université
PARIS-SACLAY

Les environnements de la toile cosmique : identification, caractérisation et quantification de l'information cosmologique

Cosmic web environments: identification, characterisation and quantification of cosmological information

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 127, Astronomie et Astrophysique d'Ile-de-France (AAIF)
Spécialité de doctorat: Astronomie et Astrophysique
Unité de recherche: Université Paris-Saclay, CNRS, Institut d'astrophysique spatiale, 91405, Orsay, France.
Référent: : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Paris-Saclay, le 16/09/2021, par

Tony BONNAIRE

Composition du jury:

Jean-Luc Starck Directeur de recherche, AIM, Université Paris-Saclay	Président
Oliver Hahn Professeur, University of Vienna	Rapporteur & Examinateur
Radu S. Stoica Professeur, Université de Lorraine	Rapporteur & Examinateur
Giulio Biroli Professeur, LPENS, Université PSL	Examinateur
Sandrine Codis Chargée de recherche, AIM, Université Paris-Saclay	Examinatrice
Simon D. M. White Professeur émérite, Max Planck Institute for Astrophysics	Examinateur

Direction de la thèse:

Nabila Aghanim Directrice de recherche, IAS, Université Paris-Saclay	Directrice
Aurélien Decelle Maître de conférence, LRI, Université Paris-Saclay	Co-encadrant

Thèse de doctorat

NNT: 2021UPASP068

Acknowledgements

Mes premiers remerciements vont évidemment à Nabila Aghanim et Aurélien Decelle. Bien plus que de simples superviseurs, vous avez non seulement modelé le contexte scientifique de cette thèse, son contenu et sa direction, mais m'avez également offert une liberté d'exploration et d'expression à laquelle je n'avais osé songer jusqu'alors. Au-delà du partage de vos vastes connaissances scientifiques dont j'ai bénéficié durant ces trois années, je vous remercie pour votre enthousiasme, votre bienveillance et votre bonne humeur qui font que cette thèse n'a pas simplement été un *travail*, mais une expérience humaine unique ayant mis à rude épreuve mes tendances misanthropes. J'ai entendu qu'un chercheur était, humainement et scientifiquement, façonné par ses directeurs de thèse. Si tel est le cas, sachez que je suis fier de ma généalogie académique.

Bien sûr, comment survivre à une thèse sans le soutien sans faille et quotidien de collègues et amis ~~avec lesquels se plaindre de nos superviseurs~~. C'est donc tout particulièrement que je remercie les thésards de l'équipe cosmo, Daniela, Thibaut et Marion avec qui j'ai eu la chance de ~~subir~~ partager ces trois années, mais aussi ceux partis trop tôt en post-doc, Adélie, Victor et Louis, ou Raphaël, arrivé trop tard en doctorat, sans lesquels le quotidien se serait trouvé bien moins amusant. Des remerciements spéciaux vont aux "hommes simples", Victor et Thibaut, pour ces échanges sains et cordiaux sur slack qui ont égayé nombreuses de mes journées pendant les sinistres périodes de confinement.

Que serait une thèse sans co-bureaux exceptionnels ? Céline et Marion, je ne pouvais pas rêver meilleur bureau que le 216, à mi-chemin entre le poste de travail et la salle de pause de l'équipe cosmo. Mélangez plaintes et chants, amour et haine, un peu de critique de ceux qui viennent de sortir, ajoutez-y une ~~once~~ tonne de travail et vous obtenez les ingrédients essentiels du "bureau des princesses". Je remercie plus généralement l'équipe de cosmologie de l'IAS pour leur accueil, à commencer par les chercheurs permanents, Marian, Mathieu, Julien, Hervé, Nabila, et désormais Laura pour avoir maintenu une ambiance à la fois chaleureuse, familiale, mais aussi propice au travail et à la reflexion. Je remercie également les post-doctorants de l'équipe, passés et présents, Hideki, Giulio, Nicola, Alex et Joseph pour leur partage.

D'un point de vue plus personnel, je remercie ma famille pour leur soutien tout au long de mes études, et surtout Anaïs pour m'avoir poussé à poursuivre ce souhait et avoir toujours cru en moi. La thèse est la consécration d'un long parcours académique durant lequel deux enseignants en particulier ont joué un rôle central dans mon intérêt et mon accès à la recherche. Pour cela, j'exprime ma sincère gratitude à Charlie Kahloun pour m'avoir communiqué, aux travers de ses cours et de mon expérience en stage, cette passion pour la physique mais également à Pierre-Yves Richard pour ses conseils précieux durant mes années à Supélec, et après. Merci aussi Guillaume, pour ces heures de discussions, ces questions et ces admirations communes qui ont alimenté mon intérêt pour la Science. Tu disais "faire une thèse par procuration". Je suis heureux d'aujourd'hui te décerner le titre de "Docteur par procuration".

J'adresse des remerciements à tous les membres de l'IAS, les scientifiques des autres équipes, le personnel administratif, et des encouragements aux futurs docteur(e)s de l'IAS.

