of the distances and peculiar velocities with respect to the actual distances (see Hoffman et al., 2021). Often this bias is treated as yet another manifestation of the Malmquist bias (see Strauss & Willick, 1995). Here we refer to it as the log-normal bias. For the WF/CRs reconstruction algorithm, the log-normal bias is treated outside of the Bayesian framework in a separate process (Sorce, 2015; Tully et al., 2014; Hoffman et al., 2016, 2021). For Monte Carlo methodologies the log-normal bias is treated within a comprehensive algorithm (Lavaux, 2016; Graziani et al., 2019).

The Constrained Local UniversE Simulations (CLUES; Yepes et al., 2009; Sorce et al., 2014; Sorce, 2015) project focuses on the reconstruction of LSS of our nearby cosmic neighbourhood from surveys of galactic distances and thereby peculiar velocities, in particular the Cosmicflows database (cf. Tully et al., 2016, and references therein). Two main methodologies have been employed by the CLUES for the reconstruction of the local LSS and the setting of initial conditions for constrained cosmological simulations - one that is based on the WF/CRs methodology (Hoffman & Ribak, 1992; Zaroubi et al., 1995) and the other on MCMC and of Hamiltonian Monte Carlo (HMC) sampling. In particular, within the WF/CRs framework the issue of the log-normal bias has been handled by two independent algorithms, that of Sorce (2015) and that of Hoffman et al. (2021) and within the Monte Carlo sampling approach by Graziani et al. (2019) and chapter 2.

Our aim in this work is to test the quality of two methods that reconstruct the LSS from peculiar velocities: the WF/CRs method with a log-normal bias correction algorithm, known as the Bias Gaussian

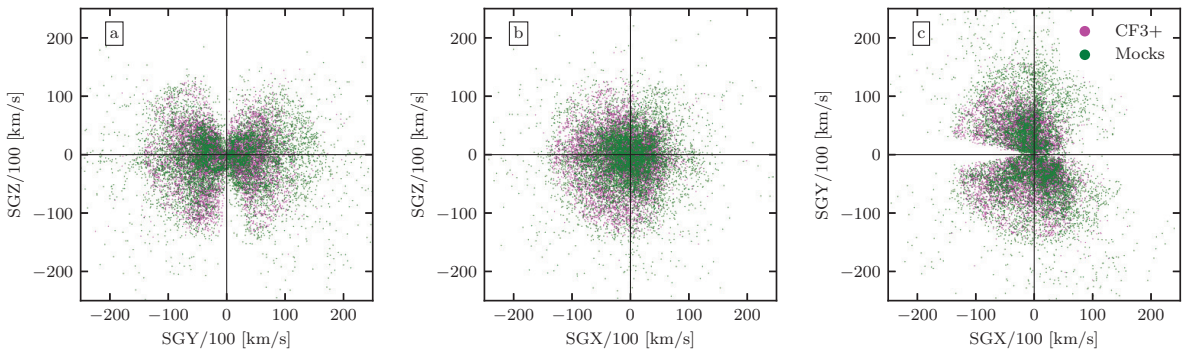

Figure 3.1: The distance, in units of km/s, of the CF3 data points (magenta) and mock data points (green) projected on the three supergalactic principal planes. Note the ZOA is accurately reproduced in the mock catalogues.

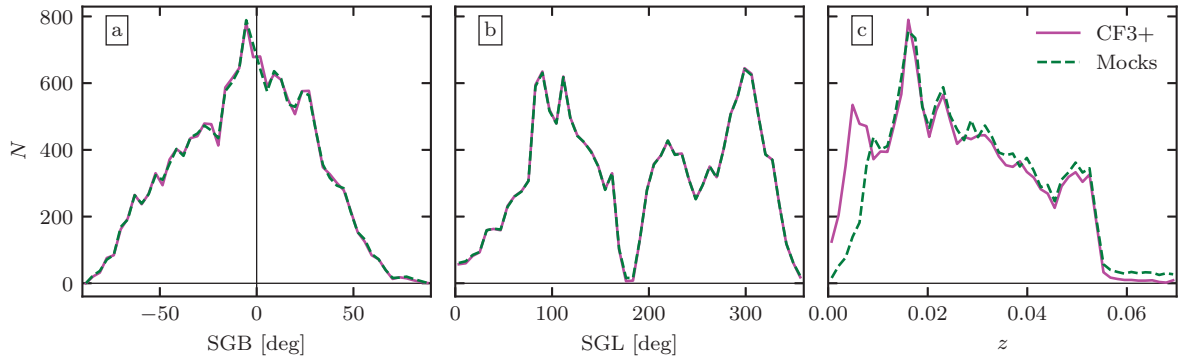

Figure 3.2: From left to right, the distribution SGB, SGL and z. Note that in panels a and b the two curves are on top of one an other.

Correction (BGc; Hoffman et al., 2021) and the HAmiltonian Monte carlo reconstruction of the Local EnvironmenT (HAMLET for short) method (see chapter 2). These two methods are applied to a mock data catalogue drawn from a cosmological simulation designed to imitate the CosmicFlows-3 data (Tully et al., 2016). The original simulation is referred to as the target simulation. The two reconstructions are compared with the target simulation to gauge their fidelity.

This paper is structured as follows. In section 3.2 the algorithm for constructing halo catalogues that mock the cosmic flows data is presented. In section 3.3 the nature of the input data as well as its biases and a bias correction scheme are presented. The results of applying to the two reconstruction methods to these mocks, as well as a comparison between them is presented in section 3.5. A summary and conclusion is offered in section 3.6.

This is work is published as Valade et al. (2023). A few editorial modifications have been made to insure the global consistency of the thesis.

3.2 Mock Catalogue construction

We wish to create a mock version of the grouped Cosmicflows catalogue that reproduces its main characteristics, since it is on these observational data that the methods studied here are supposed to be applied to. We start from the publicly available CF3 data release and add to this $\sim 4\,000$ points given to us by the authors of CF4 as a pre-release (Tully, private communications[1]), resulting in a ensemble of $\sim 15\,000$ entries, hereafter named CF3+.

A mock catalogue is constructed from the MultiDark Planck 2 simulation[2] (MDPL2, Riebe et al.,

[1] R. B. Tully provided us with an advance set of redshifts and angular positions of CF4 for the purpose of this paper. Distance modulii, and associated erros were not provided.

[2] The MultiDark simulations are publicly available: www.cosmosim.org

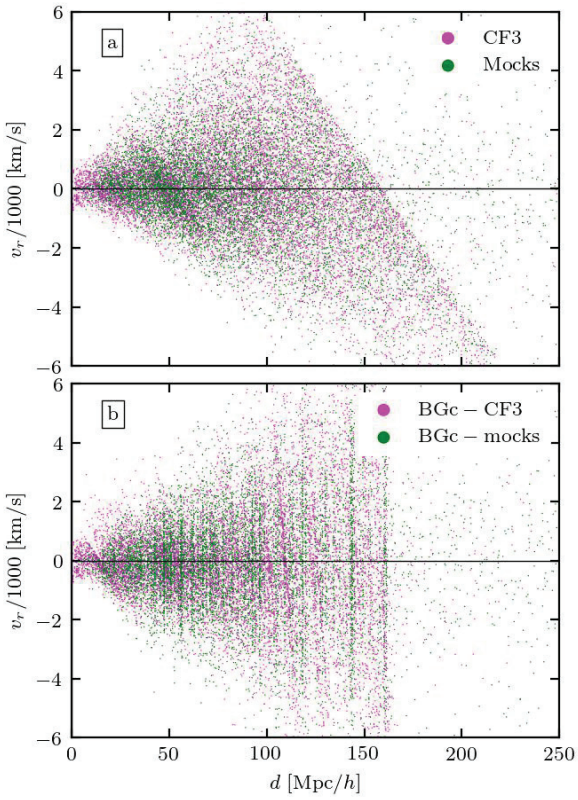

Figure 3.3: The distance of a galaxy as a function of its peculiar velocity is shown for the grouped CF3 data (magenta) as well as the mock catalogue (green). We can only make use of CF3 and not CF3+ as distances (and thus radial velocities) were not communicated for the pre-release of CF4. The log-normal bias is evident here in the lack of symmetry about $v_r = 0$; beyond around 70Mpc/h the universe appears to be systematically collapsing, in a so-called "breathing mode". Bottom panel: after application of the BGc correction, symmetry is reestablished.

2013), a dark matter only N-body run of $N = 3840^3$ particles in a periodic box of side length $L = 1\,\mathrm{Gpc}/h$. The cosmological parameters of the simulation are from the 2nd Planck data release Planck Collaboration et al. (2016) *i.e.* a flat ΛCDM Universe $\Omega_\mathrm{m} = 0.307$, $\Omega_\mathrm{b} = 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 = 0.8228$, $n_s = 0.96$ and a dimensionless Hubble parameter $h = 0.678$ where $H_0 = 100 \times h$ km/s/ Mpc. A Friend-Of-Friend's (FOF) algorithm with a linking length of 0.2 times the mean inter particle separation is used to identify haloes whose mass is roughly M_{200} (Davis et al., 1985). It is appropriate to use a FOF halo in this case since it is the *grouped* CF3 catalogue which is being mocked. Grouping the members of a virialized object together averages out nonlinear motions implying that the (*e.g.*) cluster's peculiar velocity is a better traces of the flow field. Note that the MDPL2 box size of $L = 1\,\mathrm{Gpc}/h$ is large enough to embed the CF3 catalogue, whose effective depth is roughly $160\,\mathrm{Mpc}/h$.

An "observer" is associated with a randomly selected halo of mass in the range of $[0.9 — 2.0] \times 10^{12}\,\mathrm{M}_\odot/h$. The simulation is then re-centered on this halo and the simulation's coordinate axes are then arbitrarily labelled as Supergalactic (SGX, SGY, SGZ). Furthermore a mock "sky projection" is made such that each halo is given a sky position (SGL, SGB).

The (proper) distance d of each halo from the center, is used to compute a cosmological redshift $\bar{z}$ by numerically integrating eq. (1.41). The (proper) distance d is turned into a luminosity distances d_L by

$$d_L = d \times (1 + z_\mathrm{cos}) \tag{3.1}$$

which is a simplification of eq. (1.49). The halo's distance modulus is then computed with eq. (1.5). Note that eq. (3.1) is an approximation of the full luminosity distance as described in eq. (1.49) (from Calcino & Davis, 2017). We include here only the cosmological redshift, which leads to a difference on the order of 10% at the edge of the Cosmicflows data. The corrections to the luminosity distance due to peculiar velocity redshift are of the order of $\approx 0.2\%$ for any observed data point, which is negligible

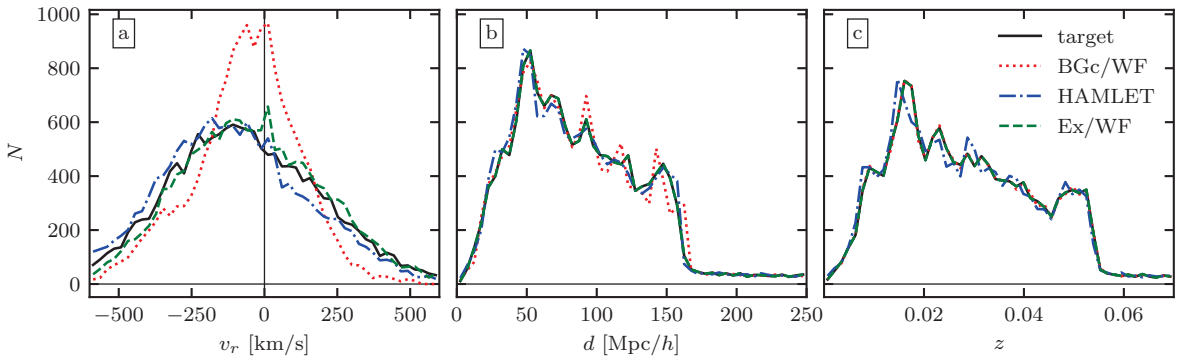

Figure 3.4: From left to right: a The distributions of the radial peculiar velocity, b the distance and c the redshift for the target (black solid lines), the BGc/WF (red dotted) and HAMLET (Blue dash dotted) reconstruction methods. The Ex/WF method is shown in green dashed.

against the 5-20% observational errors. These should however be taken into account in future works or for the estimation of cosmological parameters.

The radial peculiar velocity v_r is combined with the cosmological redshift to obtain the full redshift (see eq. (1.47))

$$z + 1 = (z_{\cos} + 1)\left(\frac{v_r}{c} + 1\right). \tag{3.2}$$

The velocities here are relative to the simulation box (which is equivalent to the Cosmic Microwave Background).

At this point each halo's position relative to the observer has been transformed into two "observable" quantities: 1. a redshift z (which includes a contribution from the radial peculiar velocity v_r) and 2. a distance modulus μ.

The mock catalogue aims to have the same Probability Distribution Functions (PDF) of P(SGB), P(SGL) and $P(z)$ as in the CF3+ data. This is accomplished with a monte-carlo style algorithm in the following way: the same number of haloes as data points in CF3+ are drawn at random from the simulation, within a sphere of around 300 Mpc/h. A merit is assigned to this initial set of haloes by computing the absolute difference between its P(SGB), P(SGL) and $P(z)$ and that of CF3+. Iterations proceed by adding and subtracting one halo at a time and evaluating the merit of the new P(SGB), P(SGL) and $P(z)$, compared to CF3+'s. If a new potential halo improves the merit of the distributions, it is kept; otherwise it is rejected. In this way, the process converges halo by halo, towards reproducing the distribution CF3+'s P(SGB), P(SGL) and $P(z)$.

Once the merit function has converged and a suitable mock catalogue has been constructed, the observational errors from the CF3 catalogue are added to the mock. Namely, the redshift and distance modulus of each CF3 data point is given as $z+\varepsilon_z$ and $\mu+\varepsilon_\mu$ where ε_z and ε_μ denote the errors associated with each measurement. ε_z is assumed to be entirely due to spectroscopic precision while ε_μ depends on which standard candle is used and may range from 5% for Supernova to 20% for scaling relations. Both ε_z and ε_μ are assumed to be Gaussian with means of zero and standard deviations of $c\sigma_z = 50$ km/s and σ_μ, respectively. The value of σ_μ associated to each halo is taken from the entry of CF3 whose redshift is the closest, so as to reproduce the dependency of the σ_μ with the distance.

The fidelity of the mock to the CF3 catalogue, is shown in figs. 3.1 and 3.2. fig. 3.1 shows the three supergalactic projections with the mock data points in green and the CF3 constraints in purple. The Zone of Avoidance (ZOA) and visual distribution of the catalogues are well recovered. Quantitatively this is shown by looking at the distributions of P(SGB), P(SGL) and $P(z)$ themselves shown in fig. 3.2a, b, and c, respectively. The distribution of SGB, SGL and cz for the mock galaxies and CF3 constraints, are largely indistinguishable from each other. fig. 3.2c shows that within ~ 20 Mpc/h, the number of CF3 constraints is much greater than the mock catalogues presented here. This is because of there are too many CF3 constraints in this region, with respect to the resolution of our simulation.

In principle the original unperturbed d, d_L, $z_{\cos}$ and v_r for each halo that in the mock can be "forgotten" and new values can be computed using the values of z and μ that include the observational errors. These new values should exhibit similar biases to the observational data by construction. This is seen in fig. 3.3, where the radial velocity as a function of distance is plotted. The diagonal cut in this

plot is indicative of the log normal bias dicussed in section 3.3. At a given distance there is an unequal number of galaxies moving towards and away from the observer, making it appear that the universe is contracting in a "breathing mode". This log normal bias and its correction are presented in section 3.3.

3.3 The log-normal bias and the Bias Gaussian correction (BGc)

One of the main purposes of constructing such a detailed mock catalogue as described above is to ensure that the log-normal bias is reproduced, thereby allowing us to gauge the ability of the two reconstruction methods to handle this bias. Much hand wringing and literature has been devoted to the handling of biases in peculiar velocity surveys, and we refer the reader to Strauss & Willick (1995) for a comprehensive explanation (or section 1.5.1 of this thesis). Here we briefly explain what the log-normal bias is and how it is handled in the context of the BGc as proposed by Hoffman et al. (2021). We refer the reader to that work for a comprehensive description of the log-normal bias and its correction by the BGc algorithm.

As mentioned, a Gaussian error on the distance modulus transforms into a log-normal error on the luminosity distance (e.g. the inverse of eq. (1.5)). In other words, if the same galaxy is observed many times, the mean of the different distance measures will not coincide with its actual value. This bias changes the spatial distribution of the galaxies as well as their inferred peculiar velocities. The log-normal bias can be seen in fig. 3.3 where the CF3 and mock catalogue peculiar velocity v_r is plotted as a function of distance. Beyond around $\sim 70\,\mathrm{Mpc}/h$, there is no longer symmetry in the distribution of v_r about zero: more galaxies have negative v_r and the universe naively appears to be collapsing, a so-called "breathing" mode. In theory it can be corrected since the standard ΛCDM model makes an explicit prediction that the expected scatter for the radial component of the velocity is roughly $\sigma_{v,\mathrm{th}} \sim 275\,\mathrm{km/s}$.

The essence of the BGc scheme is to map the log-normal distribution of the inferred distances around their respective redshift distances into a normal distribution around the median of the log-normal one. The width of that normal distribution is treated as a free parameter set to be about $3\,\mathrm{Mpc}/h$, in agreement with the ΛCDM prediction that the *theoretical* intrinsic scatter of the radial velocities is $\sigma_d = \sigma_{v,\mathrm{th}}/H_0$:

$$d_{\mathrm{BGc}} = d_{cz}^{\mathrm{med}} + \frac{\sigma_d}{\nu_\mu} \log\left(\frac{d}{d_c^{\mathrm{med}} z}\right), \tag{3.3}$$

$$d_{cz}^{\mathrm{med}} = \mathrm{median}\big\{d_{1\leq i\leq n_{\mathrm{obs}}} \text{ s.t. } cz - \Delta < cz_i < cz + \Delta\big\}. \tag{3.4}$$

where d_{cz}^{med} is the median of the distances of the constraints in the redshift neighbourhood $[cz-\Delta, cz+\Delta]$ of the considered constraint.

The same procedure is applied to the observed radial velocities, retaining the median of the distribution of the radial velocities of data points in a given redshift bin:

$$v_{\mathrm{BGc}} = v_{cz}^{\mathrm{med}} + \frac{\Sigma_v}{\nu_\mu} \log\left(\frac{v}{v_c^{\mathrm{med}} z}\right), \tag{3.5}$$

$$v_{cz}^{\mathrm{med}} = \mathrm{median}\big\{v_{1\leq i\leq n_{\mathrm{obs}}} \text{ s.t. } cz - \Delta < cz_i < cz + \Delta\big\} \tag{3.6}$$

where Σ_v is the *observational* uncertainty (as opposed to the *theoretical* scatter). Indeed, for the velocities, unlike the inferred distances, the variance of the distribution is preserved as well. Namely the log-normal distribution of the observed distances is mapped to a Gaussian distribution, while preserving the median of the log-normal distribution. It is the invariance of the median under the normal - log-normal transformation which constitutes the backbone of the BGc scheme. After the application of the BGc scheme to the data, the breathing mode dissapears and the radial peculiar velocities scatter normally about 0 as can be seen in fig. 3.3, bottom panel.

3.4 Wiener Filter reconstruction from Exact data (Ex/WF)

As there exist no possibility to homogenize the sampling (namely the Zone of Avoidance will always inhibit full sky coverage), the only source of statistical uncertainty that could, one day, be mitigated, is the observational uncertainty in the distance measurement. In order to test the methods' inherent ability to reconstruct the underlying fields, an additional "method" is compared: the exact WF (hereafter labeled Ex/WF). This is the WF applied to a mock where the error on each data point has been artificially set to zero (and thus no BGc scheme is applied). The reader will note that both the Ex/WF and the BGc/WF will never be fully accurate due to the fact that the density and velocity fields are non linear and these

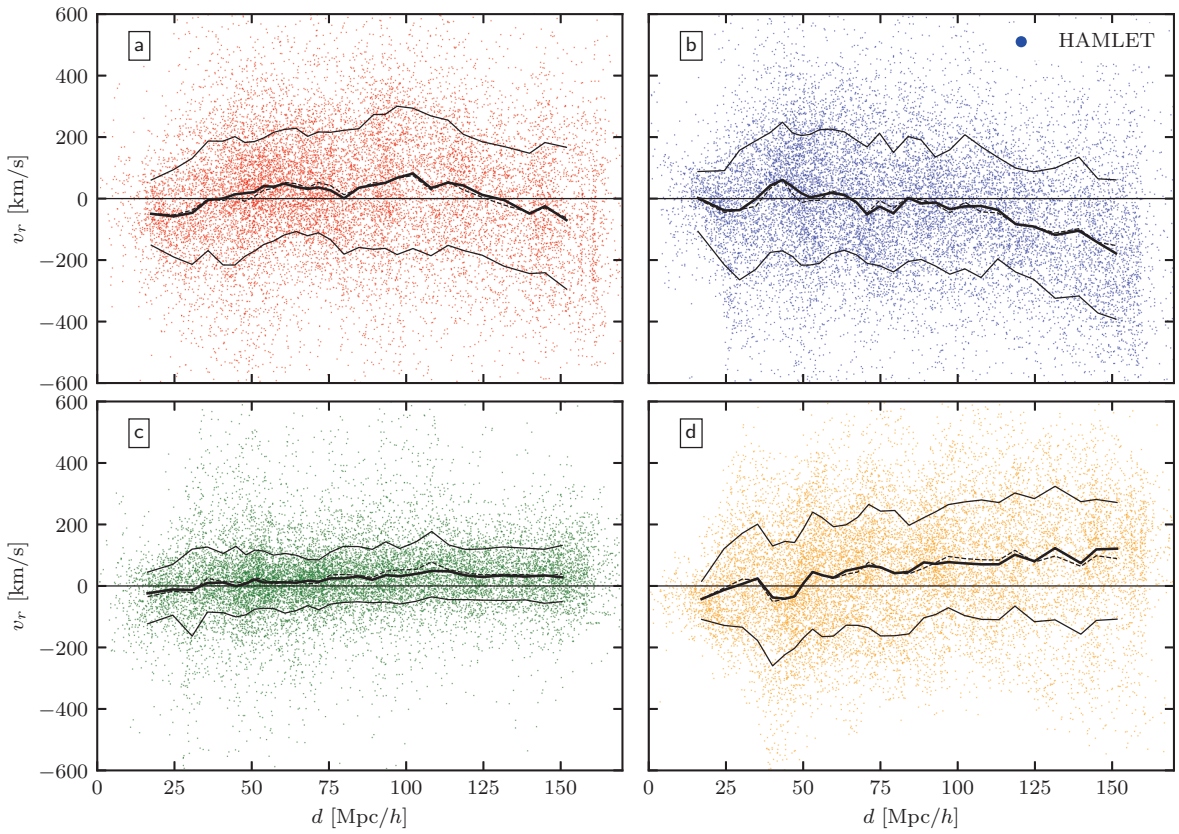

Figure 3.5: Scatter plots of the residual of the BGc/WF (panel a), of the HAMLET (panel b) and of the Ex/WF (panel c) reconstructed v_rs evaluated at the data points. The residual of the BGc/WF from the HAMLET reconstructed v_rs at the data points is shown as well (panel d). Each bins comprises the same number of points so as to avoid variations due to Poissonian statistics. The solid thick line represents the median, the thin solid lines delimit the 1-σ region around it and the dashed line shows the mean.

are linear reconstruction methods. Since these methods will never overcome this, we are only able to gauge the effect of observational errors on the reconstructions. This serves the purpose of testing the WF in the case where the only source of statistical uncertainty is the sampling. In other words in section 3.5, the reconstruction based on the BGc/WF, the Ex/WF and the HAMLET method are presented.

3.5 Results

The results are presented in three sub-sections where we (a) compare how the predicted constraints themselves differ from their real values (section 3.5.1); (b) examine the accuracy of the reconstruction of the cosmic fields (sections 3.5.2 and 3.5.3); and (c) compare the reconstructed monopole and dipole (*i.e.* bulk flow) multipoles with their target counterparts (section 3.5.4).

3.5.1 Reconstructed data

After applying the BGc/WF and HAMLET methods to the mock catalogues (as well as the WF to the exact, no error mocks), the first things to check is how well the distributions of radial peculiar velocities, distances and redshifts of the data points, match their target values. This is shown in fig. 3.4a,b and c, respectively, where the target curve represents the true distributions of the mock catalogue; namely, the closer the BGc/WF or the HAMLET curve is to the target, the more accurate the reconstructions. The values of the reconstructed v_r's of the mock data points are obtained by interpolation over the grid points. The distances are obtained differently for the different methods. The Ex/WF's distances are the true distances, thus they are identical to the target. For the BGc/WF method, the distances are the result of the application of the BGc to the data, before the WF is applied. Finally, for HAMLET , the

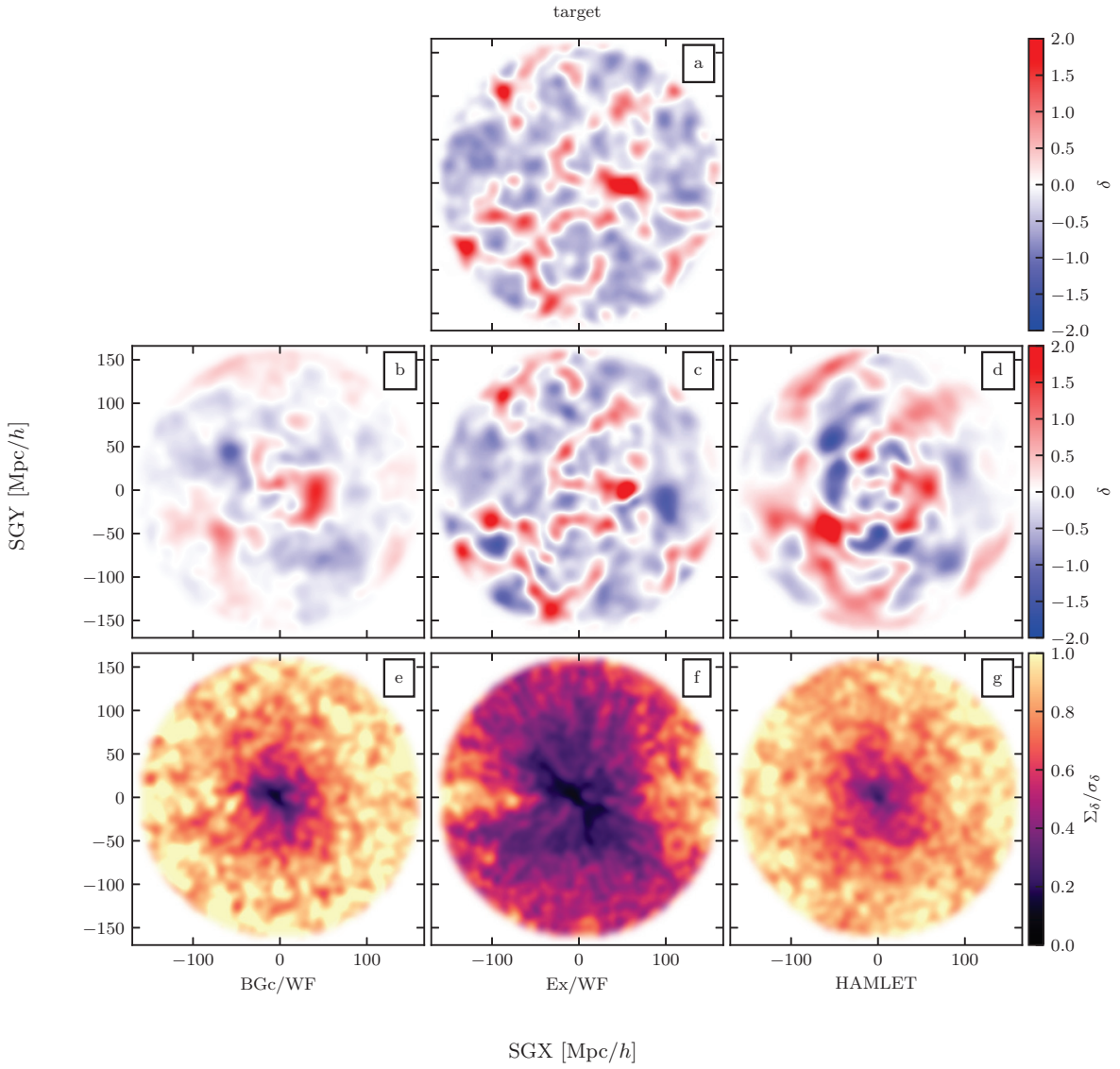

Figure 3.6: A comparison of the of the BGc/WF (left column), Ex/WF (central column) and the HAMLET (right column) reconstructed over-density fields with the target simulation. For consistency, the over-density plotted for the target field is the linear over-density *i.e.* the divergence of the velocity field. The middle panels present the reconstructed δ and the bottom ones show the constrained variance normalized by the cosmic variance, $\Sigma_\delta/\sigma_\delta$. All plots refer to the $SGZ = 0$ plane of the target simulation and all fields are Gaussian smoothed with a $5\,\mathrm{Mpc}/h$ kernel.

distance of each constraint is the mean of all the distances sampled over the Monte-Carlo steps.

We remind the reader that the Exact WF (green dashed) represents the limits of the WF method. fig. 3.4a shows that the HAMLET reconstruction method does an exceptional job at recovering the distribution of radial peculiar velocities. Note also that the WF in its purest form too recovers the target distribution. The BGc/WF struggles slightly by narrowing the data's distribution with a slight over emphasis on smaller values of the peculiar velocity at the expense of the large values. We note, as an aside, that the fact that the target (and hence the reconstructions) are not centered at $v_r = 0$ is due to the specific nature of the mock observer chosen (*i.e.* cosmic variance). The BGc/WF suppression of the reconstructed radial velocities of the data points relative to the target is inherent in the WF algorithm, where the estimated signal is the weighted "compromise" between the data and the prior model. Where the data is not very strong the WF estimator is biased towards the null field predicted by the prior.

In fig. 3.4b and c the distance and redshift distributions are examined. For both of these quantities the two reconstructions do a remarkably good job at matching the target, rendering their curves practically

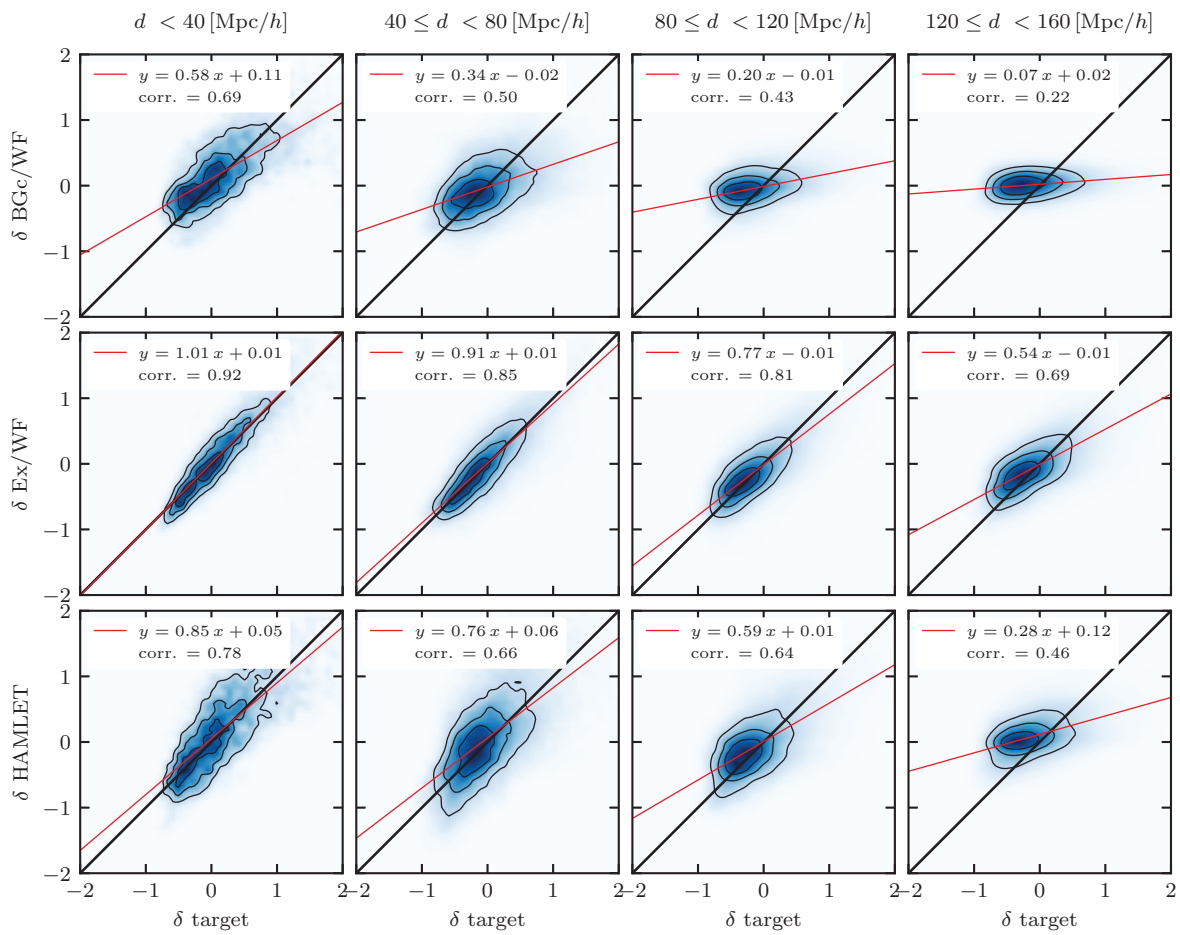

Figure 3.7: Density scatter plots of δ reconstructed versus δ target. Rows from bottom to top: HAMLET , Ex/WF BGc/WF. Columns from left to right: within spheres of 40, 80, 120, 160 Mpc/h. The red line represents the best fitted line whose line equation is $y = ax + b$. The parameters of the line and the Pearson correlation coefficient are given in the legend. The black line $y = x$ is shown for reference.

indistinguishable from the target. Note however that the BGc/WF method tends to "exaggerate" some of the peaks and valleys in the distance distributions (fig. 3.4b). All models reliably follow the input's form. In the absence of errors, *i.e.* the Ex/WF case, the reconstructed v_r's of the data points should be equal to the input constraints taken from the target simulation (Hoffman & Ribak, 1992).

The slight mismatch between the v_r histograms of the Ex/WF and the target seen in fig. 3.4 occurs because the Ex/WF histogram is an interpolation over the course grid of the WF.

It is important to understand by how much each constraint shifts during the BGc and reconstruction procedures. In fig. 3.5 the difference between the reconstructed v_r and the input v_r is compared on a constraint by constraint basis and as function of distance. From top to bottom this difference is shown for show the BGc/WF, HAMLET and the Ex/WF (respectively fig. 3.5a, b and c). The difference between the two main reconstruction methods (BGc/WF and HAMLET) is shown in the final panel, fig. 3.5d. In these plots each constraint is a dot, the median and mean values of the difference are shown as a solid and dotted black line, respectively. One standard deviation is indicated by the two thin black lines. An examination of fig. 3.5 reveals that the methods based on the WF tend to underestimate the v_r in the inner most distance shells (below $\leq 60\,\mathrm{Mpc}/h$) while overestimating it in the outer shells. This is sure even for the ideal case of the Ex/WF. The mean of HAMLET method, however (fig. 3.5b) indicates the constraints are not systematically shifted in the region $\sim 40 - 110\,\mathrm{Mpc}/h$, but underestimate v_r outside this range.

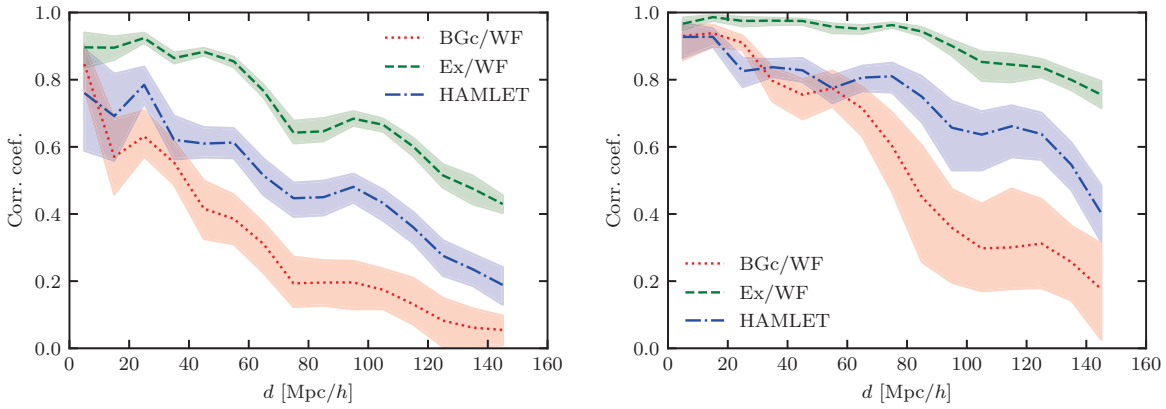

Figure 3.8: Statistics per shell of distance. Absolute value of the coefficient of correlation for δ between the different reconstructions and the target field. The error envelope represents the 2σ variation of the ensemble of realizations. **Left:** over-density field. **Right:** radial velocity field.

3.5.2 Reconstructed density maps

The non-linear density field of the target simulation cannot be directly compared with the reconstructed linear density field. To enable a meaningful comparison we compare the divergence of the velocity field of the two (*i.e.* eq. (1.38)), terming both of these δ, out of convenience.

In order to visually inspect the reconstructed density distribution, a $3.9\,\mathrm{Mpc}/h$ thick slab at the super galactic plane (SGZ=0) is chosen. This is not an arbitrary choice: given that the largest numbers of constraints are expected to lie in or close to SGZ=0, we expect this slab to be the most accurate. The fields are smoothed with a Gaussian kernel of $5\,\mathrm{Mpc}/h$.

fig. 3.6 examines the density distribution in this slab. fig. 3.6a is the target density distribution. The column below it (namely the middle column, fig. 3.6c, f shows the Ex/WF results, while the left column (fig. 3.6b, e) shows the BGc/WF results and the right column (fig. 3.6d, g) shows the HAMLET results. The middle row (panels b, c, d) shows the reconstructed density distribution. Some conclusions may be drawn from a visual examination of fig. 3.6b, c, d. The Ex/WF generally recovers the features of the local cosmography at all distances. The reconstruction is not exact; given that there are no "observational" errors here, this implies that the mismatch between the Ex/WF and the target (*i.e.* between fig. 3.6c and fig. 3.6a), on both small and large scales is due to the finite, inhomogenous and anisotropic sampling. On small scales there is an additional contribution due to non-linearities that are not modeled by the WF. Comparing the BGc/WF (fig. 3.6b) with the target indicates a decline of power of the reconstructed density field with the distance from the observer, yet the general structure of the cosmic web of over- and under-dense regions is recovered. The HAMLET reconstructed δ field does not exhibit the same loss of power as in the BGc/WF case but it suffers from a loss of spatial resolution with distance (fig. 3.6d). The more distant structures become more fuzzy and diffuse.

The bottom panels of fig. 3.6 present the constrained variance Σ^2_δ of the three reconstructed δ fields. It is defined as the local, cell by cell, variance calculated over an ensemble of CRs for the Ex/WF and BGc/WF case and over a set of independent states of the Monte Carlo chain in the HAMLET case. The panels show the square root of constrained variance normalized by the cosmic variance, $\Sigma_\delta/\sigma_\delta$. The cosmic variance is calculated by calculating the variance over all CIC cells in each reconstructed δ field. The value of $\Sigma_\delta/\sigma_\delta$ gauges the constraining power of the constraints and the assumed prior model. When this equals to 0 the region is highly constrained and when it equals unity the reconstructions are as random as cosmic variance. Thus one expects it to be small close to the observer and to approach unity asymptotically with distance.

fig. 3.6e, f, g quantifies what is visually apparent from (fig. 3.6b, c, d) namely that the inner regions are well constrained but that this fades with increasing distance. The reconstruction methods that include errors (*i.e.* fig. 3.6e,g) are never "perfect", while the Ex/WF method fig. 3.6f, does obtain values of $\Sigma_\delta/\sigma_\delta$ close to 0. Interestingly, the impact of the ZOA on the reconstruction method is apparent in fig. 3.6f. Here it causes a very clear limitation of the expected ability to reconstruct the density field.

The accuracy of the density field reconstructions - specifically their accuracy *as a function of distance* - is shown in fig. 3.7. These are scatter plots which compare, on a cell by cell basis, the density of the target

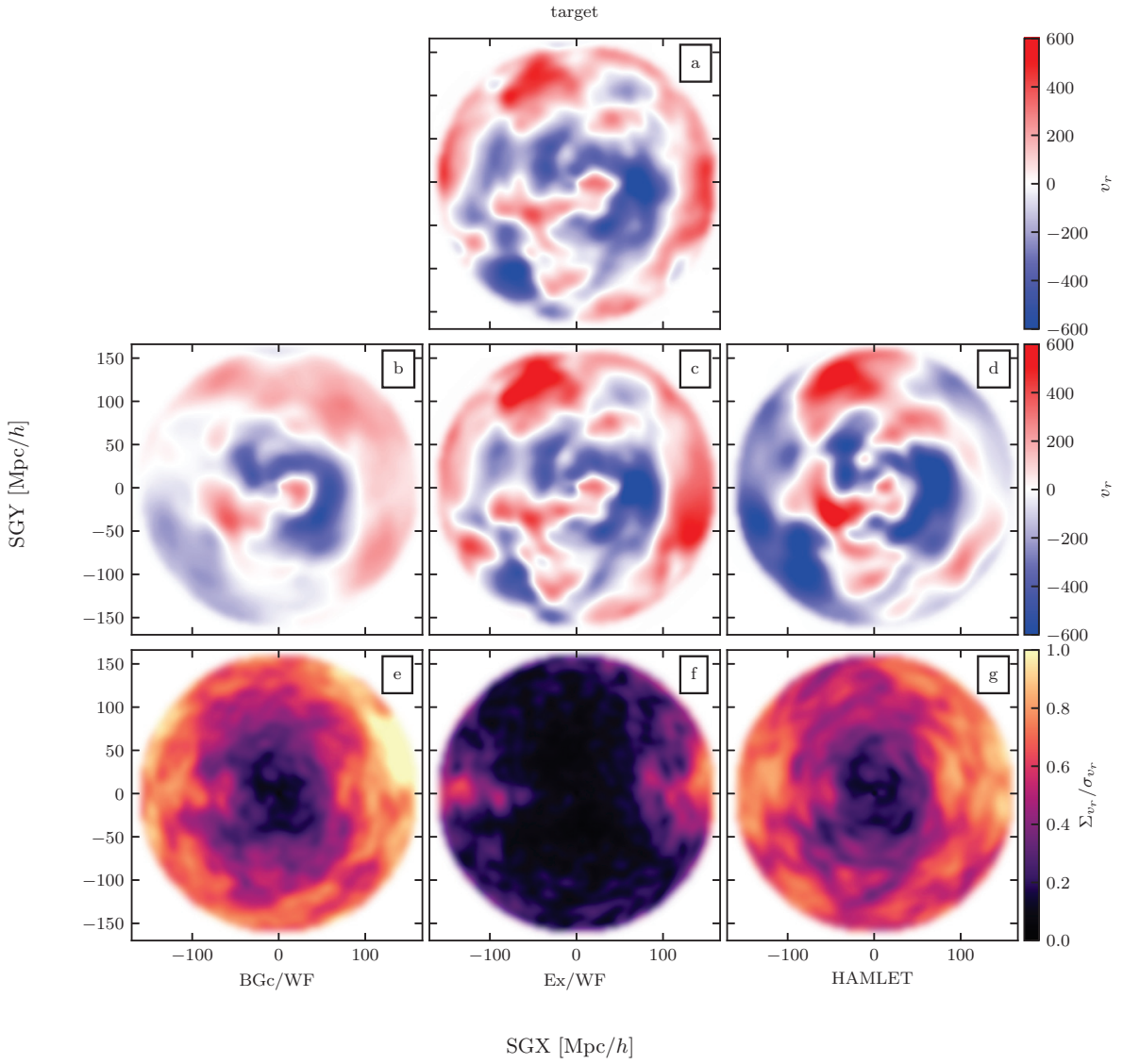

Figure 3.9: Same as fig. 3.6 for the radial component of the velocity field.

with the BGc/WF (top row), Ex/WF (middle row) and HAMLET (bottom row). The line, $y = ax + b$, (or $\delta_{\rm method} = a\delta_{\rm target} + b$, to be more precise), which best describes the scatter is shown in red; its slope, y intercept and the Pearson correlation coefficient is given in each sub-panel. In the ideal case where a reconstruction method perfectly matches the target this would simply be a slope of $a = 1$ and an offset or bias of $b = 0$ line with zero scatter (shown in black), with a Pearson correlation coefficient of unity. The columns in this figure denote different $40\,{\rm Mpc}/h$ thick radial shells under consideration. Note that a slope less than unity indicates that the reconstruction under-estimates the over-dense regions and over estimates the under-dense regions. A slope greater than unity represents the opposite (exaggerates over- and under-dense regions). An offset of $b \neq 0$ means a biased reconstruction.