Enfin, je souhaite remercier le jury de thèse pour le temps consacré à la lecture de mes travaux, mais aussi pour leurs commentaires sur le manuscrit et la présentation orale.

Cosmic web environments: identification, characterisation, and quantification of cosmological information

Abstract: The late-time matter distribution depicts a complex pattern commonly called the *cosmic web*. In this picture, the spatial arrangement of matter is that of dense anchors, the nodes, linked together by elongated bridges of matter, the filaments, found at the intersection of thin mildly-dense walls, themselves surrounding large empty voids. This distribution, shaped by gravitational forces since billions of years, carries crucial information on the underlying cosmological model and on the evolution of the large-scale structures. Detecting and studying elements of cosmic web, playing also a key role in the formation and evolution of galaxies, are challenging tasks requiring the elaboration of optimised methods to handle the intrinsic complexity of the pattern made of multi-scale structures of various shapes and densities.

With the aim of identifying and characterising the cosmic web environments, we propose several approaches to analyse spatially structured point-cloud datasets, not restricted to cosmological ones, by means of unsupervised machine learning methods based on mixture models. In particular, we use principles emanating from statistical physics to get a better understanding of the learning dynamics of a clustering algorithm and expose how statistical physics can be used to explore the data distribution and obtain key insights on its structure. In order to identify the filamentary part of the pattern, its most prominent feature, we propose a regularisation of the clustering procedure to iteratively learn a non-linear representation of structured datasets, assuming it was generated by an underlying one-dimensional manifold. The method models this latent structure as a graph embedded as a prior in the Bayesian formulation of the problem to estimate a principal graph passing in the ridges of the matter distribution as traced by galaxies or halos. We show that this formulation is especially well-suited for the description of the filaments since it allows the description of their geometrical properties (lengths, widths, etc.) and associates to each tracer a probability of residing in a given filament. The resulting algorithm is successfully used to detect filaments in state-of-the-art numerical simulations. It also allows us to study the relation between the connectivity of galaxy clusters to the cosmic web and their dynamical and morphological properties. Finally, based on a large suite of N-body simulations, we perform a comprehensive analysis of the cosmological information content based on the two-point statistics derived in the cosmic web environments (nodes, filaments, walls and voids). We show that they can break some degeneracies among key parameters of the model making them a suitable alternative probe to significantly improve the constraints on cosmological parameters obtained by standard analyses.

Keywords: Cosmology: Large-scale structure of Universe, Cosmic web; Methodology: Statistical methods, Pattern analysis, Unsupervised machine learning, Mixture models.

Les environnements de la toile cosmique : identification, caractérisation et quantification de l'information cosmologique

Résumé : La distribution de matière dans l'Univers se présente sous une structure complexe que l'on appelle la *toile cosmique*. Dans cette disposition spatiale, des régions denses, les nœuds de la toile cosmique, sont reliés par des ponts de matière, les filaments, qui se trouvent à l'intersection de structures planaires moyennement denses appelées murs qui définissent eux-mêmes les bords de vastes régions vides. Cette distribution, façonnée par la gravité depuis des milliards d'années, contient de précieuses informations sur le modèle cosmologique sous-jacent mais également sur les conditions initiales de l'Univers et son évolution. La détection et l'étude des éléments de la toile cosmique, qui jouent également un rôle fondamental dans la formation et l'évolution des galaxies, constituent de véritables défis nécessitant la conception d'outils sophistiqués pour traiter la complexité des structures multi-échelles qui la compose.

Avec pour ambition d'identifier et de caractériser les différents environnements, cette thèse propose plusieurs approches pour analyser des jeux de données spatialement organisés au moyen de méthodes d'apprentissage non supervisé fondées sur les modèles de mélanges. En particulier, des principes dérivés de la physique statistique sont utilisés pour mieux appréhender et comprendre la dynamique d'apprentissage d'un algorithme de classification non supervisé. Nous exposons comment utiliser ce parallèle avec la physique statistique afin d'explorer le jeu de données et obtenir des informations sur sa structure. Afin d'identifier la structure filamentaire de la toile cosmique, nous construisons ensuite une version régularisée de la procédure de classification pour apprendre itérativement une représentation du jeu de données, que l'on suppose généré par une structure uni-dimensionnelle sous-jacente. La méthode modélise cette structure latente par un graphe qui est intégré comme un *a priori* dans la formulation Bayésienne du problème menant à l'estimation d'un graphe principal passant au centre de la distribution de matière tracée par les galaxies. Nous montrons que cette formulation est particulièrement adaptée à la description des filaments cosmiques puisqu'elle permet la description de leurs propriétés géométriques (longueurs, épaisseurs, etc.) ainsi que l'association, pour les traceurs (galaxies, halos), d'une probabilité d'appartenir à un filament donné. L'algorithme proposé dans la thèse est appliqué avec succès à des simulations numériques. Ces applications ont notamment permis l'étude des relations entre la connectivité des amas de galaxies dans la toile cosmique et leurs propriétés dynamiques et morphologiques. Enfin, nous réalisons, à partir d'un ensemble de simulations à N-corps, une étude approfondie de l'information cosmologique contenue dans les environnements de la toile cosmique (nœuds, filaments, murs et vides). Il est notamment montré que l'analyse des environnements permet de lever les dégénérescences entre certains des paramètres du modèle faisant de la toile cosmique une sonde alternative permettant d'améliorer significativement les contraintes sur les paramètres cosmologiques vis-à-vis des analyses conventionnelles.