There are a number of important features of this fig. 3.7. First, considering the inner most bin (leftmost column) the Ex/WF reconstruction recovers very well the density of the target. A slope of unity and practically null offset of $b = 0.01$ and a correlation coefficient of 0.92 indicates that in general in this region the Ex/WF reconstructions is very well recovered. This implies that the nearby sampling of the CF3 catalog is almost optimal. Obviously, the HAMLET and the BGc/WF methods do worse in recovering the density field. Moving to the outer shells all three density reconstruction systematically degrade with slopes and correlation coefficient decreasing. The slope in all cases is less than unity, indicating that the reconstructions suppress the power of the recovered density field. This diminishing of the power increases with the distance from the observer. The BGc/WF suffers more from the loss of

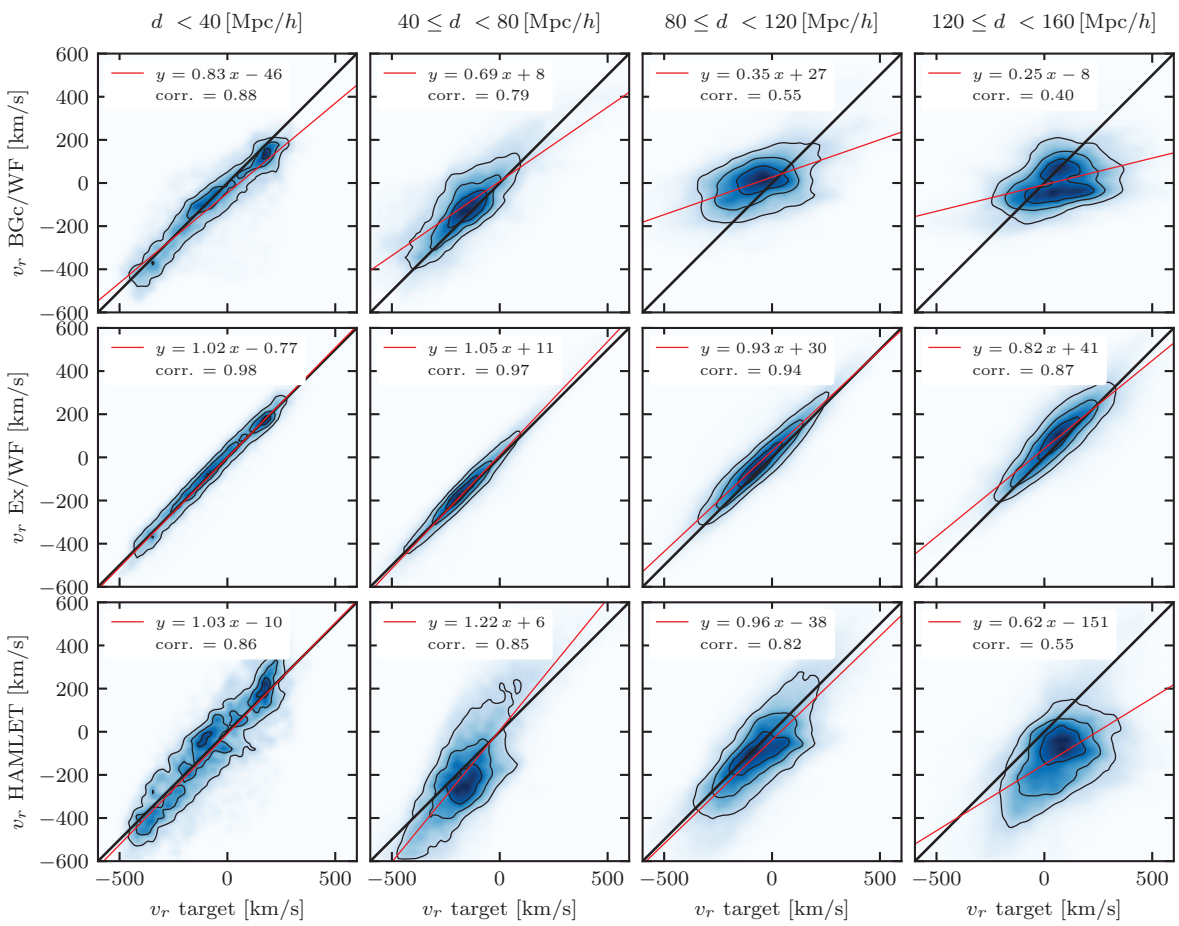

Figure 3.10: Same as fig. 3.7 for the radial component of the velocity field.

correlation with distance than the HAMLET . Yet, the latter reconstruction is at large distances with $b = 0.12$ for the distance range of $120 \leq d \leq 160\ Mpch$. The BGC/WF behaves, on the other hand, by the "Bayesian book" - where the sampling is very sparse and the errors are much larger than the signal, the unbiased ΛCDM prior is recovered.

The correlation between the reconstructed mean field and the target is shown as a function of distance in fig. 3.8. These are the correlation coefficients from the scatter plots (fig. 3.7) plotted as a function of distance in order to gauge the degradation of the reconstruction methods as data becomes more sparse and volumes become large. Note that the binning is different hence the non identical values of the correlation coefficient between the two plots. The solid lines in fig. 3.8 represent the mean correlation coefficient between the reconstruction and the target; the error corridor represents the 2σ variance about this mean. As expected the Ex/WF is always a superior to the BGc/WF and the HAMLET method. With the exception of the inner most bin, the HAMLET method achieved higher correlation coefficients than the BGc/WF method. At the edge of the data, no method achieves a correlation coefficient of greater than 0.5.

3.5.3 Reconstructed radial velocity maps

The examination of the radial component of the velocity field follows here that of the density field (see section 3.5.2). The same $SGZ = 0$ and $4\,\mathrm{Mpc}/h$ thick slab is shown in fig. 3.9. Again, the top panel is the target radial peculiar velocity field while the left column shows the BGc/WF reconstruction, the middle column the Ex/WF reconstruction and the right most column, the HAMLET reconstruction. The fields are smoothed with a Gaussian kernel of $5\,\mathrm{Mpc}/h$.

The radial velocity field (fig. 3.9b, c, d) appears much more accurately reconstructed than the density field. The same outflows and inflows are generally visible and the cosmographic landscape is recognisable in all three cases. Although the reader will note that the accuracy of the velocity reconstruction, like the

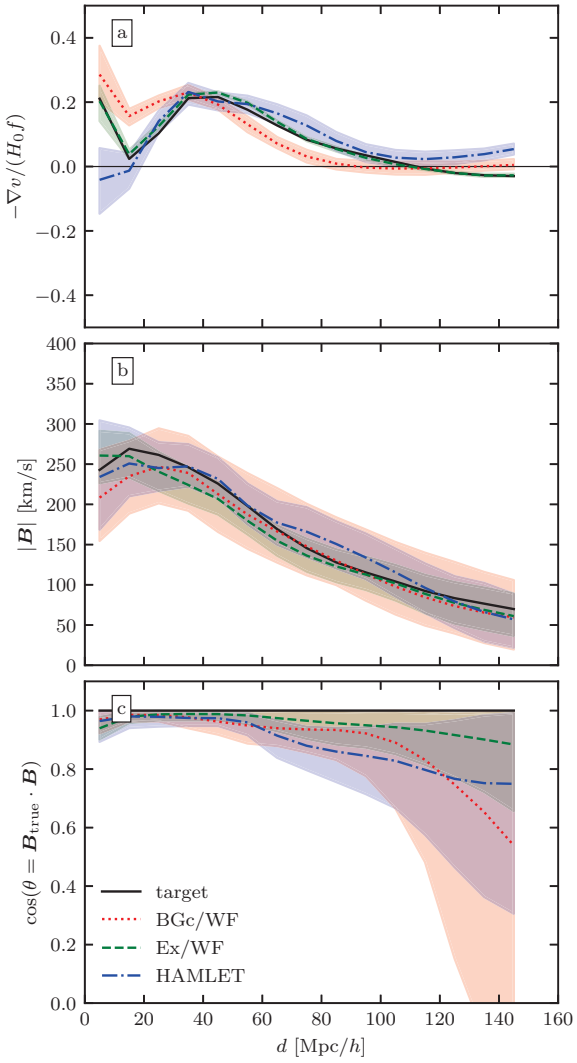

Figure 3.11: The monopole moment (upper panel), the amplitude of the dipole moment (*i.e.* the bulk velocity; middle panel), and the cosine of the angle of alignment between the reconstructed and target bulk velocities (lower panel) are shown. The profiles present the mean and "2σ" scatter of the mean profile in spheres of radius d. The reconstructions correspond here to the Ex/WF (green dashed line), the BGc/WF (red dotted line) and the HAMLET (blue dot-dashed line) case. The scatter is the constrained variance of the different reconstruction and the target simulation is presented by the black solid line (middle and upper panels).

density field, deteriorates at larger distances. Features are recognisable but distorted and smoothed out.

The constrained and cosmic variances of the radial velocity, Σ_{v_r} and σ_{v_r}, are calculated much in the same way as in for the density field (see section 3.5.2). The imprint of the ZOA is clearly seen in the $\Sigma_{v_r}/\sigma_{v_r}$ map of the Ex/WF map. Yet, in all cases considered here the constrained variance, normalized by the cosmic variance, is much smaller than in the density case. Namely, the velocity field is much more constrained by the CF3 data than the density field. In general the reconstructed HAMLET velocity field bares a closer resemblance to the target than the BGc/WF reconstruction.

A close inspection of fig. 3.9d uncovers one troubling feature. At the edge of the reconstructed volume, at distances close to 150 Mpc/h the reconstructed is "bluer" than in the corresponding target and the Ex/WF maps. Namely the HAMLET reconstructed velocity field has a spurious negative infall. This is a manifestation of the limitation of the method as described by Hinton et al. (2017).

Again, we turn to a scatter plot, on a cell by cell basis to quantify the quality of the reconstruction as a function of distance in fig. 3.10, which is structured identically to fig. 3.7 - namely radial extent increasing column wise from left to right, while the rows from top to bottom being BGc/WF, Ex/WF and HAMLET .

This figure is qualitatively identical to its density field counter part (fig. 3.7) in that the same behavioural trends between the different reconstructions methods and as a function of distance exist. The correlation analysis of the Ex/WF and BGc/WF cases behaves much in the same way as for the density field - a degradation of the correlation with distance, a slope (a) that is close to unity nearby and diminishes with distance, and essentially with zero offset ($b \sim 0$). Yet, the quality of the reconstruction of the radial velocity is much better than that of the density. The HAMLET reconstruction shows a somewhat unexpected behaviour. The slope of the best fit line for the distance range of $40 \leq d \leq 80$ Mpc/h exceeds unity, $a = 1.22$, *i.e.* there is an excess of power compared with the target and the Ex/WF cases. This effect is likely symptomatic of the so-called Inhomogenous Malqmist Bias (see section 1.5). Indeed, in the current model of HAMLET , the positions of the probes are not directly correlated to the density field, whereas in reality, galaxies tend to be found in over-dense regions (*i.e.* they tend to cluster). This bias leads to an over-estimation of the contrast, seen here as a slope $a > 1$. This bias doesn't show itself in the outer bins (*i.e.* 80 - 160 Mpc/h) since here it is counteracted by the loss of contrast due to the decrease in sampling. The best linear fit forthe range of $120 \leq d \leq 160$ Mpc/h yields a significant negative offset of $b = -151$ km/s, in agreement with the visual inspection of fig. 3.9g.

The correlation of the radial component of the velocity field between the reconstructions and the target is shown as a function of distance in fig. 3.8, right panel. Similar to fig. 3.8 left panel, these are the correlation coefficients computed from scatter plots (fig. 3.10) plotted as a function of distance in order to gauge the degradation of the reconstruction methods as data becomes more sparse and volumes become large. The solid lines in fig. 3.8 represent the mean correlation coefficient between the reconstruction and the target; the error corridor represents the 2σ variance about this mean. As expected the Ex/WF is always a superior to both the BGc/WF and the HAMLET method. The Ex/WF reconstruction is very well correlated with the target out until ~ 80 Mpc/h, beyond which it begins to drop, although it is worth noting that it stays correlated for the full sample. This drop is a manifestation of the sampling and the decreasing number of the data (per volume) at these distances. The HAMLET and BGc/WF method are roughly equal in the inner regions out to ~ 70 Mpc/h, and beyond it the HAMLET method provides better correlation. At the edge of the data, no method achieves a correlation coefficient of greater than 0.5.

3.5.4 Multipole moments of the reconstructed velocity field

The first two moments of the velocity field, the monopole and dipole, are examined here. The effect of errors and sampling on the fidelity of these two physical quantities is of particular interest since the monopole and dipole are often used as probes of the scale of homogeneity and can affect probes of the cosmological model in particular.

fig. 3.11a shows the target and reconstructed velocity monopole as a function of distance. The same colouring and line style convention used in fig. 3.8 is adopted here too, with the moments of the target simulation plotted in black. Note that the monopole - the mean infall or outflow of matter, is the zeroth order moment of the velocity field. It is the mean of the divergence of the velocity field in spheres of radius d and as such is called the "breathing mode" of the velocity field. In the linear theory of the cosmological gravitational instability the density and velocity fields are related by eq. (1.38), hence we opted here to present the monopole term by means of this equations. Thereby, fig. 3.11a effectively presents the mean linear density with spheres of radius d. The Ex/WF is nearly indistinguishable from the target here: the error corridor (which corresponds to variance across all the constrained realisations) is tiny and the black and green dashed line are practically on top of each other.

The BGc/WF curve overestimates the monopole in the inner parts (within ~ 50 Mpc/h) while underestimating it outside that range. This increased monopole implies an overestimation of the density in the inner parts of the mock universe, which is confirmed by examining the equation of the best fit line to the scatter plot fig. 3.7 (upper row, left column, $d < 40$ Mpc/h). The best fit line has has an offset of $b = 0.11$, meaning that there is a systematic increase in the estimated densities, consistent with the higher monopole. Both the reconstructions and the target tend to zero infall at these large scales. The HAMLET method on the other hand behaves inversely to the BGc/WF method, underestimating the target monopole at small scales and over estimating it at large scales. The HAMLET 's monopole term at the edge of the data reveals an excess of density at $d \sim (120 - 150)$ Mpc/h, in agreement with fig. 3.7 (lower/right panel). Otherwise, the HAMLET method succeeds in tracking the target monopole over a large range from ~ 20 to ~ 100 Mpc/h.

fig. 3.11b and c shows the second moment of the velocity field, namely the dipole or the bulk flow. fig. 3.11b refers to the magnitude of the bulk flow, while fig. 3.11c refers to its direction. Accordingly

all methods do a fine job of recovering the magnitude of the bulk flow beyond around $\sim 30\,\mathrm{Mpc}/h$. The Ex/WF has, predictably, a smaller error corridor than the other two methods, which are roughly similar in size. With respect to direction, fig. 3.11c shows the dot product between the target bulk flow direction and the reconstructed one (hence in this plot there is no black target line). The bulk flow directions for the Hamlet and BGc/WF method are aligned to within ~ 15 deg of the target out to a distance of $\sim 50\,\mathrm{Mpc}/h$, while the Ex/WF is well aligned to greater distances. Note that however, even the Ex/WF curve begins to deviate significantly at the reconstructed edge. This indicates that even in the best case scenario of zero errors, sampling at these great distances is a limiting factor in terms of recovering the direction of the cosmic dipole. Note that the accuracy recovered here is also restricted in its ability to recover the underlying dipole direction by the limited depth of the survey (Nusser, 2014). The problem is exacerbated when examining the BGc/WF and Hamlet curves at large distances. fig. 3.11 indicates that although the monopole and dipole are well recovered across a large range, the direction of the reconstructed dipole begins to deteriorate when the sampling drops.

3.6 Summary

The reconstruction of the large scale density and velocity fields from Cosmicflows-like databases of galaxy distances, and hence peculiar radial velocities, is challenging. The data is sparse, extremely noisy with Noise/Signal ratio larger than a few for the majority of the data, non-uniformly and anisotropically distributed. Furthermore the data suffers from the log-normal bias, which leads to a non-linear bias in the estimated distances and velocities.

A number of independent methods have been developed to reconstruct the local LSS and to produce constrained initial conditions for cosmological simulations designed to reproduce our local patch of the Universe (i.e Sorce, 2015). What is generally missing from the literature in this field is an understanding of the accuracy of these methods. Often the reconstructions are applied directly to observational data and only very limited conclusions can be drawn on the viability of a cosmography. The present paper compares the BGc/WF (Hoffman et al., 2021) and the Hamlet algorithms (see chapter 2) by testing them against a carefully crafted mock of an observational catalogue (an improved CF3-like survey) drawn from one the MultiDark cosmological simulations.

The quality of the reconstruction is gauged by studying the residual between the reconstructed and target density and velocity fields. The residual is mostly analyzed by quadratic measures and as such it is characterized by the mean and variance of the distribution. An optimal reconstruction should make the mean of the residual to be as close as possible to the null field and aim at minimizing its variance. A related measure is the linear correlation analysis which yields the best "line", $y = ax + b$, that fits the linear dependence of reconstructed field on the target one, and the Pearson correlation coefficient. The values of the offset, b, for the case of the linear over-density and for the radial velocity are consistent with zero fro the BGc/WF, in agreement with the theoretical expectations. The distant data points are extremely noisy and very sparsely distributed, hence the WF reconstruction is dominated by the ΛCDM prior model. The Hamlet 's significant offset is however inconsistent with the prior model.

We define here three different regions: the nearby ($d \lesssim 40\,\mathrm{Mpc}/h$), the intermediate ($40 \lesssim d \lesssim 120\,\mathrm{Mpc}/h$) and the distant one ($d \gtrsim 120\,\mathrm{Mpc}/h$). Based on the above criteria we conclude that nearby the BGc/WF and the Hamlet methods are doing roughly equally well. The methods diverge at large distance - with the Hamlet outperforming the BGc/WF with a tighter correlation and smaller variance but underperforming in terms of the bias. This is most noticeable for distant region (the right columns of fig. 3.7 and fig. 3.8).

The three panels of Fig. 3.11 deserve a special attention here. The upper panel shows the radial profile of the monopole moment. The four profiles shown there - target, Ex/WF, BGc/WF and Hamlet - are all constructed under the assumption of ΛCDM value of $H_0 = 67.7\,\mathrm{km/s/Mpc}$. Yet, the negative offset of the monopole moment at the edge of the data implies that the local value of H_0 is somewhat smaller than its global value. A phenomenon expected for any finite volume realization in the ΛCDM cosmology (see Hoffman et al. (2021) for a quantitative assessment). A proper adjustment of the local value of H_0 would bring the target and Ex/WF profiles to converge to zero at the edge of the data, together with the BGc/WF asymptotic value. This would leave the Hamlet positive offset standing out with a systematic bias. The amplitude of the dipole moment, namely the bulk velocity, is recovered equally well by the three reconstruction and is in very good agreement with the target. The bottom panel shows the cosine of the angle between the reconstructed and the target bulk velocities. The BGc/WF behaves as expected - the mean misalignment is consistent with the full alignment to within one σ of the

constrained variance. This is not the case with the HAMLET reconstruction, where the misalignment is more than 2σ away from the expected alignment.

Our overall assessment of the HAMLET and the BGc/WF reconstructions is that the former outperforms the latter one in terms of reduced scatter and tighter correlation between the reconstructed and the target density and velocity fields. Yet, the HAMLET suffers from biases in the reconstructed LSS at the distant regime - ones that do not appear in the BGc/WF reconstruction. It follows that the HAMLET should be the method of choice for the reconstruction of the LSS and the study of the cosmography of our local patch of the Universe. The BGc/WF reconstruction is the preferred tool for performing quantitative analysis and parameters estimation and possibly also for setting initial conditions for constrained cosmological simulations. One last comment is due here. The WF/CRs is a very well tested approach that is based on a solid theoretical foundations (Hoffman & Ribak, 1992; Zaroubi et al., 1995; Zaroubi et al., 1999). As such it provides an attractive framework for performing Bayesian reconstruction of the nearby LSS. Yet, any bias in the observational data and in particular the log-normal one needs to be addressed and apply outside that framework in some ad-hoc and approximate way. The HMC methodology, and in particular its HAMLET implementation, still suffer from some teething problems that need to be overcome, such as a proper modeling of the so-called Inhomogenous Malmquist Bias and of the selection function. The ability of the MCMC methodology in general and the HMC in particular to address the issue of reconstruction of the LSS, the handling of observational biases and the estimation of cosmological parameters within one computational self-consistent framework makes HAMLET a very attractive tool in the CLUES' toolbox (Yepes et al., 2009; Doumler et al., 2013c; Sorce et al., 2014; Sorce et al., 2017; Sorce & Tempel, 2017). The incredible improvement in the computational efficiency of the HAMLET compared with previous implementation of MCMC algorithms makes it even more promising for future implementations within the CLUES project.

Chapter 4

Application to Cosmicflows-4

4.1 Introduction

Our knowledge of the Local Universe has grown at an ever increasing pace over the last decades. Since the discovery by the CfA survey in the 1980s that galaxies are not uniformly distributed but rather that they form a Large Scale Structure (LSS) (CfA Wall, Sloan Great Wall, etc; *e.g.* de Lapparent et al., 1986; Gott et al., 2005), until today, our understanding of how our local cosmographic landscape has grown out of an initially smooth homogeneous distribution of perturbations has deepened considerably. In cosmographic efforts to chart maps of the heavens, we have come to name such features of the galaxy distribution.

Peculiar velocities of galaxies can be used as tracers of the full matter density distribution of the Universe. Unlike reconstruction from pure redshift surveys, they do not suffer from Redshift Space Distortion (RSD; see section 1.4.7) nor need any assumption on the galaxy bias (see section 1.3.3). Instead, exactly because peculiar velocities are gravitational velocities and are due to the net gravitational field, a relatively small sample accurately trace the full density distribution of the universe.

In order to recover the peculiar velocity of a galaxy (on the order of 300km/s in ΛCDM , see sections 1.6 and 1.8) two measurements need to be made: its redshift and its distance. From these two measurements a peculiar velocity can be inferred (see section 1.4.6). Estimations of peculiar velocities are primarily curbed by the difficulty in estimating the distance of a galaxy (since redshift measurements, when available, tend to be fairly precise; see sections 1.1.2 and 1.1.3). This leads peculiar velocity data to be (relatively) rare and extremely noisy.

Although some standard candles are highly accurate (*i.e.* Surface Brightness fluctuations or the magnitude of stars measured at the Tip of the Red Giant Branch (TRGB; Ferrarese et al., 2000)), all of which are accurate to 5% to 10%) these have their limits. TRGB tends to be unobtainable at large distances due to inherent technological limitations (*i.e.* the ability to resolve stellar populations at cosmological distances; see section 1.1.2). Other methods, that allow for a distance measurement at cosmological distances include scaling relationships such as Tully-Fisher (Tully & Fisher, 1977, TF;) and Fundamental Plane (Faber & Jackson, 1976; Djorgovski & Davis, 1987). Here however the error is around 18% to 25% and exceeds the signal at $cz \approx 10^4$ km/s. It increases with distance reaching a factor of 30 at at the edge of the CF4 data, $cz \approx 30\,000$ km/s.

However, despite the large errors associated with using scaling relations as distance measures and hence peculiar velocity, what they lack in precision they make up for in sample size. Scaling relations are relatively "cheap" and outnumber more accurate measures immensely (*e.g.* Tully et al., 2016; Tully et al., 2023). For all the reasons mentioned above, powerful methods of reconstructions need to be employed in order to extract the peculiar velocity signal from such a complex, inhomogeneous and error prone data.

In this work, we apply the Hamlet method of chapter 2 (published as Valade et al., 2022) to the Cosmicflows-4 catalogue (CF4; Tully et al., 2023): the largest and the most complete catalogue of peculiar velocities to date. With more than 38 000 groups of galaxies, it extends the previous Cosmicflows catalog (CF3; Tully et al., 2016) by adding more than 5 000 distance measurements within $cz \lesssim 10\,000$ km/s, as well as the entire SDSS-PV sample bad citation (Howlett et al., 2022), which contains more than 20 000 FP measurements up to $cz \lesssim 30\,000$ km/s in the SDSS region.

As seen in previous chapters, the Hamlet code is a very efficient GPU-accelerated implementation of the Hamiltonian Monte Carlo exploration of a posterior probability designed to reconstruct the linear over-density and velocity fields from measurements of the radial component of peculiar velocities. The

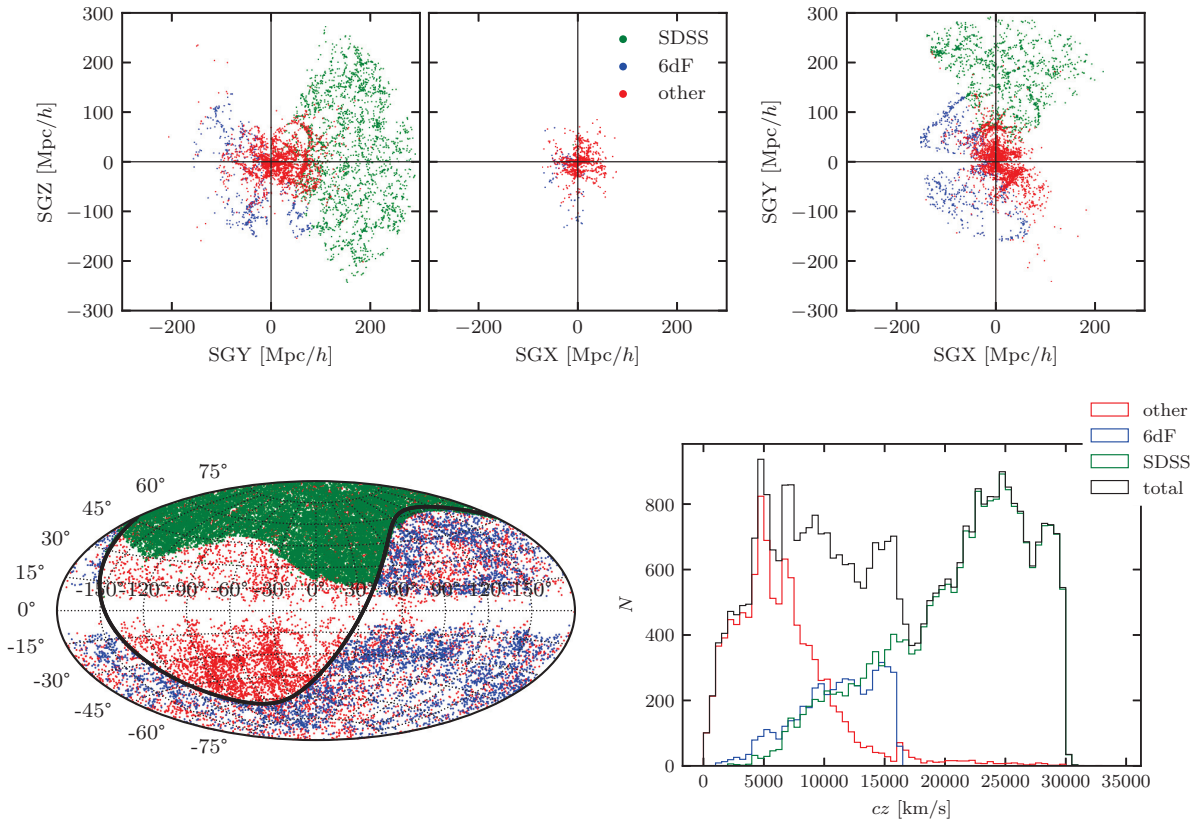

Figure 4.1: The distribution of the various subsamples of the grouped CF4 data: SDSS (green), 6dF (blue) and all the others (red): The distribution in ± 10 Mpc/h slices around the three principal Super galactic planes, where galaxies redshift distances are taken as proxy to actual proper distances (upper row); The angular distribution is presented in Aitoff projection in Galactic (l, b) coordinates (lower-right panel, the black line indicates the delimitation between celestial north and south hemisphres); The redshift distribution of the CF4 data points, with the clear redshift cuts for the SDSS and 6dF respectively at $cz = 30\,000$ km/s and $cz = 16\,000$ km/s (lower-right panel.

method has been tested on mocks in chapters 2 and 3 (published as Valade et al., 2023).

In this chapter we first introduce our interpretation of the data in section 4.2. We then briefly introduce an improvement of the modelization of the selection function in section 4.3. A cosmographic description of our reconstruction of the matter distribution in the Local Universe in done in section 4.4.3, where a qualitative comparison with the LEDA galaxy catalogue is carried. We then proceed to extensively analyses the velocity field, first by studying its Basins of Attractions (BoA) in section 4.4.4, its moments in section 4.4.4 and the LSS classification by the V-Web in section 4.4.4. Finally, a series of visualizations done by Daniel Pomarède[1] are presented in section 4.4.5.

4.2 Data

This work presented here is entirely based on the grouped Cosmicflows-4 catalogue (CF4; Tully et al., 2023). This catalogue of galaxies is not one survey but rather the careful compilation of several independent surveys. This makes it the biggest self-consistent catalogue of galaxies peculiar velocities with about 38 000 entries. However its footprint and selection function are very complex to describe given the heterogeneous nature of the catalogs it is based on. For each entry in the CF4 catalog (ie galaxy or group), we consider 5 quantities: its angular sky position (RA and dec), its redshift (cz) where c is the speed of light, its distance modulus (μ) and its uncertainty on the distance modulus (σ_μ). An uncertainty on the redshift of $\sigma_{cz} = 50$ km/s is assumed for all entries and aims at reproducing the error

[1]Institut de Recherche sur les Lois Fondamentales de l'Univers, CEA Université Paris-Saclay, F-91191 Gif-sur-Yvette, France, daniel.pomarede@cea.fr.

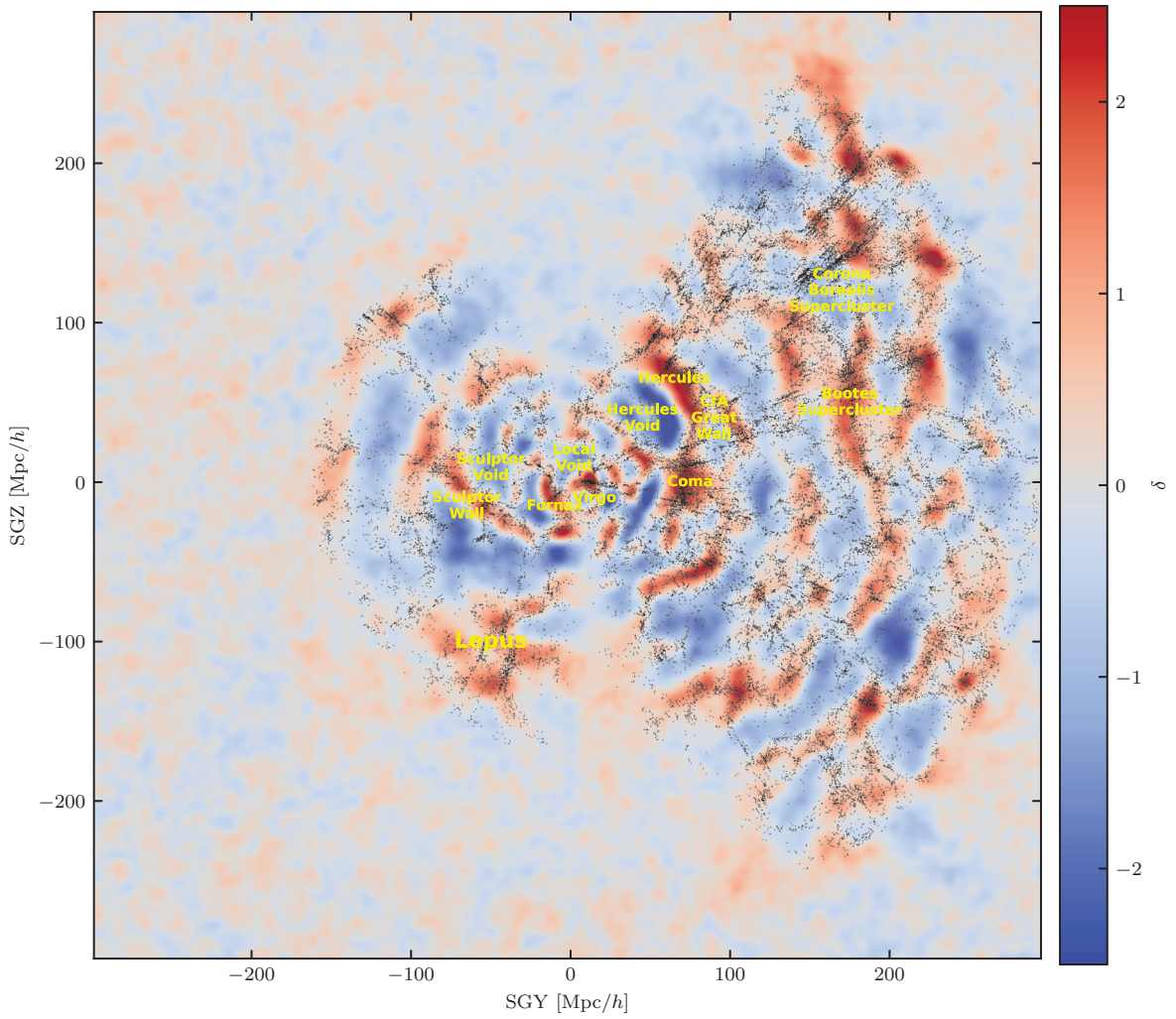

Figure 4.2: Over-density field at SGX = 0 and LEDA galaxies within $-5 < \mathrm{SGX} < 5\,\mathrm{Mpc}/h$ at their redshift position. The labels of some prominent features of the Local Universe are given in yellow.

size of spectroscopic redshift measurements. Errors both on the distance modulus and on the redshift are considered as normally distributed. The catalogue comes with an estimation of $H_0 = 74.6\,\mathrm{km/s/Mpc}$ that minimizes the flow at the edge of the data in accordance with the hypothesis of an homogeneous universe. This is the value taken for our reconstruction. The grouping of the data is a complex topic that is not discussed in this chapter. Groups and single galaxies are treated equally by our method, and indifferently named "constraint" or "entry".

CF4 is composed of about a dozen different sources, some have thousands of constraints, some just a few. We separate the two main sources from the rest, thus splitting CF4 in three sub-catalogues: *SDSS*, 6dF and *other* . An entry (either a single galaxy or a group of galaxies) is associated to

— **SDSS** if it contains more than one Fundamental Plane (FP) measurement (Howlett et al., 2022), has a declination greater than $-3.5\,\mathrm{deg}$ and a positive galactic latitude;

— **6dF** if it contains more than one FP measurement (Campbell et al., 2014), has a negative declination

— **other** in the remaining cases.

The small overlap between SDSS and 6dF in the region $-3.5 > \mathrm{dec} > 0$ is resolved in favor of 6dF.

This division is motivated by the very different sky coverages and redshift distributions of the different surveys, which should ideally be treated separately. The *other* sub-catalogue is fairly isotropic and

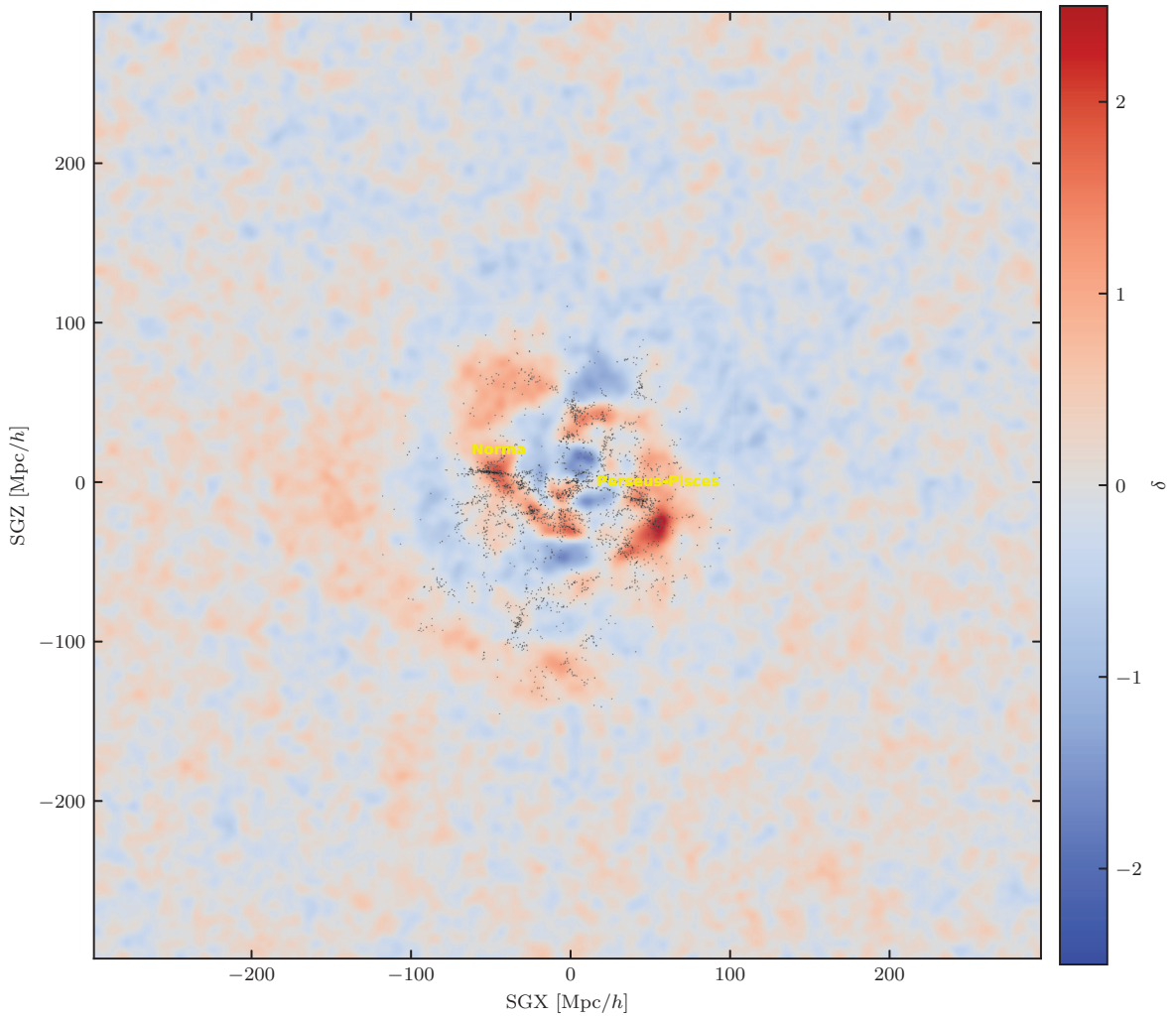

Figure 4.3: Same as fig. 4.2 but for the SGY = 0 plane.

does not need sub-dividing, and contains mostly TF measurements (Kourkchi et al., 2020) but also higher quality Ssupernovae and TRGB measurements. The SDSS, 6dF and *other* sub-catalogue comprise respectively about 22 000, 5 000 and 10 500 entries.

Figure 4.1 shows three different views of the data. First, we focus on the distribution in the three principal super-galactic planes, and plot for each plane the points in a 10 Mpc/h slice. This gives a good overview of the spatial coverage of CF4 as a whole, but it also highlights the differences between the sub-catalogues. SDSS data covers a very large volume in the galactic north, but is however restricted to a rather small solid angle. The 6dF sub-catalogue covers the celestial south almost entirely and stops in the galactic north where SDSS begins. The *other* catalogue is dense in all direction up to 100 Mpc/h. The Zone of Avoidance (ZoA) obscures the SGY = 0 plane, leaving only a few points within less than 100 Mpc/h. The second view is the projection of the data on the sky. Here, the angular separation of the different sub-catalogues is striking, and the limitations of SDSS and 6dF respectively to the celestial north and south becomes clear. The ZoA can be seen here obstructing the Universe by about ±10 deg around the galactic disc. Finally, looking at the redshift distributions, the distinct depths of the sub-catalogues become apparent. The *other* sub-catalogue dominates within 10 000 km/s ≈ 100 Mpc/h. Both SDSS and 6dF show dramatic redshift cuts at respectively 16 000 km/s ≈ 160 Mpc/h and 30 000 km/s ≈ 300 Mpc/h.

Note that the 6dF selection function drops almost immediately following the local maxima (a mere 3 bins later the histogram has dropped to zero) while the SDSS selection function has a slightly "fuzzier" cut off. As we will see, this sharp cut off has an affect on the reconstructed density and velocity filed at these distances.

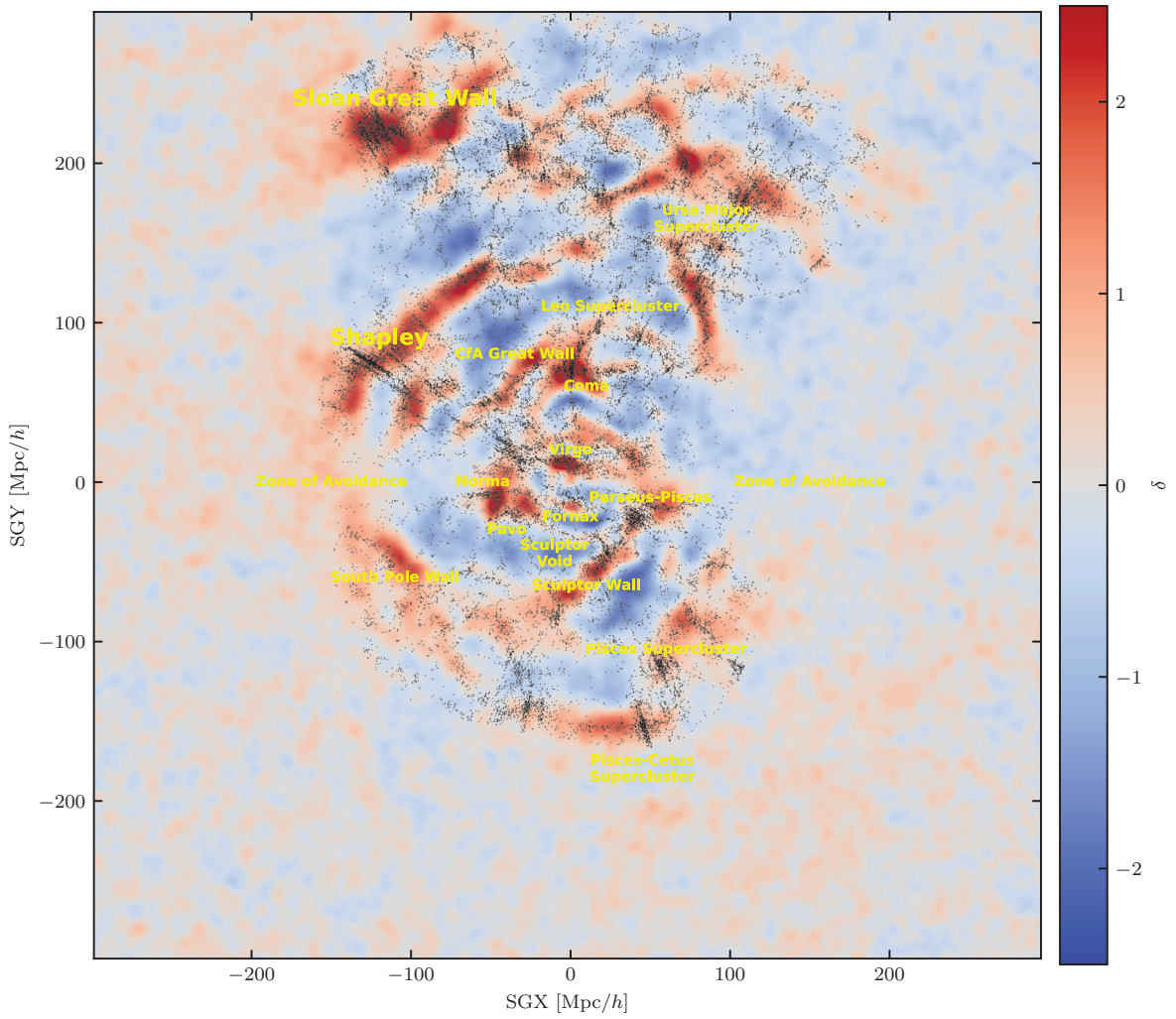

Figure 4.4: Same as fig. 4.2 but for the SGZ = 0 plane.

4.3 Adapting the model to CF4

4.3.1 Prior on distances

Modeling the selection function of the CF4 catalogue is very tedious for the reasons mentioned above. Therefore, after separating the full catalogue into SDSS, 6dF and *other* (as described in section 4.2), we model the prior on each entry's distance as the histogram on the redshifts distances smoothed with a 300 km/s Gaussian kernel (to account for peculiar motions). Such a model helps reduce the Inhomogeneous Malmquist Bias described in section 1.5.4 and Boruah et al. (2022).

4.3.2 Modeling the redshift cuts

The edge of the SDSS and 6dF data are characterized by sharp redshift cuts. If not taken into account, these may lead to a spurious infall at the edge of the data due to the "inaccessible" values of redshifts and thereby to an overestimation of the density field at the edge of the data section 3.6. Therefore it is necessary to properly undo this effect.

The likelihood $L(z_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL})$ of measuring a redshift (z_i) given the model's parameters has to be renormalized in the case that not all values of (z_i) are accessible. Following Strauss & Willick (1995);

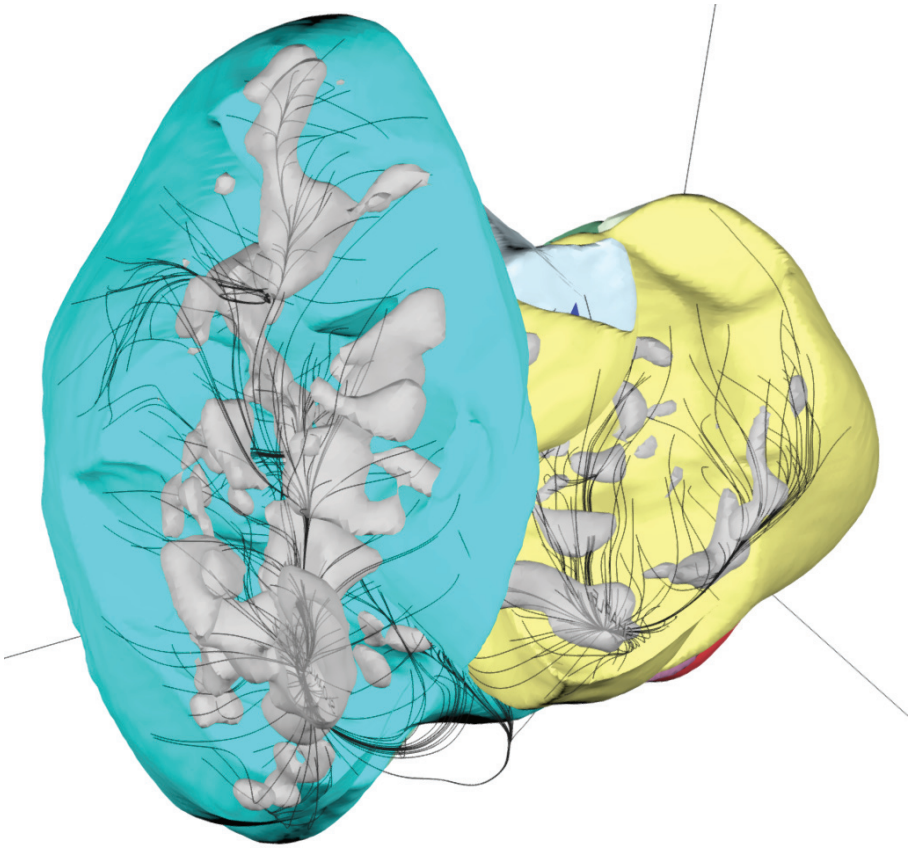

Figure 4.5: Screenshot of the 3D visualization of the flow lines (in black), with different Basins of Attraction (BoAs) colored. The blue and yellow region delimits respectively the BoAs converging on the Sloan Great Wall and Shapley. The other BoAs are named in the following figs. 4.6 and 4.7. The visualization is cropped to the reconstructed volume. The 3D visualization can be found here.

Hinton et al. (2017), we rewrite eq. (2.19):

$$L(z_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL}) \to \frac{L(z_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL})}{\operatorname{erfc}\left(\frac{z(\Delta_k, d_i) - z_i^{\rm cut}}{\sqrt{2}\gamma(d_i)}\right)}, \tag{4.1}$$

$$z(\Delta_k, d_i) = \big(1 + \bar{z}(d_i)\big)\big(1 + \boldsymbol{v}(d_i\hat{\boldsymbol{r}}_i|\Delta_k) \cdot \hat{\boldsymbol{r}}_i/c\big) - 1, \tag{4.2}$$

$$\gamma^2(d_i) = \frac{1}{c^2}\left(\sigma_{cz}^2 + \big(1 + \bar{z}(d_i)\big)^2\sigma_{\rm NL}^2\right) \tag{4.3}$$

where $z_i^{\rm cut}$ depends on the sub-catalogue to which the entry belongs. This value is set to $cz_{\rm SDSS}^{\rm cut} = 30\,000\,\mathrm{km/s}$, $cz_{\rm 6dF}^{\rm cut} = 16\,000\,\mathrm{km/s}$, while the sub-catalogue *other* does not have any redshift cut and thus does not undergo this re-normalization process. To insure full consistency after this renormalization procedure has been completed, if a SDSS or 6dF redshift is greater than the redshift cut of that catalogue, the constraint is removed from the catalogue, which in practice concern a few dozen of galaxies. The renormalization effectively bends the likelihood $L(z_i|\Delta_k, \mathcal{D}, \sigma_{\rm NL})$ towards higher values of reconstructed redshifts $z(\Delta_k, d_i)$, allowing for a larger outflow at the edge of the data.

4.4 Results

4.4.1 Hyper-parameters of the reconstruction

A grid of size 256^3 covering a $1\,[\mathrm{Gpc}/h]^3$ cubic volume is employed in the reconstruction. The resulting cell size is $3.9\,\mathrm{Mpc}/h$ per grid cell side. Such a grid can capture density wave numbers in the range $k_{\rm min} = 6.28 \cdot 10^3\,h/\mathrm{Mpc}$ to $k_{\rm max} = 1.39\,h/\mathrm{Mpc}$, *i.e.* wavelengths in the range $\lambda_{\rm min} = 4.5\,\mathrm{Mpc}/h$ to

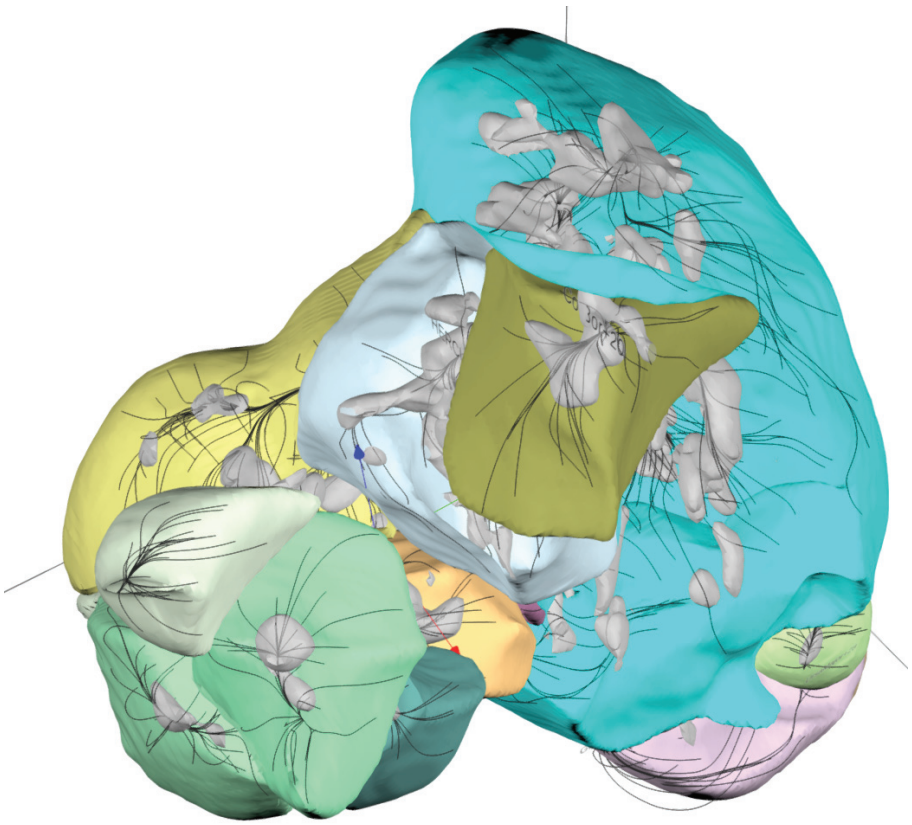

Figure 4.6: Same as fig. 4.5 with another view angle. The BoAs visible in this figure are: Sloan Great Wall (blue), Shapley (yellow), Hercules (white), Corona-Borealis (olive green), Perseus-Pisces (orange), Pisces-Cetus (light green, left), Pisces (light green, right) and Funnel (dark green). The yellow green BoA on the left of the reconstructed region is unnamed.

$\lambda_{\max} = 1\,\mathrm{Gpc}/h$. All the modes of the grid are taken into account, leading to $256^3 = 16\,777\,216$ free parameters representing the reconstructed field. The cosmology is fixed to a flat Universe with $H_0 = 74.6\,\mathrm{km/s/Mpc}$, $\Omega_m = 0.3$, $\Omega_b = 0.0471$, $\sigma_8 = 0.8228$ and the linear growth factor is $f = 0.51$.

4.4.2 Comparison to an independent redshift galaxies catalogue

The reconstruction method has been thoroughly tested against mocks in chapters 2 and 3. However, when it comes to reconstructing the real Local Universe, quantifying the credibility or the quality of the recovered map is slightly nuanced. Borrowing from previous work with the same goal (*e.g.*; Graziani et al., 2019), the most direct way to do so is to over-plot on the field a catalogue of galaxies independent of the one used for the reconstruction. Indeed, the galaxy distribution is supposed to be strongly correlated to the full matter distribution down to the galaxy-bias (see section 1.3.3), and galaxies are thus expected to be preferentially found in high-density regions, *i.e.* the galaxies and the over-density field are expected to align.

For that purpose, the Lyon Meudon Extragalactic database (LEDA) catalogue is chosen (Paturel et al., 2003). Only the galaxies whose sky position, redshift (spectroscopic or photometric) are known are selected. The LEDA catalogue, like the Cosmicflows catalogues, does not result from one survey but is rather the compilation of many sources. Even though its selection function is complex, it can be described as roughly flux limited. We limit our comparison to the LEDA galaxies that are in the constrained volume. In this volume, the LEDA catalogue is pseudo-complete[2] for galaxies such that $M < -19$. We limit thus our study to these bright galaxies, of which there are more than 360 000 in the reconstructed region.

These galaxies are subject to the effects of Redshift Space Distortion discussion in section 1.4.7,

[2]Again, the selection is complex, and provide a true completeness of the catalogue is out the scope of this work.

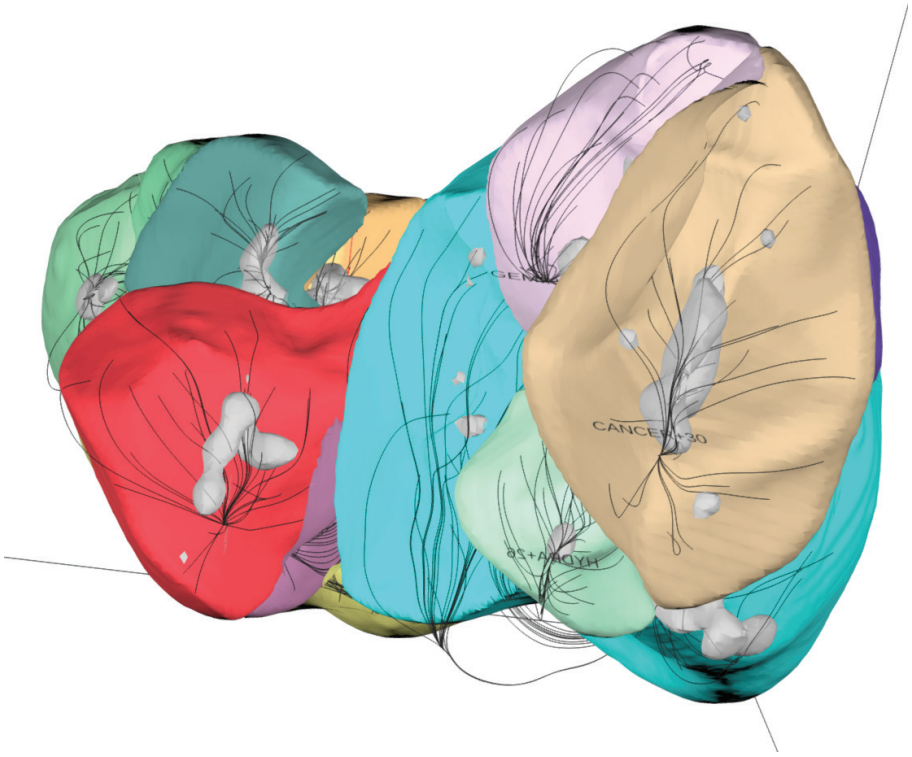

Figure 4.7: Same as fig. 4.5 with another view angle. The BoAs are: Sloan Great Wall (blue), Lepus (red), Funnel (dark green), Perseus-Pisces (orange), Pisces-Cetus (light green, left). Other BoAs do not have a known or named over-density as attractor, and are thus named after the constellation there are in + distance in 10 000 km/s. It is the case in the SDSS region of Cancer+30 (light orange), Hydra+26 (light green) and Gemini+24 (pink).

namely the "Kaiser's pancakes" and "Bull's eye" effects for field galaxies and the Finger of Gods for galaxies in clusters (*e.g.*; Jackson, 1972; Kaiser, 1987; Praton et al., 1997; Thomas et al., 2004).

4.4.3 The over-density field

Figures 4.2 to 4.4 show the SGX = 0, SGY = 0 and SGZ = 0 planes containing the entire volume of the reconstruction, with the LEDA galaxies over-plotted. In the rest of this section, we will simplify the notation by naming these planes respectively the X, Y and Z plane. The footprint of the CF4 catalogue is easily recognizable. The ZoA shadowing the Y plane is visible is both X and Z planes along the Y-axis and in Y plane by the very limited surface reconstructed. In the Z plane, a slight asymmetry along the X-axis below SGY $\approx 160\,\mathrm{Mpc}/h$ is due the contribution of 6dF. In both the X and Z planes, the large volume covered by SDSS is striking, leaving to a strong asymmetry between the SGY > 0 and the SGY < 0 hemispheres of the reconstruction.

Structures at the edge of the data (*e.g.* SGY $> 250\,\mathrm{Mpc}/h$ at the end of the SDSS region, visible both the in X and Z planes) are quite well defined. There is only a small lessening of the sharpness of the contrast with distance, a phenomenon usually observed in reconstructions with the Wiener Filter (see chapter 3). It shows that Hamlet is able to precisely recover the fields in the entirety of the constrained volume. The quality and the density of the constraints severely decreases with the distance: this ability is thus all the more remarkable that it holds up to the edge of the SDSS region, around $cz \approx 30\,000\,\mathrm{km/s}$, a distance never reached by velocity based reconstructions.

Description of the Local Universe and comparison with the LEDA galaxies

The comparison of the predicted linear fractional over-density (δ) field with the observed galaxy distribution should be conducted within the following framework. The reconstructed over-density field is actually the divergence of the reconstructed velocity field, properly normalized and as such does not reflect the

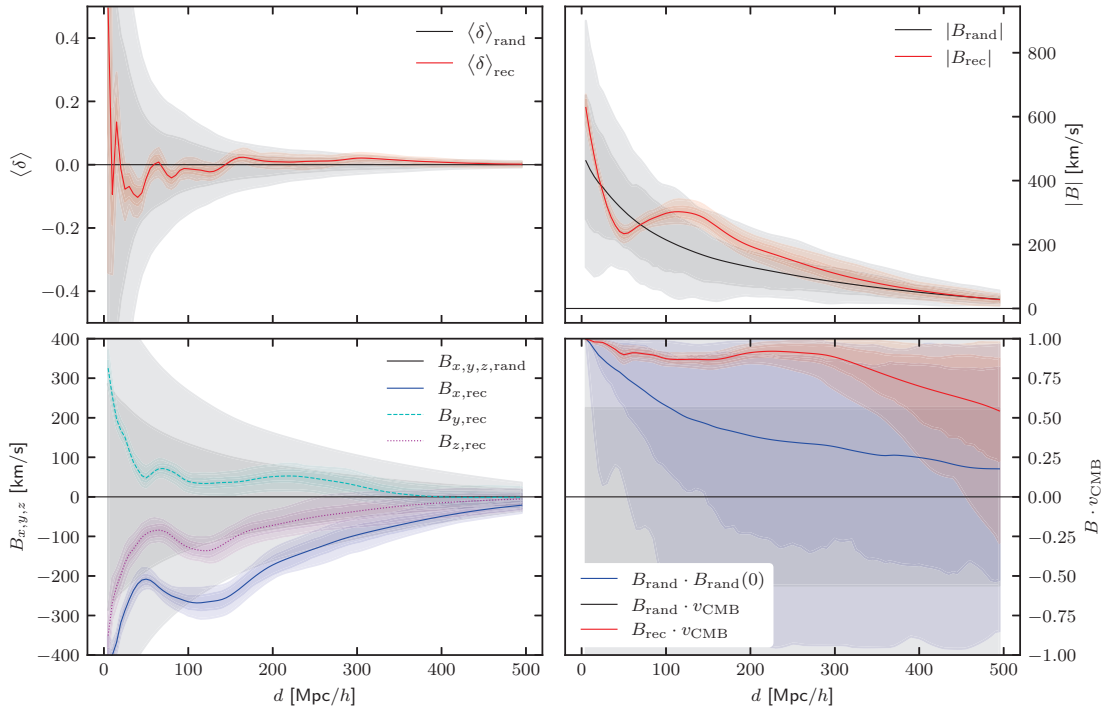

Figure 4.8: Moments of the reconstructed velocity field. **Top left:** Monopole, computed as the mean over-density within a sphere. The mean monopole over the reconstructions is in red full line enclosed withing two shades for the 1- and 2-σ intervals of confidence. The full black line is the mean of the random signal, while the grey shades in the background are the associated 1- and 2-σ intervals of confidence. They are computed over a hundred random realizations of the power spectrum over the same grid ($n = 256$, $L = 1\,\mathrm{Gpc}/h$). **Top right:** Amplitude of the dipole, following the same color convention as for the monopole. **Bottom left:** components of the dipole, with the same convention as for the monpole, except the use of different colors and line styles to differentiate the components. **Bottom right:** cosine of the angle between the dipole and the CMB velocity $v_{\rm CMB} = (-410, 353, -324)\,\mathrm{km/s}$. The red line and shades represent the results from the reconstruction, while the black line and shades stand for the random signal. The blue line and shades show the alignment of the random dipoles with their value at the observers' position (*i.e.* the CMB velocity in this random universe, assuming it stems from the peculiar velocity of the observer w.r.t. the universe).

non-linear evolution of the density field in general, and the virial collapsed clusters of galaxies in particular. Rich clusters are to be associated and compared with local over-density maxima, smoothed on the scale of a few Mpc. Galaxies, are on the other hand, the end product of very complicated non-linear gravitational and other - yet to be understood - galaxy formation processes. Moreover, the LEDA galaxies, used here as a benchmark, are distributed in redshift space. The comparison with the linear over-density field is further hampered by RSD effects, the Finger of God in particular (see section 1.4.7. Given all that the qualitative matching between the galaxies and the over-density field is quite remarkable.

The known nearby clusters, superclusters and voids structures are labeled on figs. 4.2 to 4.4. A short description of these structures follows:

— The nearest clusters, Virgo and Fornax are robustly recovered.

— The Perseus-Pisces (PP) supercluster, the Pavo and Norma clusters and the Sculptor Wall are recovered in the region solely covered by the *other* catalogue.

— In the Z plane, the Centaurus and the 4-cluster region is not well captured by the reconstruction. The causes of this discrepancy are not understood yet.

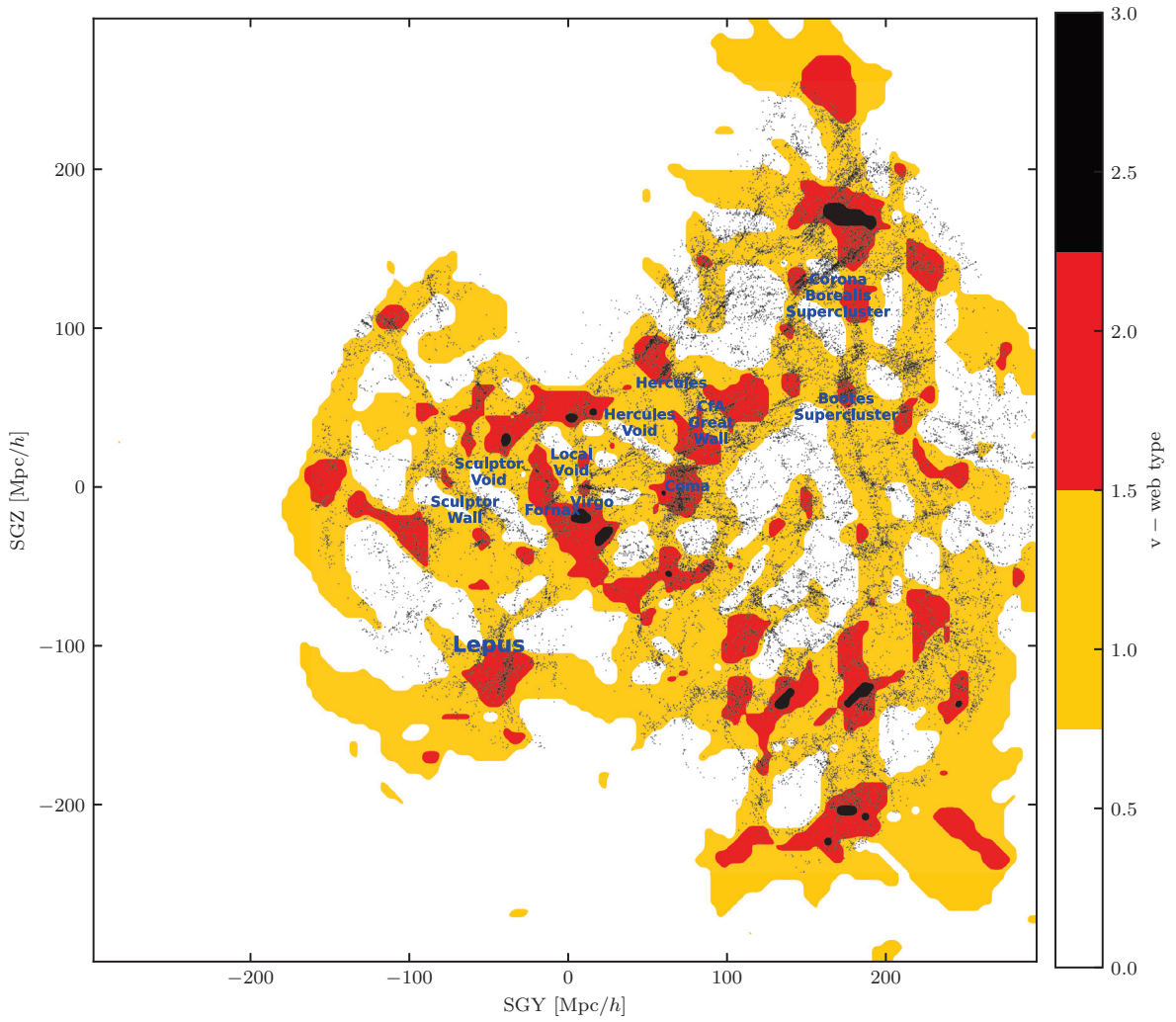

Figure 4.9: Same as fig. 4.2 but the V-Web is shown here. Each cell is given a type associated to the colors: (0) for void in white, (1) for wall in yellow, (2) for a filament in red and (3) for clusters in black. The computation of the V-Web is based on the mean field smoothed with a $5\,\mathrm{Mpc}/h$ kernel and uses $\lambda_{\mathrm{th}} = 0.1$. The V-Web is cropped to the reconstructed volume.

— In the volume constrained by the 6dF component of CF4, a multimodal structure is found around Lepus in the X plane, along with other over-densities mirrored in the galaxy distribution at the edge of the data on the SGY < 0 end. Most of the structures are seen on the Z plane, with Pisces-Cetus at the SGY < 0 edge of the data, the SGX < 0 end of the Sculptor-Wall, the South Pole Wall which cuts the plane and last but not least, the massive Shapley concentration. Shapley is here split in two: a foreground and a core. A long, very dense filamentary-like structures spans from Shapley to another strong over-density also appearing in the galaxy distribution, midway to the Sloan Great Wall.

— Where the SDSS and the *other* catalogues overlap, the reconstruction captures well the complexity of the very rich region comprising Coma, Leo, Hercules supercluster along with the CfA Great Wall. This region is caught in both the X and Z planes. In the same planes, deeper alongside the SGY > 0, in the region of SDSS, an intricate structure of voids over-densities can be seen. In the X plane, the most prominent are the Bootes and the Corona-Borealis superclusters and other unnamed strong although smaller peaks in the SGZ < 0 region. The notorious Sloan Great Wall cuts the Z plane, where Ursa-Major supercluster can also be found. There again, many other clusters are found in accordance with the LEDA galaxy distribution.

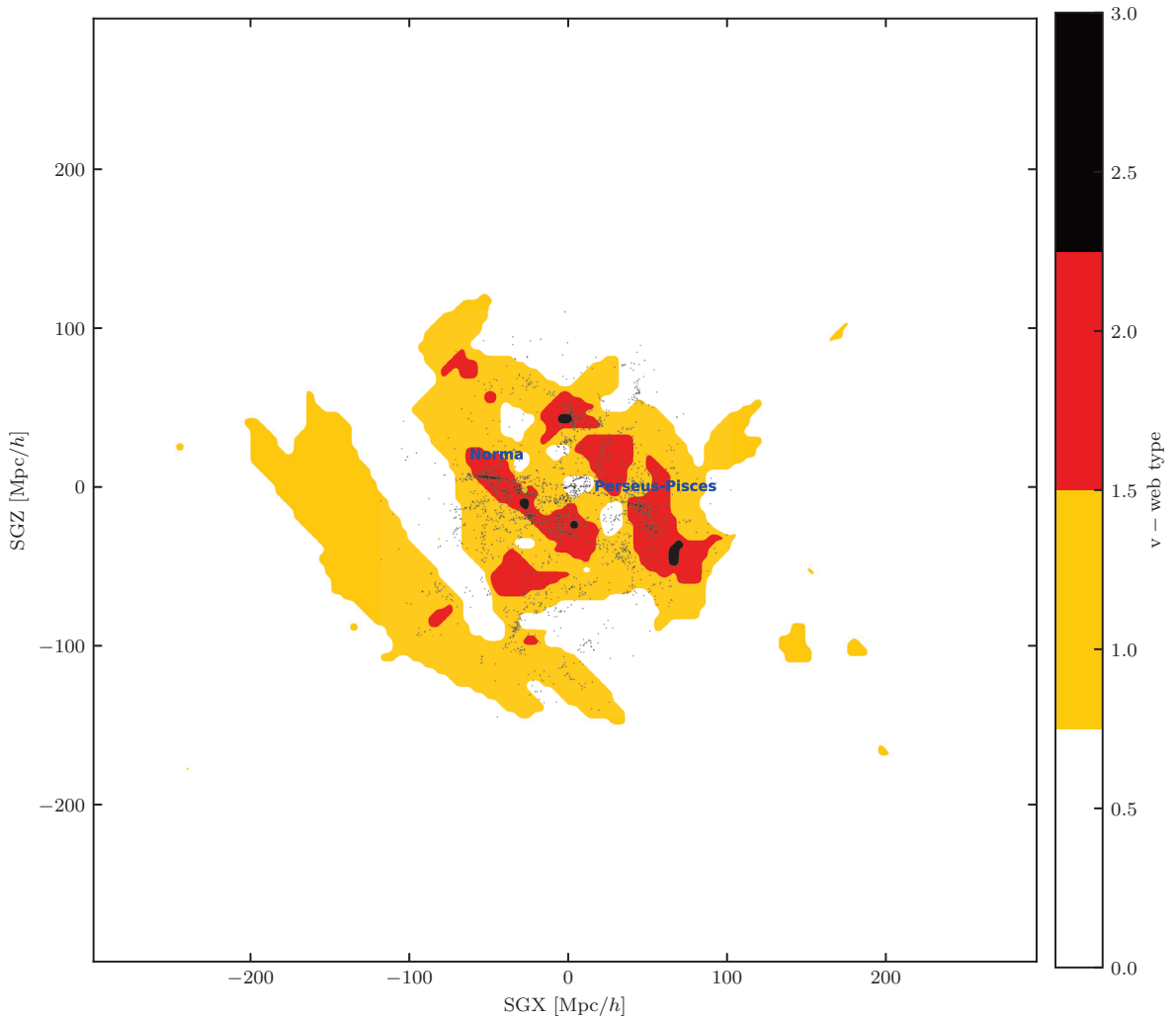

Figure 4.10: Same as fig. 4.9 but for the SGY = 0 plane.

— Voids are important features of the Universe and they are also properly retrieved by HAMLET . In the X plane, the Sculptor void, the Local void as well as Hercules void and Bootes void (showing between the CfA great Wall and Bootes superclusters, the void is not exactly aligned with the X plane) can be recognized. Only the Sculptor void cuts the Z plane.

The detailed cosmographical inspection of the HAMLET reconstruction and its comparison with the observed distribution of galaxies is still in preparation (Valade, Tully, Pomarede, Libeskind and Hoffman, to be submitted).

4.4.4 The velocity field

Basins of Attraction (BoA)

Basins of Attraction (BoA) are large scale gravitational watersheds. A BoA is defined as a volume of space in which, if the expansion were permanently frozen and all galaxies followed their gravitational trajectories, all points would move toward a common *attractor*, like the drops of water of a watershed all converge to the same river. They can be easily derived from the three-dimensional velocity field through the computation of streamlines. A streamline (or flow line) is a method to visualize a vector field, *i.e.* the velocity field in our application. A flow line is the trajectory in space of a particle $\boldsymbol{q}(s)$ alongside (a

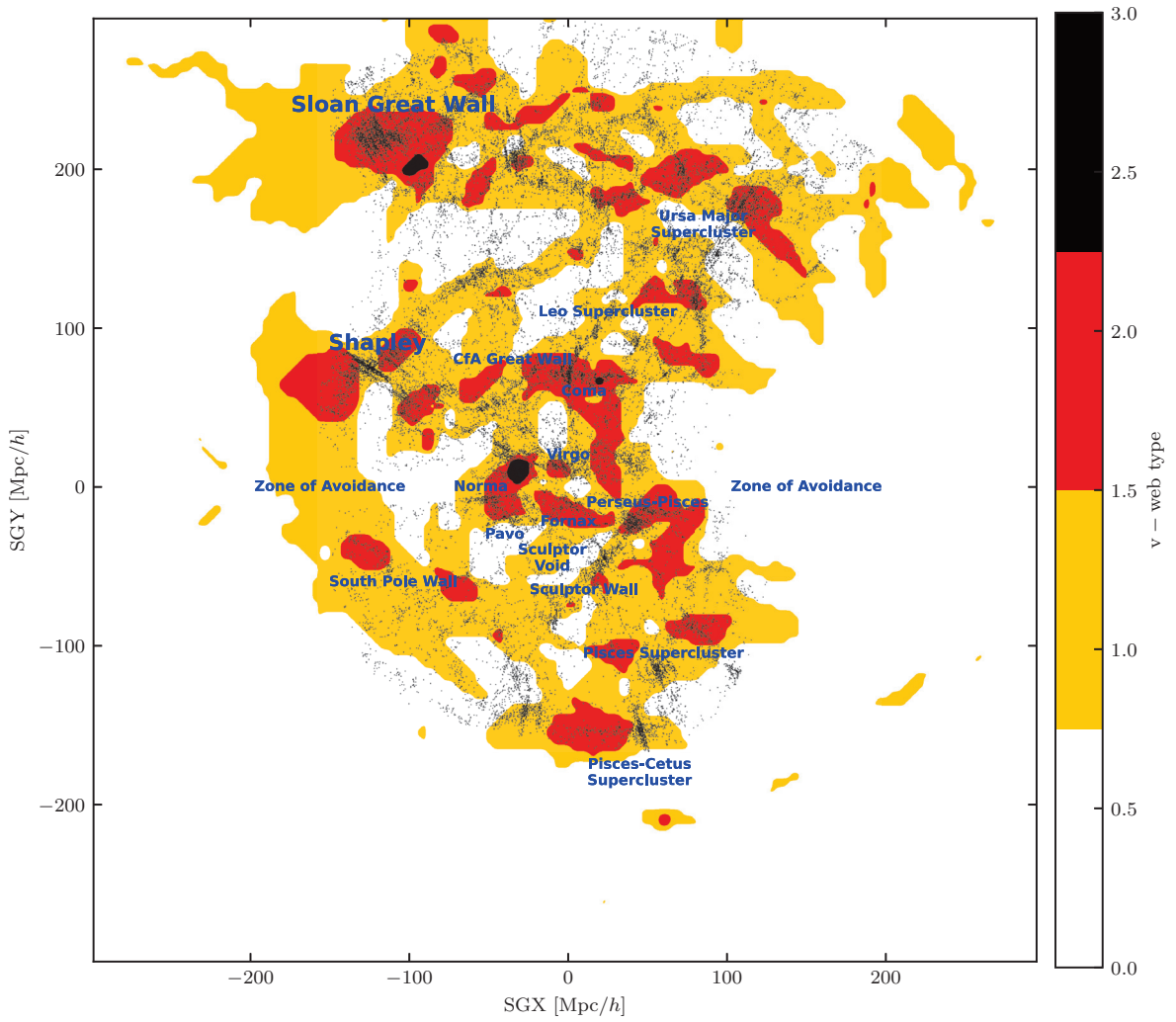

Figure 4.11: Same as fig. 4.9 but for the SGZ = 0 plane.

pseudo) time, starting at position $\boldsymbol{q}_0$ and moving alongside the velocity field:

$$\begin{cases} \mathrm{d}_s \boldsymbol{q} & = \boldsymbol{v}(\boldsymbol{q}) \\ \boldsymbol{q}(s=0) & = \boldsymbol{q}_0. \end{cases} \tag{4.4}$$

A BoA is then the ensemble of cells whose streamlines converge to the same point. The concept of BoA has been introduced by Tully et al. (2014) to propose a new definition of a super-cluster as the attractor of a BoA, rather than an unusually large cluster.

The quality of the retrieved BoAs through our reconstruction methods is yet to be extensively studied. Not knowing the limits of our methodology, they are thus quite hard to be properly analysed. It is however known that smoothing leads to the merging of neighbor BoAs (Dupuy et al., 2020). One of the effects of the large errors and the scarcity of the data points of distant regions is an apparent smoothing of the structures in the mean field. It is thus expected that true BoAs tend to be merged in the reconstruction into larger artificial BoAs in regions where constraints are poor. This artificial merging of BoAs may explain why the BoAs tend to span until the edge of the reconstruction (or even leak outside of it if not cropped). In the absence of good constraints at the edge of the reconstruction, the possibly existing BoAs are merged to the well defined ones whose attractors are well retrieved. Dupuy et al. (2020) also demonstrates that the mass contained in a BoA ranges from roughly $10^{15}\,\mathrm{M}_\odot/h$ to $10^{17}\,\mathrm{M}_\odot/h$. There is however no statement on the span of their volume, but, given that they are all roughly of the same (mean) density, it is reasonable to expect quite a large variety in volume as well.

Figures 4.5 and 4.6 are screenshots of the 3D visualization of the velocity flow lines with the Basins

of Attraction (BoA) marked in different colors.

The BoA of the Sloan Great Wall is by far the most prominent is the SDSS region, and in the whole reconstruction. It covers the vast majority of the volume above the Hercules, CfA Great Wall and Coma complex. This large BoA may possibly the results of the merging of several BoAs. Other BoAs are found in the SDSS region: Corona-Borealis with an identified attractor, Cancer+30, Hydra+26, and Gemini+24 without an identified attractor. Although these BoAs are at the edge of the reconstruction, nothing indicates that they are not to be trusted. Indeed, in the absence of constraints, BoAs are expected to merge, not to split. Although this is purely hypothetical, the attractors of these BoAs could be outside of the constrained volume.

Most of the region of 6dF is enclosed in the BoA of Shapley, which spans from the edge of the reconstruction to the Local Group, absorbing the South Pole Wall, Virgo, Norma and all the very nearby Universe. Lepus, Pisces and Pisces-Cetus have their own BoAs. Another unnamed BoA is found at the SGY < 0 end of the reconstruction, with no identified attractor, which could again possibly be outside of the constrained volume.

In the volume covered primarily by the *other* catalogue, Funnel and Perseus-Pisces have neighbor BoAs going deep into the reconstruction. The boundaries of the later come quite close to the Milky Way and the Local Group ($\approx 10 - 15\,\mathrm{Mpc}/h$). The BoA of Hercules, which includes the CfA Great Wall and Coma, is of great interest. At the border of both Shapley's and the Sloan Great Wall's BoAs, it lies in a very well constrained region, almost at the core of the reconstruction. This BoA might be one of the main findings of this chapter. It would make of Hercules a supercluster under the definition of Tully et al. (2014), *i.e.* that a supercluster is the attractor of a large BoA. Further work is however necessary to confirm this result.

Moments of the velocity field

Another common approach to analyze the velocity field is through its moments. Here, we limit ourselves to the monopole and the dipole, also called bulk flow. Although more complicated definitions exist, the monopole can be computed as the average over-density within concentric spheres centered on the observer, and the dipole as the average velocity of that same volume:

$$\boldsymbol{B} = \big(B_x, B_y, B_z\big), \qquad B = |\boldsymbol{B}|, \qquad B_\alpha(r) = \langle v_\alpha(\boldsymbol{x})\rangle_{|\boldsymbol{x}|<r} \tag{4.5}$$

Both the magnitude and the separated components of the dipole are discussed in the literature. These metrics are thus well studied, and are relatively stable. They are as such a good tool to quantify the quality of our reconstruction. The periodic boundary conditions employed by the HAMLET algorithm affect the reconstructed velocity field. Not only that it misses the tidal contribution of the structures outside of the computational box but also distant structures close to the edge of the box gets correlated by the periodic boundary conditions. It is estimated here that the bulk velocity on scales exceeding $\sim 250\,\mathrm{Mpc}/h$ are strongly affected by the boundary conditions (Hellwing et al., 2018).

To compare the results of our reconstruction to the predictions of ΛCDM , we create 100 random realizations (see section 1.6.5) on which we compute each metric discussed in this section. We show these results in fig. 4.8 as a black line for the mean of the metric, augmented by two grey shades that represent the 68% and 95% quantiles of the distribution centered on the median (which we abusively name 1- and 2-σ interval of confidence).

The monopole is displayed in fig. 4.8 (top left) stays in the 2-σ prediction of ΛCDM and shows no strong deviation from the expected signal. Except the first data point, whose deviation is large compared with the grey 1 and 2 σ corridors, the peaks around $10\,\mathrm{Mpc}/h$ and $80\,\mathrm{Mpc}/h$ respectively due to Virgo and Coma, the Universe within $100\,\mathrm{Mpc}/h$ and even up to $150\,\mathrm{Mpc}/h$ appears under-dense. Two bumps at $160\,\mathrm{Mpc}/h$ and $300\,\mathrm{Mpc}/h$ are of interest, reaching respectively the 1- and the 2-σ limits. Indeed, these coincide respectively with the edge of the 6dF and the SDSS samples. They thus may be resulting from the selection function of the CF4 catalogue rather than features of the Universe.

The magnitude of the dipole seen in fig. 4.8 (top right) lies as well in the 2-σ expectation from ΛCDM. However, a prominent bump spanning from $80\,\mathrm{Mpc}/h$ to $180\,\mathrm{Mpc}/h$, reaching the 2-σ expectation limit around $160\,\mathrm{Mpc}/h$. This feature is also found in the components of the dipole: as shown in fig. 4.8 (bottom left), the x-component leaves the 2-σ expectations shortly before $100\,\mathrm{Mpc}/h$ and re-enters it around $200\,\mathrm{Mpc}/h$. The z-component briefly leaves the 1-σ shade which it re-enters after $150\,\mathrm{Mpc}/h$.

Our result is strengthened by the finding of similar features in the bulk flow in (1) a reconstruction of the CF2 catalogue (Hoffman et al., 2015) (2) in the 6dFGSv catalogue (Magoulas et al., 2016) and (3) in the SDSS-PV (Howlett et al., 2022). Shapley is often cited as the source of this major flow in the Local

Universe, which would indeed be consistent with our analysis of the BoAs of section 4.4.4. However, Watkins et al. (2023) argues that the bulk flow of CF4 is extremely inconsistent with ΛCDM .

The alignment of the dipole with the CMB velocity is studied in fig. 4.8 (bottom right). As the velocities of our reconstruction are the velocities with respect to the CMB, the CMB velocity to consider is not the one with respect to the Sun of eq. (1.44), but with respect to the Milky Way:

$$v_{\rm CMB} = (-410, 353, -324)\,{\rm km/s}. \tag{4.6}$$

Three different metrics are shown. In grey, the mean, 1- and 2-σ zones of confidence of the alignment of the bulk flow of each random universe with the true CMB velocity (*i.e.* the alignment between two random vectors, as they are expected to be completely independant). In blue, the alignment of the bulk flow of each random universe with the bulk flow at the center of of box (taken as a proxy to the CMB velocity of the observer in this random universe). Finally, in red, the alignment of the bulk flow of our reconstruction with the CMB velocity.

The alignment between the reconstructed bulk flow and the CMB velocity remains out of the 1-σ region of the random signal until the edge of the box. This, and the red shades, tighter than their grey equivalent, demonstrate a constraining of this angle on the entire volume of the reconstruction. More than non-random, this alignment is very strong even compared to the self-alignment of bulk flows of random universes: it borders on the 1-σ limit on the most of the reconstructed distance range and exceeds it between 200 Mpc/h and 300 Mpc/h. After 300 Mpc/h, the alignment of the reconstruction decreases rapidly, as expected in the absence of constraints.

The remarkable deviation of the alignment of the dipole and the CMB velocities resonates with the tension in the bulk flow amplitude. A deeper analysis of the moments of the velocity field with more accurate methods of computation, a discussion on the selection function and of course tests on mocks are due.

The V-Web

The large scale distribution of galaxies seems to span a range of structures that form a continuity from voids, walls/sheets, filaments an clusters - the so-called cosmic web. Numerous methods have been suggested for the classification and the construction of the cosmic web (Libeskind et al., 2018).

A method of classification has been developed from the projection of the velocity field on a regular grid (Hoffman et al., 2012). The method is based on the evaluation of the velocity shear tensor on a regular grid and classification of the cosmic web elements by means of the eigenvalues of the shear tensor. Similarly the cosmic web can be constructed by means of the tidal tensor (Forero-Romero et al., 2009).

The detail of the classification algorithm can be found in Hoffman et al. (2012), a very brief summary is given here. First, the shear tensor of the velocity field

$$\Sigma_{ab} = -\frac{1}{H_0}\left(\frac{\partial v_a}{\partial x_b} - \frac{\partial v_b}{\partial x_a}\right), \qquad a, b = x, y, z \tag{4.7}$$

where the tensor is evaluated on a regular Cartesian grid. The $-\frac{1}{H_0}$ normalize is introduced so as to make the tensor dimensionless and the minus sign to associate positive eigenvalues to a contraction along eigenvectors. Eigenvalues are assumed to be ordered by decreasing order ($\lambda_1 > \lambda_2 > \lambda_3$). The classification is done with respect to an assumed positive threshold value ($\lambda_{\rm th}$ such that: voids (type 0): $\lambda_{\rm th} > \lambda_1$; sheets (type 1): $\lambda_1 > \lambda_{\rm th} > \lambda_2$; filaments (type 2): $\lambda_2 > \lambda_{\rm th} > \lambda_3$; and clusters (type 3): $\lambda_3 > \lambda_{\rm th}$;

In this work, we limit our interpretation of the V-Web to its visual inspection. The value of the threshold directly influences the volume occupied by the V-Web: a lower value of $\lambda_{\rm th}$ is less restrictive and leads to a larger number of cells to be marked as a given type. The values of the V-Web on the three super-galactic planes are presented in figs. 4.9 to 4.11, with the LEDA galaxies over-plotted. The value of the threshold is $\lambda_{\rm th} = 0.1$. The first remark that can be made about the V-Web is its relative bulkiness, with respect to both the galactic Cosmic Web (of the LEDA galaxies, for instance) and of the structures of the over-density field.

The present visual analysis of the cosmic web is to be extended to more qualitative studies, relating the distribution of DM halos (in simulations) and galaxies (in simulation and observed ones) to the various components of the cosmic web. This lies outside the scope of the present thesis.

The alignment between the V-Web and the LEDA galaxy might not be as remarkable as for the over-density field but it remains significant. Galaxies are found preferentially in colored regions (*i.e.* part of

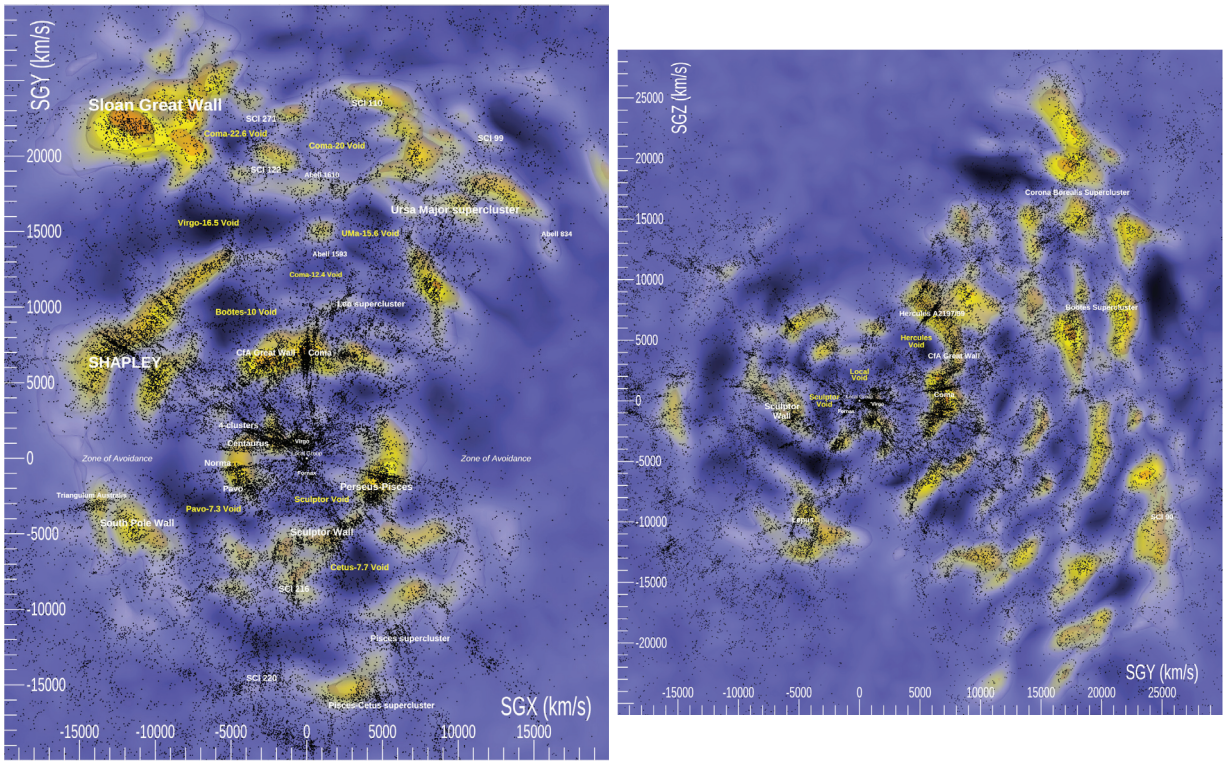

Figure 4.12: Over-density field (from dark violet to orange) with the XSCz galaxies over-plotted. The left panel shows the SGZ = 0 plane (also called super galactic plane), while the right one is the SGX = 0 plane. The names of the major over-densities and voids are labeled respectively in white and yellow. The slices are both 4 000 km/s thick. The over-density field is smoothed with a Gaussian kernel of 5 Mpc/h.

the V-Web), and cluster in red regions, *i.e.* filaments. There is thus a gradation between the galaxy density and the V-Web type. Better, the V-Web seems to follow better radial filaments, *i.e.* filaments parallel to the line of sight, than the over-density field. The black regions (*i.e.* clusters) are however not aligned with the clusters of the LEDA galaxies. Even though the velocity and the density field are not fully coherent with one another (see section 1.8), the separation between the knots of the velocity field and the peak of the LEDA galaxy density do seem a bit distant.

To continue this analysis of the reconstruction of the V-Web, analysis on mocks should be run. Indeed, the V-Web classification has been separately applied to simulations (*e.g.* Forero-Romero et al., 2009; Hoffman et al., 2012; Tempel et al., 2014; Pfeifer et al., 2022, and Hoffman et al., 2023, *in prep.*) and to the Universe (Pomarède et al., 2017). Yet, its stability through our methods of reconstructions has never been studied.

4.4.5 Gallery

In this section we present figs. 4.12 to 4.15, which show alternative visualizations of the reconstruction. The building of these pictures is quite advanced, it associates 3D projections of the over-density field and ray casting. Even though they are based on the same reconstuction, they might thus differ from the "simple" slices presented in section 4.4.3. The description and few comments on the figures are found in the respective captions.

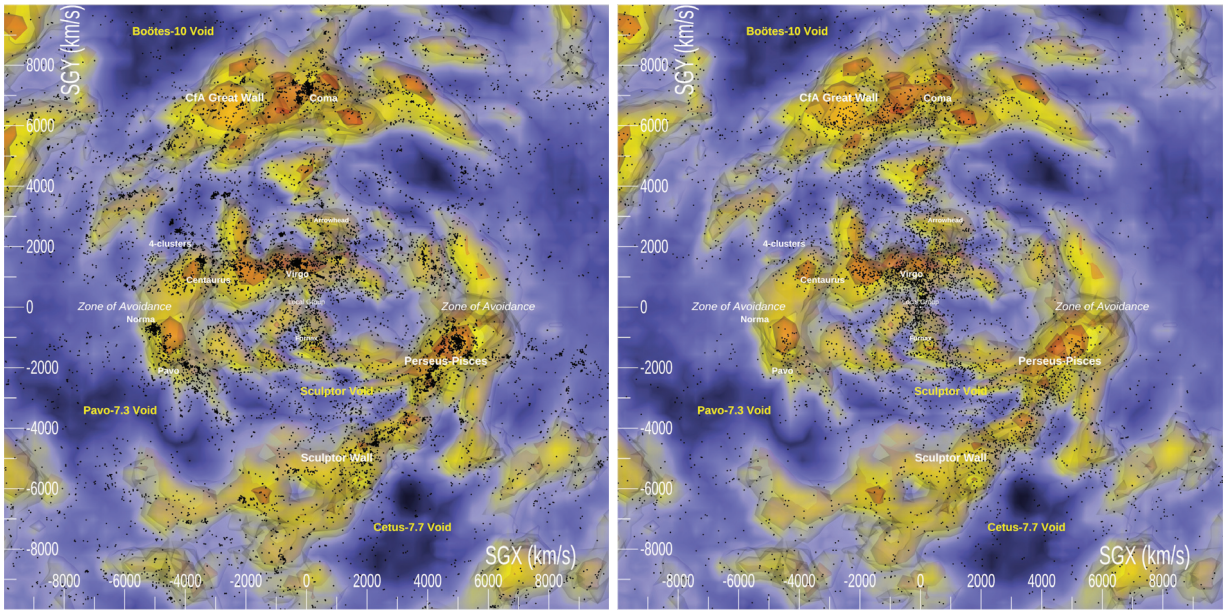

Figure 4.13: Zoom in on the left panel of fig. 4.12. On the left (*resp.* right) panel, the redshift corrected 2MASS catalogue (*resp.* CF4 non-grouped catalogue) is over-plotted. To sharpen the structures, the over-density field shown here is not smoothed.

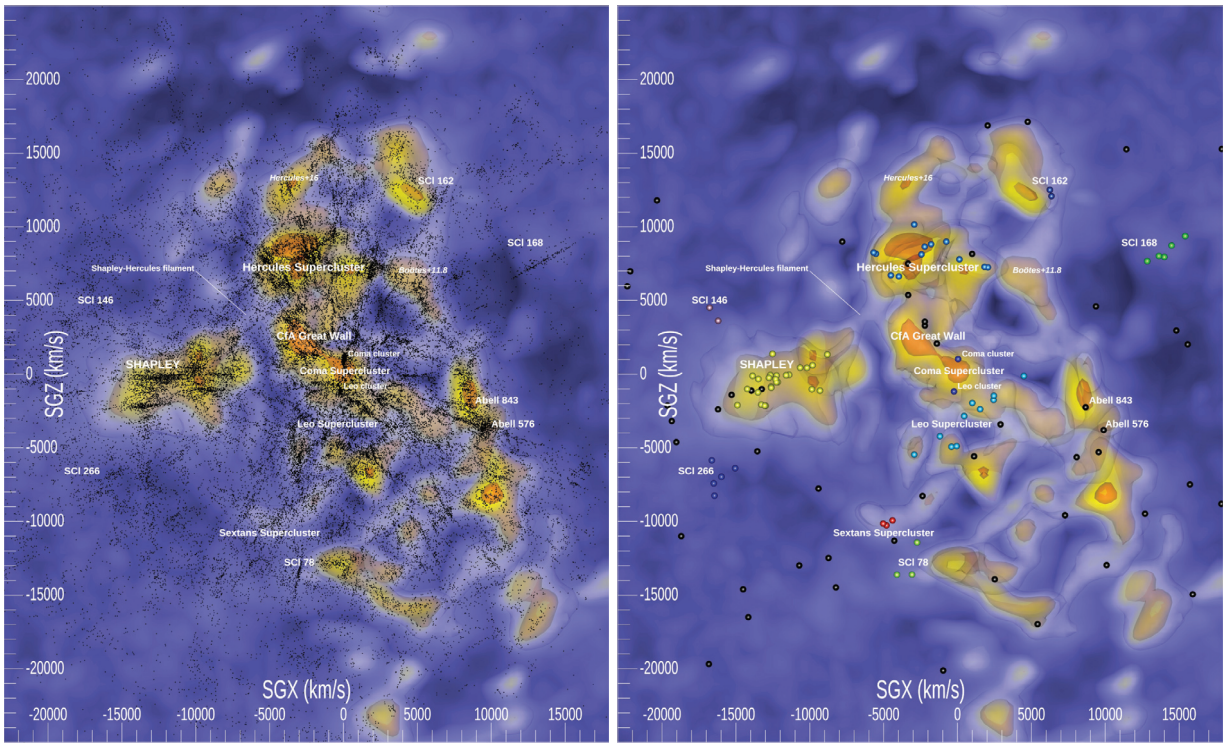

Figure 4.14: Slice of the over-density field between SGY = 5 000 km/s and SGY = 12 000 km/s. On the left (*resp.* right) panel, the XSCz galaxies (*resp.* Abel clusters) are over-plotted. The color of the Abel clusters marks their belonging to identified groups of clusters.

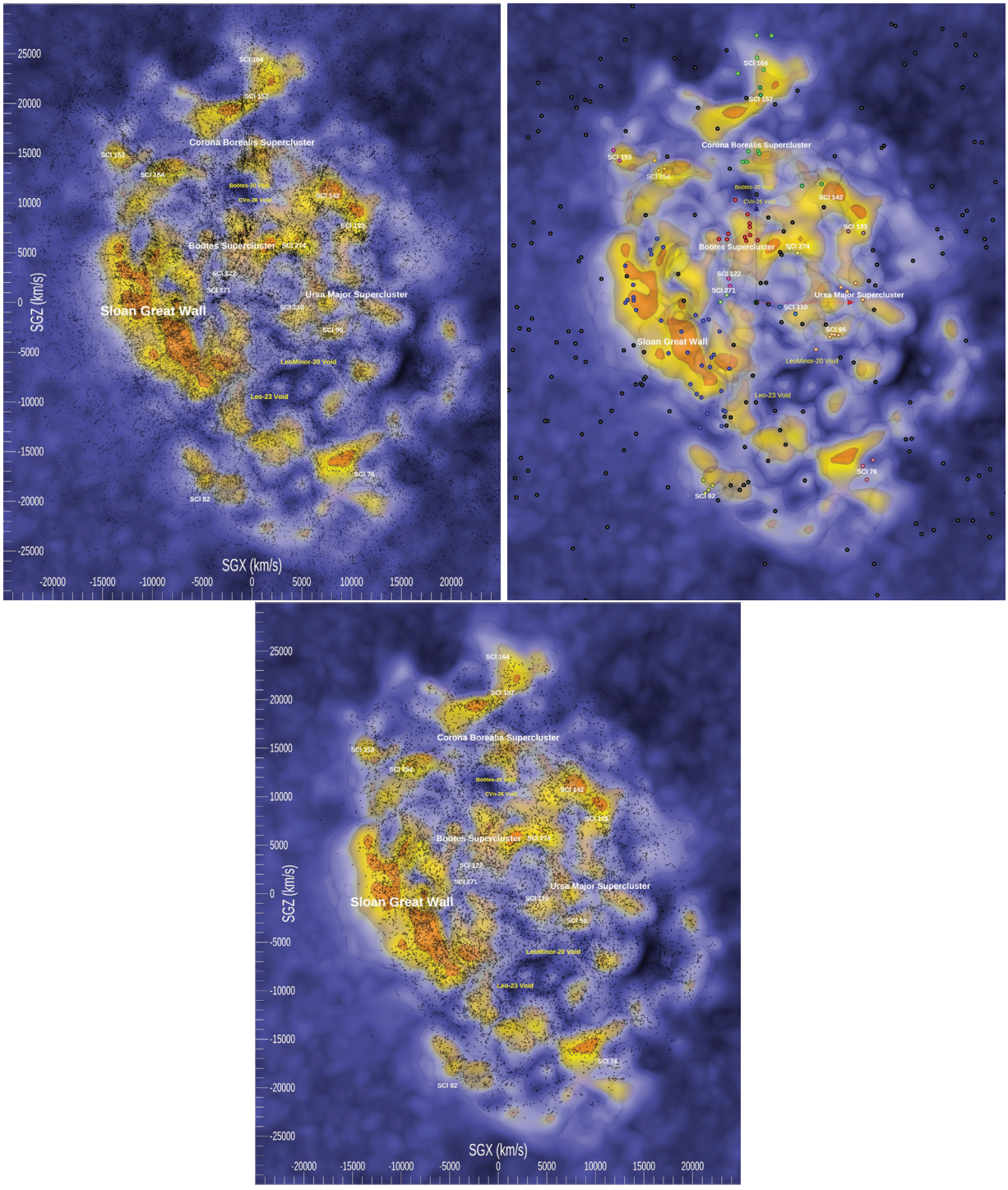

Figure 4.15: Slice of the over-density field between SGY = 16 000 km/s and SGY = 26 000 km/s. On the top left (*resp.* top right) panel, the XSCz galaxies (*resp.* Abel clusters) are over-plotted. The bottom plot shows the CF4 galaxies over-plotted on the over-density field. The color of the Abel clusters marks their belonging to identified groups of clusters.

Chapter 5

Summary and outlook

5.1 Summary

5.1.1 Motivations

Cosmology has been a flourishing field since the beginning of the XX$^{\text{th}}$ century. In the first half of that century, the fundamental mathematics necessary for the modern physical and the statistical description of the cosmos were developed (*e.g.* Einstein, 1916; Friedmann, 1922), while the first the existence of several galaxies and the expansion of the Universe were evidenced (Hubble, 1929; Lemaître, 1933). In the second half of the XX$^{\text{th}}$ century, the increasing quantity and quality of observational data lead to four major discoveries: the cosmological Microwave Background (CMB; Penzias & Wilson, 1965; Smoot et al., 1977, 1992; Fixsen et al., 1996; Komatsu et al., 2011; Planck Collaboration et al., 2016), the Large Scale Structures formed by galaxies (LSS; *e.g.* Peebles, 1980; de Lapparent et al., 1986), the accelerated expansion of the Universe (Λ constant or dark energy; Riess et al., 1998) and the problem of gravitation, resolved in the current model by the presence of dark matter (Zwicky, 1937; Rubin et al., 1980) and cold dark matter (Blumenthal et al., 1984; Davis et al., 1985). The last thirty years have seen the incredibly rapid expansion of computational capabilities offered by fast growing technologies, which opened the door to much more frontal and computational approaches of problems that had before to be simplified to be analytically solved.

Since the discovery of the LSS, redshifts surveys of increasing precision, depth and sky coverage have been carried out (*e.g.* SDSS, York et al. 2000; 2dF, Colless et al. 2001; 6dF, Jones et al. 2009; DESI DESI Collaboration et al. 2016, to name a few), confirming the existence and the structure of the Cosmic Web. In parallel, surveys of galaxies distances have been conducted, which, associated to redshift measurements, allow for the construction of catalogues radial peculiar velocities (*e.g.* Great-Attractor, Lynden-Bell et al. 1988; SAMURAI, Han & Mould 1992; Mark III, Willick et al. 1997; SFI++ Springob et al. 2007; Cosmicflows Tully et al. 2009, 2013; Tully et al. 2016; Tully et al. 2023).

The LSS of the Universe unveils itself by the distribution and the velocities of galaxies. This has driven cosmologists to study the nearby Universe by means of galaxy redshift surveys and by radial velocity surveys. Two families of stratagies for uncovering the LSS have emerged.

The first one aims at retrieving the matter density field from the distribution of the galaxies in redshift space. Reconstruction from redshift surveys - start from the Wiener Filter reconstruction from the IRAS Lahav et al. (1994) and Zaroubi et al. (1995), then a later paper from Erdoğdu et al. (2004) on the reconstruction of the 2dF galaxy survey with the same method. The MCMC approach to the Bayesian reconstruction from redshift survey was pioneered by Kitaura et al. (2009), followed by Jasche & Wandelt (2013) and Wang et al. (2014).

However its central assumption, the relationship between the presence of galaxies and the underlying matter density, is poorly understood (Kaiser, 1984; Bardeen et al., 1986; Mo & White, 1996), which threatens the trustworthiness of their estimations. The second family of reconstructions methods aims at constraining the velocity field of the Universe from the velocities of the galaxies (Bertschinger & Dekel, 1989; Dekel et al., 1999; Zaroubi et al., 1999; Lavaux, 2016; Graziani et al., 2019). On these scales, the only source of motion being gravitational suction, the velocity field is expected to trace the full matter distribution. These methods do not suffer from the poor understanding of the galaxy density versus matter density relationship. They are however curbed by the poor quality and the scarcity of the data stemming from the difficulty to estimate the velocities of distant galaxies. These methods are thus

subject to numerous observational biases (Strauss & Willick, 1995).

Not only have data a very low signal to noise ratio but they are also tainted with strong biases. This has motivated the development of several algorithm to correct the data before the reconstruction of the velocity field by a separated program (Sorce, 2015; Hoffman et al., 2021). However, the biases are many and their modeling complex: all methods differ in their approach and none frontally solves the whole problem.

5.1.2 Hamiltonian Monte Carlo reconstruction of the Local EnvironmenT (HAMLET)

This work consists in the designing, the extensive testing and the application to real data of a method that reconstructs the linear velocity field from measurement of peculiar velocities of galaxies. It follows the lead of Lavaux (2016); Graziani et al. (2019), who respectively developped and applied an algorithm that *conjointly* correct the biases in the data and reconstruct the velocity at once, in a probabilistic, Bayesian, self-consistent manner. The novelty of this work is threefold and constists of (1) the improvement of the Monte Carlo exploration algorithm (2) the implementation of a GPU accelerated code and (3) the better modelization of the data, namely of its selection function. These innovations enable us to reconstruct the last data release of the Cosmicflows catalogues (CF4; Tully et al., 2023) and open the door to future developpements.

Bayesian inference is the mathematical answer to the question "provided that I have made certain observations, and provided that I have a model depending on an ensemble of parameters (or degrees of freedom), what are the values of the model's parameters that can explain the observations best?". The approach taken by forward modeling, detailed under the angle of its application to the reconstruction of the LSS in chapter 2, is twofold.

The first step is the writing of the conditional probability law of the set of parameters given the ensemble of the observations and the model, which is called the posterior distribution. It is the product of a Likelihood function which is the probability that the observations spawn from the set of parameters, and a prior which is the probability of the parameters, independently of the observations. In our work, the posterior distribution models observations of redshifts and distances from a linear velocity field, derived in the context of ΛCDM. The free parameters considered are the Fourier modes of the linear over-density field project on a grid as well as the distances of the constraints.

Although the posterior can be analytically written, its complexity prevents analytical or even simple numerical calculus of summary statistics (mean fields, monopole, dipole, etc). The second aspect of forward modeling is thus the exploration of the posterior probability by a Monte Carlo method. Such a method generates an arbitrary long series of realizations of the posterior probability, on which summary statistics can be computed. One of the innovations of this work is the replacement of the Gibbs sampling used by Graziani et al. (2019) by a cutting edge exploration method: the Hamiltonian Monte Carlo (HMC; Hoffman & Gelman, 2011). The HMC utilizes the Hamiltonian equations to integrate trajectories in the parameters space, allowing the exploration process to make large steps in highly dimensional parameters' spaces and thus partially tackle the so-called "curse of dimensionally" – the rapid escalation of the cost of the exploration of the parameters space as its dimensionality increases. This innovation is not merely technical: it allows for a leap in the applicability of the method to great distances and making the largest reconstructions to date of the universe, from peculiar velocities.

The second key improvement of this work is the implementing of a code, HAMLET , which is designed to run on GPUs. This speed up reduces the running time by several orders of magnitudes with respect to a previously designed algorithm that were limited in their abilities due to inefficient use of computational resources.[1]. More, the physical and the mathematical frameworks are extremely supple and allow for further improvements of both the physical model and the exploration methods. This flexibility is reflected by the extreme modularity of the code.

The last major enhancement of this thesis is the modeling of the redshift cuts in the data following Hinton et al. (2017). It is presented in the later chapter 4, before the application to CF4.

[1]The methodological and numerical differences between this code and the one of Graziani et al. (2019) make the exact computation of acceleration factor non trivial. It seems to range about more than four orders of magnitude. In practice, a result can be achieved in a matter of minutes with the present method, whereas months were necessary with the code of Graziani et al. (2019).

5.1.3 Testing HAMLET

The HAMLET method is applied in this work against data with increasing complexity: in chapters 2 and 3 on mocks and in chapter 4 on real, partially ill-described, data.

Linear mock data

In chapter 2, directly after the presentation of the method, it is tested against mock data that are fully consistent with what the model expects (linear field and same description of the errors). A reference mock catalogue is built with a size of and an error modeling that replicates the Cosmicflows-3 catalogue, the most up-to-date catalogue of distance available at the time of the writing of this chapter. The selection function is isotropic and aims at a fixed number of points per shell of distance (thus a density decreasing as the square of the distance) which corresponds roughly to that of Cosmicflows-3. Other mock catalogues are created by varying the number of constraints (by a factor 1/2 and 2) and the amplitude of the errors (by a factor 1/2 and 1/10) of the reference catalogue, leading to a total number of 9 mock catalogues. The results focus on (1) the mean and standard deviation of the over-density field (2) the mean and standard deviation of the velocity field and (3) the first moments of the velocity field (*i.e.* monopole and dipole).

We show that for these quantities, our code converges well and in an expected fashion with the quantity and the quality of the constraints. This study also allows us to give a first quantitative estimation of the uncertainty on these metrics and predictions for future applications.

Non-linear mock data and comparison with the BGc/WF

In chapter 3, published as Valade et al. (2023), HAMLET is tested against mocks issued from a simulated universe. The size of the catalogue and the modeling of the errors are again linked to that of Cosmicflows-3, extended by a few thousand of points. While the selection function was very basic in the first test, this mock catalogue reproduces with a great fidelity the footprint of Cosmicflows-3 in redshift space, with notably a hemispheric asymmetry and a zone of avoidance. The data effectively spawns up to 160 Mpc/h. The construction of this mock catalogue is done by an advance Monte-Carlo like algorithm that performs the selection the dark matter halos of the simulation (which represent galaxies and groups of galaxies).

Two other methods are applied to the same mock data in order to quantitatively assess and compare the quality of the reconstructions: (1) the Bias Gaussianization correction (BGc) is applied to the data, which are then fed to the Wiener Filter to retrieve the fields (BGc/WF pipeline) and (2) the mocks *without* errors are fed directly to the Wiener Filter (Ex/WF pipeline). While the two first methods (HAMLET and BGc/WF) are methods that can be effectively applied to real data, the last one is used as an hypothetical best-case scenario, in which the position and the velocities of each constraint has been perfectly retrieved. It allows us to quantity the errors and uncertainties resulting from (1) the scarcity of the data and (2) the application of the linear theory to a non-linear universe. The results focus again on mean and standard deviation of both the over-density and velocity fields, as well as the moments of the velocity fields.

We demonstrate that in absence of observational error, the quality of the reconstruction is close to perfect on a very large volume, even with a simplified field theory and a limited set of constraints. In presence of error, the picture is however different. Both HAMLET and the BGc/WF methods yield very similar results within a volume of 80 Mpc/h, outside of which HAMLET displays a higher (and thus a priori better) contrast than the BGc/WF, until the end of the data at 160 Mpc/h. Yet, HAMLET tends to "over-shoot" within 80 Mpc/h and 160 Mpc/h as it produces velocities whose amplitude exceed the ones of the target simulation. Finally, the moments of the velocity are better reconstructed with the BGc/WF than with HAMLET .

The conclusion of this study is that while HAMLET seems to be extract much more information from the same data, the BGc/WF remains a more conservative method. Indeed, HAMLET seems to be subject to some biases that curb its promising potential and that are yet to be understood and corrected.

5.1.4 Application to Cosmicflows-4

After having been extensively tested and compared to another methods in chapters 2 and 3, HAMLET is applied to real data in chapter 4: the peculiar velocity catalogue Cosmicflows-4 (Tully et al., 2023).

In this section, again we push HAMLET further: the size of the Cosmicflows-4 catalogues exceeds the one of the previous release by a factor three, and roughly doubles the volume (by extending the constrained region from 160 Mpc/h to slightly less than 300 Mpc/h in a direction of the sky). More than bringing HAMLET outside of its "computational zone of comfort", the application to real data is also trying for the physical modeling.

Indeed, the data is extremely complex: the Cosmicflows-4 catalogue is not the result of a single survey but rather the merging of several redshifts surveys, extended by different distance catalogues based on different methods. The spatial footprint is thus very asymmetrical, and the errors on each measurement are non trivial. Although these different components of the whole catalogue are inter-calibrated, the possibility of an error in the process or in the sources remains.

The real distribution of the matter in the Universe is unknown. The validity of our reconstruction on this new set of data can thus not be assessed in the same fashion as in chapters 2 and 3. In place of a proper quantitative study, we chose to qualitatively compare our reconstructed fields to the distribution in redshift space of the Lyon Meudon Extragalactic galaxy database (LEDA; Paturel et al., 2003). Even without assuming a specific form for the galaxy bias, a certain correlation between the over-density field and the galaxy distribution can be expected. The redshift being a quite precise estimator of the distance, this constitutes a fair first test for the application of our method on actual data.

The correspondence between the LEDA galaxies and our estimation of the over-density field is satisfactory, the main deviations being imputable to the redshift space distortion. The main features of the Universe are recovered, and for the first time, the matter distribution is directly measured in the region of the Sloan Digital Sky Survey (SDSS). This volume encloses notably the famous Sloan Great Wall, which appears to be the most prominent over-density of our reconstruction, more than the Shapley concentration and the Hercules–CfA-Great-Wall–Coma complex. The over-density field displays a very rich filamentary structure filling the newly mapped volume of SDSS.

The reconstructed velocity field is also closely inspected. An analysis of the Basins of Attraction (BoA) shows that the two major attractors in the reconstructed volume are Shapley and the Sloan Great Wall in the region of SDSS. The Local Group appears to be embedded in the Shapley BoA. The Hercules supercluster is also the attractor of a relatively large, well constrained BoA. The moments of the velocity field are then discussed. While the monopole shows no unexpected behavior and is fully consistent with ΛCDM and the power spectrum, the dipole manifest a somewhat surprising behavior around 160 Mpc/h, where the both the component along the super-galactic X axis and amplitude of the dipole deviate from the 2-σ expectation of ΛCDM . This result in confirmed in the literature (Hoffman et al., 2015; Magoulas et al., 2016; Howlett et al., 2022; Watkins et al., 2023). The alignment of the dipole with the CMB velocity is also remarkably high and constant and may deserve being investigated. Finally, the V-Web is computed and presented.

5.2 Future work

The method and the code presented in this work have a promising potential. Indeed, the method allow the expansion of the physical model to correct existing biases or add new sources of information, while the mathematical algorithm and the high efficiency computing implementation leave to door open to more complex and costly models. In this final section, we briefly review a few outlooks that may be the directions taken in further works. ‘

5.2.1 The need for a CF4 mock catalogue

First of all, the testing of the method done so far was mimicking the properties of Cosmicflows-3 (Tully et al., 2016). However, for the coming years, the cutting edge data in term of size, errors and more generally selection footprint is Cosmicflows-4 (Tully et al., 2023). Moreover, this work leads not only to the development of a novel reconstruction method, but its extensive testing also gave us a deep understanding of its strengths and shortcomings.

New mocks will thus be built based on Cosmicflows-4 and with a proper modeling of the important characteristics that influence the HAMLET reconstructions. To represent well the galaxy bias(es) depending of the components of the Cosmicflows-4 catalogue (SDSS, 6dF and *other*), the dark matter halos used in the previous mocks will be replaced by galaxies. However, given the size of our catalogue, hydrodynamical simulations can not be employed. Thus, a dark matter only simulation with a semi-analytical galaxy modelling will be used.

Among other applications, these mocks will enable us to

— check the stability the basins of attraction and the V-Web through our reconstructions;
— test different hypotheses of the ortho-radial bending of the LSS and more generally quantify the robustness of HAMLET with respect to the galaxy bias and other biases;
— construct and test methods to correct the RSD with the maps of density and velocity;
— construct and test a constrained simulation pipeline with HAMLET (see section 5.2.2).

5.2.2 Quasi-linear maps and initial conditions with HAMLET

Already in preparation, the very next step is to go beyond the linear theory, which is one of the limitations of our method. This will be done in the first place by the modification of eq. (2.16) by a first Largrangian Perturbation evolution model (1-LPT; Zel'dovich, 1970) which gives an estimation of the position of particle in the modern universe $\boldsymbol{x}$ given its initial position $\boldsymbol{q}$ and the initial velocity field $\boldsymbol{u}$

$$\boldsymbol{x} = \boldsymbol{q} + \frac{1}{H_0 f}\boldsymbol{u}(\boldsymbol{q}). \tag{5.1}$$

The detail of the analytical implications of this approximation is discussed in Nusser et al. (1991).

The method and its application will be the subject of a future publication, but the big lines are as follow. The 1-LPT evolved density and velocity fields can be computed on a grid from the initial over-density field on the same grid as follow

1. the initial velocity field on the grid is computed using eq. (1.38) (alternatively eq. (1.85));
2. the nodes of the grid are moved following eq. (5.1);
3. a Cloud In Cell (CiC) algorithm is employed to reconstruct the evolved density field[2];
4. the evolved velocity field is computed by applying eq. (1.38) to the evolved density field.

The replacement of eq. (2.16) by this method directly implicates that the parameters are not the Fourier modes of the evolved linear over-density field but rather the ones of the initial linear over-density field. Two goals are thus achieved at once: (1) a quasi linear map of the evolved Universe is drawn (2) constrained initial conditions for cosmological simulations are created as a "by-product". These two applications should lead hopefully to a better mapping of the Local Universe, and a better constraining of the evolution models through the constrained simulations.

Note that this method is fundamentally different from the Reverse Zeldovich approximation developed in Doumler et al. (2013a); Doumler et al. (2013b,c), where the Zeldovich approximation is applied to the constraints using a linear estimation of the evolved velocity field. What is proposed in this section is to use a linear description of the initial velocity field, apply the Zeldovich approximation to a grid, and compare the resulting evolved fields with the constraints.

Nothing prevents the implementation of a 2-LPT (Buchert & Ehlers, 1993) or higher order even a rough Particle Mesh (Darden et al., 1993) or Particle-Particle Particle-Mesh (Couchman, 1991) models of evolution. Including these higher order methods in the furture is one way the current model could be improved.

5.2.3 A better modeling of the galaxy bias

The modeling of the galaxy bias will be at the core of future works. Moving away from the linear theory as proposed in section 5.2.2 would be a step towards it, as the actual density field (as opposed to the divergence of the velocity field) would be constructed. This estimation would however be relatively rough and diverge in high density regions.

Yet, the problem of the resolution of the grid remains, independently of the quality of the evolution model. Either future works will demonstrate that a good estimation of the density field on a course grid is sufficient for a good correction of the IHM and the galaxy bias, or grid independent methods should be developed.

An alternative solution would be to model the two-points correlation between galaxies, the probability to observe a galaxy at a position $\boldsymbol{x}$ knowing there is agalaxy at $\boldsymbol{y}$. This quantity has been studied extensively in the 1970s and 1980s. It is written in an isotropic universe as (Peebles, 1980)

$$P(\boldsymbol{x}, \boldsymbol{y}) = n_g(\boldsymbol{x}) n_g(\boldsymbol{y})[1 + \xi(r)], \quad r = |\boldsymbol{x} - \boldsymbol{y}|, \tag{5.2}$$

where n_g is the galaxy density. The two-point correlation is well modeled by (Peebles, 1980)

$$\xi(r) = \left(\frac{r}{R_0}\right)^{-\alpha}. \tag{5.3}$$

This probability could be injected in our code, however two main obstacles are on the way. First, the values of R_0 and α depend on the type of constraints (Bahcall & Soneira, 1983). As there are several types of galaxies in the Cosmicflows catalogues and that the data is grouped, a very careful modelization has to be done. Secondly, the number of pairs of constraints increases as $\mathcal{O}(n_{\text{obs}}^2)$. Such an interaction matrix takes too long to compute and too heavy to store in memory. As both short and long distance interactions are important, a mixture of accurate description for the close by pairs and Taylor development for distant pairs should be developed and tested.

Note that this solution may solve the IHM bias, *i.e.* it constraints the distribution of galaxies in space, yet, it does not bind it to the underlying over-density field. Finally, the two-points correlation captures only part of the statistics of the LSS: the three-point, four-point and more generally n-point correlation functions should be taken into account for a full description (Peebles, 1980). These can however not easily written (Fry, 1984) and their computation is out of reach in our framework, as the number of interactions grows as $\mathcal{O}(n_{\text{obs}}^p)$ where p is the order of the correlation function.

Bibliography

Alam S., et al., 2017, MNRAS, 470, 2617

Bahcall N. A., Soneira R. M., 1983, ApJ, 270, 20

Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15

Barnes J., Hut P., 1986, Nature, 324, 446

Bartelmann M., Schneider P., 2001, Phys. Rep., 340, 291

Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109

Bertschinger E., Dekel A., 1989, ApJL, 336, L5

Betancourt M., 2017, arXiv e-prints, p. arXiv:1701.02434

Betoule M., et al., 2014, A&A, 568, A22

Beutler F., et al., 2011, MNRAS, 416, 3017

Birdsall C. K., Fuss D., 1969, Journal of Computational Physics, 3, 494

Blakeslee J. P., Davis M., Tonry J. L., Dressler A., Ajhar E. A., 1999, ApJL, 527, L73

Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, 517

Bolzonella M., Miralles J. M., Pelló R., 2000, A&A, 363, 476

Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603

Bonnor W. B., 1957, MNRAS, 117, 104

Boruah S. S., Lavaux G., Hudson M. J., 2022, MNRAS, 517, 4529

Buchert T., Ehlers J., 1993, MNRAS, 264, 375

Calcino J., Davis T., 2017, J. Cosmology Astropart. Phys., 2017, 038

Campbell L. A., et al., 2014, MNRAS, 443, 1231

Cole S., Lacey C., 1996, MNRAS, 281, 716

Colless M., et al., 2001, MNRAS, 328, 1039

Couchman H. M. P., 1991, ApJL, 368, L23

DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036

Darden T., York D., Pedersen L., 1993, JCP, 98, 10089

Davis M., Peebles P. J. E., 1983, ApJ, 267, 465

Davis T. M., Scrimgeour M. I., 2014, Monthly Notices of the Royal Astronomical Society, 442, 1117

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371

Dekel A., Eldar A., Kolatt T., Yahil A., Willick J. A., Faber S. M., Courteau S., Burstein D., 1999, The Astrophysical Journal, 522, 1

Deupree R. G., Wallace R. K., 1987, ApJ, 317, 724

Djorgovski S., Davis M., 1987, ApJ, 313, 59

Doumler T., 2012, Theses, Université Claude Bernard - Lyon I, https://tel.archives-ouvertes.fr/tel-01127294

Doumler T., Gottlöber S., Hoffman Y., Courtois H., 2013a, MNRAS, 430, 912

Doumler T., Hoffman Y., Courtois H., Gottlöber S., 2013b, Monthly Notices of the Royal Astronomical Society, 430, 888

Doumler T., Courtois H., Gottlöber S., Hoffman Y., 2013c, Monthly Notices of the Royal Astronomical Society, 430, 902

Draine B. T., 2003, ARA&A, 41, 241

Dressler A., 1980, ApJ, 236, 351

Dupuy A., Courtois H. M., Libeskind N. I., Guinet D., 2020, MNRAS, 493, 3513

Einasto M., et al., 2016, A&A, 587, A116

Einstein A., 1914, Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, pp 1030–1085

Einstein A., 1915a, Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, pp 778–786

Einstein A., 1915b, Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, pp 844–847

Einstein A., 1916, Annalen der Physik, 354, 769

Einstein A., de Sitter W., 1932, Proceedings of the National Academy of Science, 18, 213

Eisenstein D. J., Hu W., 1998, ApJ, 496, 605

Erdoğdu P., et al., 2004, MNRAS, 352, 939

Event Horizon Telescope Collaboration et al., 2019, ApJL, 875, L1

Faber S. M., Jackson R. E., 1976, ApJ, 204, 668

Fabian A. C., 2012, ARA&A, 50, 455

Ferrarese L., et al., 2000, The Astrophysical Journal Supplement Series, 128, 431

Fixsen D. J., Cheng E. S., Gales J. M., Mather J. C., Shafer R. A., Wright E. L., 1996, ApJ, 473, 576

Forbes D. A., Kroupa P., 2011, , 28, 77

Forero-Romero J. E., Hoffman Y., Gottlöber S., Klypin A., Yepes G., 2009, MNRAS, 396, 1815

Friedmann A., 1922, Zeitschrift fur Physik, 10, 377

Friedmann A., 1924, Zeitschrift fur Physik, 21, 326

Fry J. N., 1984, ApJ, 279, 499

Ganon G., Hoffman Y., 1993, ApJL, 415, L5

Gorski K., 1988, ApJL, 332, L7

Gott J. Richard I., Jurić M., Schlegel D., Hoyle F., Vogeley M., Tegmark M., Bahcall N., Brinkmann J., 2005, ApJ, 624, 463

Graziani R., 2018, Theses, Université de Lyon, https://tel.archives-ouvertes.fr/tel-02068966

Graziani R., Courtois H. M., Lavaux G., Hoffman Y., Tully R. B., Copin Y., Pomarède D., 2019, Monthly Notices of the Royal Astronomical Society, 488, 5438

Hamana T., Kayo I., Yoshida N., Suto Y., Jing Y. P., 2003, MNRAS, 343, 1312

Hamana T., Kayo I., Yoshida N., Suto Y., Jing Y. P., 2005, MNRAS, 357, 1407

Han M., Mould J. R., 1992, ApJ, 396, 453

Heath D. J., 1977, MNRAS, 179, 351

Hellwing W. A., Bilicki M., Libeskind N. I., 2018, Phys Rev D, 97, 103519

Hinton S. R., Kim A., Davis T. M., 2017, arXiv e-prints, p. arXiv:1706.03856

Hoffman M. D., Gelman A., 2011, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo (arXiv:1111.4246)

Hoffman Y., Ribak E., 1992, The Astrophysical Journal, 384, 448

Hoffman Y., Metuki O., Yepes G., Gottlöber S., Forero-Romero J. E., Libeskind N. I., Knebe A., 2012, MNRAS, 425, 2049

Hoffman Y., Courtois H. M., Tully R. B., 2015, MNRAS, 449, 4494

Hoffman Y., Nusser A., Courtois H. M., Tully R. B., 2016, MNRAS, 461, 4176

Hoffman Y., Nusser A., Valade A., Libeskind N. I., Tully R. B., 2021, MNRAS, 505, 3380

Howlett C., Said K., Lucey J. R., Colless M., Qin F., Lai Y., Tully R. B., Davis T. M., 2022, MNRAS, 515, 953

Hubble E. P., 1926, ApJ, 64, 321

Hubble E., 1929, Proceedings of the National Academy of Sciences, 15, 168

Hubble E., 1934, ApJ, 79, 8

Jackson J. C., 1972, MNRAS, 156, 1P

Jasche J., Lavaux G., 2019, A&A, 625, A64

Jasche J., Wandelt B. D., 2013, MNRAS, 432, 894

Jones D. H., et al., 2009, Monthly Notices of the Royal Astronomical Society, 399, 683

Kaiser N., 1984, ApJL, 284, L9

Kaiser N., 1987, MNRAS, 227, 1

Kant I., 1755, Allgemeine Naturgeschichte und Theorie des Himmels

Khintchine A., 1934, Mathematische Annalen, 109, 604

Kitaura F. S., Jasche J., Li C., Enßlin T. A., Metcalf R. B., Wandelt B. D., Lemson G., White S. D. M., 2009, MNRAS, 400, 183

Klypin A. A., Shandarin S. F., 1983, MNRAS, 204, 891

Knebe A., et al., 2011, MNRAS, 415, 2293

Kolb E. W., Salopek D. S., Turner M. S., 1990, Phys Rev D, 42, 3925

Komatsu E., et al., 2011, ApJS, 192, 18

Kourkchi E., et al., 2020, ApJ, 902, 145

Lahav O., Lilje P. B., Primack J. R., Rees M. J., 1991, MNRAS, 251, 128

Lahav O., Fisher K. B., Hoffman Y., Scharf C. A., Zaroubi S., 1994, ApJL, 423, L93

Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193

Lavaux G., 2016, Monthly Notices of the Royal Astronomical Society, 457, 172

Leavitt H. S., Pickering E. C., 1912, Harvard College Observatory Circular, 173, 1

Lee M. G., Freedman W. L., Madore B. F., 1993, ApJ, 417, 553

Lemaître G., 1927, Annales de la Société Scientifique de Bruxelles, 47, 49

Lemaître G., 1931, MNRAS, 91, 483

Lemaître G., 1933, Annales de la Société Scientifique de Bruxelles, 53, 51

Libeskind N. I., Hoffman Y., Gottlöber S., 2014, MNRAS, 441, 1974

Libeskind N. I., et al., 2018, MNRAS, 473, 1195

Libeskind N. I., et al., 2020, Monthly Notices of the Royal Astronomical Society, 498, 2968

Lynden-Bell D., Faber S. M., Burstein D., Davies R. L., Dressler A., Terlevich R. J., Wegner G., 1988, ApJ, 326, 19

Magoulas C., Springob C., Colless M., Mould J., Lucey J., Erdoğdu P., Jones D. H., 2016, in van de Weygaert R., Shandarin S., Saar E., Einasto J., eds, Vol. 308, The Zeldovich Universe: Genesis and Growth of the Cosmic Web. pp 336–339, doi:10.1017/S1743921316010115

Malmquist K. G., 1922, Meddelanden fran Lunds Astronomiska Observatorium Serie I, 100, 1

McBride C. K., Connolly A. J., Gardner J. P., Scranton R., Scoccimarro R., Berlind A. A., Marín F., Schneider D. P., 2011, ApJ, 739, 85

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, JCP, 21, 1087

Mo H. J., White S. D. M., 1996, MNRAS, 282, 347

Monin A., Yaglom A., 2007, Statistical Fluid Mechanics, Volume 2: Mechanics of Turbulence

Naidu R. P., et al., 2022, ApJL, 940, L14

Neal R., 2011, in , Handbook of Markov Chain Monte Carlo. pp 113–162, doi:10.1201/b10905

Nusser A., 2014, ApJ, 795, 3

Nusser A., Dekel A., Bertschinger E., Blumenthal G. R., 1991, ApJ, 379, 6

Paturel G., Petit C., Prugniel P., Theureau G., Rousseau J., Brouty M., Dubois P., Cambr ésy L., 2003, A&A, 412, 45

Peacock J. A., 2007, Cosmological Physics

Peacock J. A., Dodds S. J., 1994, MNRAS, 267, 1020

Peacock J. A., Smith R. E., 2000, Monthly Notices of the Royal Astronomical Society, 318, 1144

Peebles P. J. E., 1980, The large-scale structure of the universe

Penzias A. A., Wilson R. W., 1965, The Astrophysical Journal, 142, 419

Pfeifer S., Libeskind N. I., Hoffman Y., Hellwing W. A., Bilicki M., Naidoo K., 2022, MNRAS, 514, 470

Planck Collaboration et al., 2016, A&A, 594, A13

Pomarède D., Hoffman Y., Courtois H. M., Tully R. B., 2017, ApJ, 845, 55

Praton E. A., Melott A. L., McKee M. Q., 1997, ApJL, 479, L15

Prideaux-Ghee J., Leclercq F., Lavaux G., Heavens A., Jasche J., 2022, Field-Based Physical Inference From Peculiar Velocity Tracers, doi:10.48550/ARXIV.2204.00023, https://arxiv.org/abs/2204.00023

Riebe K., et al., 2013, Astronomische Nachrichten, 334, 691

Riess A. G., et al., 1998, AJ, 116, 1009

Riess A. G., et al., 2018, ApJ, 855, 136

Robertson H. P., 1933, Reviews of Modern Physics, 5, 62

Robertson H. P., 1935, ApJ, 82, 284

Robertson H. P., 1936a, ApJ, 83, 187

Robertson H. P., 1936b, ApJ, 83, 257

Rubin V. C., Ford W. K. J., Thonnard N., 1980, ApJ, 238, 471

Scrimgeour M. I., et al., 2012, MNRAS, 425, 116

Sheth R. K., Diaferio A., 2001, MNRAS, 322, 901

Skrutskie M. F., et al., 2006, The Astronomical Journal, 131, 1163

Smith R. W., 1982, The expanding universe: astronomy's "Great Debate" 1900 - 1931.

Smoot G. F., Gorenstein M. V., Muller R. A., 1977, PRL, 39, 898

Smoot G. F., et al., 1992, ApJL, 396, L1

Sohn S. T., Besla G., van der Marel R. P., Boylan-Kolchin M., Majewski S. R., Bullock J. S., 2013, ApJ, 768, 139

Sorce J. G., 2015, MNRAS, 450, 2644

Sorce J. G., Tempel E., 2017, Monthly Notices of the Royal Astronomical Society, 469, 2859

Sorce J. G., Courtois H. M., Gottlö ber S., Hoffman Y., Tully R. B., 2014, MNRAS, 437, 3586

Sorce J. G., Hoffman Y., Gottlöber S., 2017, Monthly Notices of the Royal Astronomical Society, 468, 1812

Springel V., 2005, MNRAS, 364, 1105

Springob C. M., Masters K. L., Haynes M. P., Giovanelli R., Marinoni C., 2007, ApJS, 172, 599

Strauss M. A., Willick J. A., 1995, Phys. Rep., 261, 271

Tempel E., Libeskind N. I., Hoffman Y., Liivam ägi L. J., Tamm A., 2014, MNRAS, 437, L11

Thomas B. C., Melott A. L., Feldman H. A., Shandarin S. F., 2004, ApJ, 601, 28

Tonry J. L., 1997, in Livio M., Donahue M., Panagia N., eds, The Extragalactic Distance Scale. p. 297

Tonry J. L., Blakeslee J. P., Ajhar E. A., Dressler A., 2000, ApJ, 530, 625

Tonry J. L., Dressler A., Blakeslee J. P., Ajhar E. A., Fletcher A. B., Luppino G. A., Metzger M. R., Moore C. B., 2001, ApJ, 546, 681

Tsujikawa S., 2013, Classical and Quantum Gravity, 30, 214003

Tully R. B., Fisher J. R., 1977, A&A, 500, 105

Tully R. B., Shaya E. J., Karachentsev I. D., Courtois H. M., Kocevski D. D., Rizzi L., Peel A., 2008, ApJ, 676, 184

Tully R. B., Rizzi L., Shaya E. J., Courtois H. M., Makarov D. I., Jacobs B. A., 2009, AJ, 138, 323

Tully R. B., et al., 2013, AJ, 146, 86

Tully R. B., Courtois H., Hoffman Y., Pomarède D., 2014, Nature, 513, 71

Tully R. B., Courtois H. M., Sorce J. G., 2016, The Astronomical Journal, 152, 50

Tully R. B., Pomarède D., Graziani R., Courtois H. M., Hoffman Y., Shaya E. J., 2019, The Astrophysical Journal, 880, 24

Tully R. B., et al., 2023, ApJ, 944, 94

Valade A., Hoffman Y., Libeskind N. I., Graziani R., 2022, MNRAS, 513, 5148

Valade A., Libeskind N. I., Hoffman Y., Pfeifer S., 2023, MNRAS, 519, 2981

Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2019, Cosmological Simulations of Galaxy Formation, doi:10.48550/ARXIV.1909.07976, https://arxiv.org/abs/1909.07976

Walker A. G., 1937, Proceedings of the London Mathematical Society, 42, 90

Wang H., Mo H. J., Yang X., Jing Y. P., Lin W. P., 2014, ApJ, 794, 94

Wang H., et al., 2016, ApJ, 831, 164

Watkins R., Feldman H. A., 2015, MNRAS, 450, 1868

Watkins R., et al., 2023, arXiv e-prints, p. arXiv:2302.02028

Weingartner J. C., Draine B. T., 2001, ApJ, 548, 296

Wiener N., 1930, Acta Mathematica, 55, 117

Willick J. A., Courteau S., Faber S. M., Burstein D., Dekel A., Strauss M. A., 1997, ApJS, 109, 333

Yepes G., Martínez-Vaquero L. A., Gottlöber S., Hoffman Y., 2009, in Balazs C., Wang F., eds, American Institute of Physics Conference Series Vol. 1178, 5th International Workshop on the Dark Side of the Universe. pp 64–75, doi:10.1063/1.3264558

York D. G., et al., 2000, AJ, 120, 1579

Zaroubi S., Hoffman Y., Fisher K. B., Lahav O., 1995, The Astrophysical Journal, 449, 446

Zaroubi S., Hoffman Y., Dekel A., 1999, ApJ, 520, 413

Zel'dovich Y. B., 1970, A&A, 5, 84

Zeldovich I. B., 1978, in Longair M. S., Einasto J., eds, Vol. 79, Large Scale Structures in the Universe. p. 409

Zwicky F., 1937, ApJ, 86, 217

de Lapparent V., Geller M. J., Huchra J. P., 1986, ApJL, 302, L1

van der Wel A., Bell E. F., Holden B. P., Skibba R. A., Rix H.-W., 2010, ApJ, 714, 1779

Acknowledgments

I would like to thank first and foremost Dr. Noam Libeskind, for his valuable scientific input, for his exceptional support, for his – sometimes deserved – trust, for his neverending patience, and for the great freedom he offered me in my work and in my – sometimes lack of – organization.

I am also very thankful to Dr. Prof. Anne Ealet, who immediatly took over the direction of my PhD in difficult times, and helped me again and again navigate the tumultuous seas of the French administration. I would as well like to thank Dr. Prof. Matthias Steinmetz for overseeing the supervision of my PhD in Potsdam.

I must too warmly thank Dr. Prof. Yehuda Hoffman, who took a crucial leading role very early on in my project, who thaught me so much and who never held back on praises – or criticisms. A thousand thanks to Dr. Romain Graziani who put me on tracks for this work, and with whome I wished I could have collaborated longer. Thank you to Dr. Simon Pfeifer for the hours of interesting discussions on scientific and sometimes less scientific topics. Thank you to Dr. Stefan Gottlöber for his persisting interest in my work, for his very relevant questions, and for his many letters of recommandation. Thank you to Dr. Daniel Pomarède and Dr. Prof. Brent Tully for sharing with me their immense knowledge about the Universe that surrounds us.

I am infinitely thankful to my partner Pauline, not only for her comfort on tough days or for sharing the happier ones with me, but also for the immense amount of time she has invested by taking care of our newborn daughter, so that I could simply carry this thesis to completion. Thank you to Anouk too: I had some of my best ideas when I had nothing else to do than silently think next to you in a dark room while you were – slowly – falling asleep.

Last but not least, I must thank my family for their years of moral and material support, from my birth to the one of this work. Thank you to my father for giving me the taste of science and the curiosity to understand the world out there. Thank you to my mother for trying – although mostly failing – to teach me how to be rigorous in my work and that lazyness is never worth it, amonst many other things.