Mots-clefs : Cosmologie: Structures grandes échelles de l'Univers, Toile cosmique; Méthodes: Méthodes statistiques, Reconnaissance de motifs, Apprentissage automatique non supervisé, Modèles de mélange.

Contents

List of Figures

Introduction

"Accepte-la, cette thèse."
A. Doyen

Cosmology in the era of data science

The first mappings of galaxies in the sky [Joeveer et al., 1978; Einasto et al., 1980; de Lapparent et al., 1987] together with first simulations [Zel'dovich, 1970; Doroshkevich & Shandarin, 1978] show that today's matter, may it be dark or baryonic, is not filling the Universe uniformly. Instead, it is spreading over a gigantic well-organised structure shaped by billion years of gravitational forces. This pattern, commonly referred to as the *cosmic web* [Bond et al., 1996], exhibits massive nodes that are linked together by uni-dimensional bridges of matter, the filaments, themselves found at the intersection between thin and mildly-dense planes of matter labelled walls forming the vast shells enclosing underdense volumes almost devoided from galaxies called voids. This complex spatial pattern provides a rich amount of information about the content of the Universe but also contains an imprint on its history, how it formed and evolved through time. One of the aims of modern cosmology is to understand this spatial distribution and extract all the possible information to link it with theoretical principles. To do so, the effort of various communities is being put together that allows: (i) deeper, wider and always more complete observations of different matter tracers [e.g. York et al., 2000; Colless et al., 2001; Driver et al., 2009; Laureijs et al., 2011; Abbott et al., 2016]; (ii) accurate simulations of the Universe and its evolution from large to small scales enabling to model the physics of stars and galaxies to clusters and filament [e.g. Springel et al., 2005; Vogelsberger et al., 2014; Dubois et al., 2014; Nelson et al., 2019; Villaescusa-Navarro et al., 2020]; (iii) the development of state-of-the-art methods and algorithms to analyse the large amount of data (web finders [see the review of Libeskind et al., 2017, and reference therein], statistical descriptors of non-Gaussian fields [e.g. Hahn et al., 2020; Cheng et al., 2020], robust and accurate cosmological constraints etc.). All these analyses then provide paramount information to confront to theory thus validating or invalidating some models on the formation and evolution of the Universe itself and its components at all scales.

As such, data analysis is a pillar of modern observational cosmology since its beginning. The interplay between the two fields is even more emphasised by the recent and upcoming collections of always larger and more complex datasets allowing the measurement of cosmological quantities with an unprecedented accuracy. New galaxy surveys such as Euclid [Laureijs et al., 2011], the Vera Rubin Observatory [Collaboration et al., 2009] or the Dark Energy Survey Instrument [DESI, Levi et al., 2013] and upcoming observations of physical quantities like the cosmic microwave background with the Simons Observatory [Ade et al., 2019] or the 21cm neutral Hydrogen line with the Square Kilometer Array will require the development and application of state-of-the-art methods to analyse and interpret the huge amount of complex data. By providing orders of magnitudes more data than previous generations, this soon-available

flood of data is a double challenge in both designing the right tools to deliver meaningful and precise cosmological analyses but also in their optimisation to obtain reasonable time complexity. Standard statistics, for instance based on the n-point correlations, are heavy to derive for such large datasets and require computational enhancements [see e.g. Philcox & Eisenstein, 2020; Philcox, 2021]. In parallel, the growth of machine learning algorithms of the past decade proposes a unique opportunity to handle such data. Their first applications in astrophysics showed outstanding results in many tasks, ranging from the refinement of redshift estimates in photometric surveys [e.g. Carrasco Kind & Brunner, 2013] to the encoding of high-order information [e.g. Cheng et al., 2020; Allys et al., 2020] and constraining cosmological parameters [e.g. Ribli et al., 2019]. Despite these successes, it is essential that, when applied to physics, machine-learning-based algorithms, and in particular deep-learning ones, are well-controlled and understood. Beyond the fact that machine learning can be particularly opaque to the user in the way it *learns* a representation of a dataset, a special attention must be drawn on the inclusion of errors and uncertainties on the estimated quantities. Understanding the features of importance for the learning, the biases that can be induced when models are trained from predefined datasets and how to exploit correctly the output of such algorithms are key questions to which several communities are jointly trying to answer.

With the motivation of enriching our knowledge on large-scale structures of the Universe for cosmological analyses, we will explore, throughout this thesis, these two linked facets of data science and cosmology. The work presented in the manuscript is in particular addressing two main questions that are:

1. How can the matter distribution at present time be used to identify and characterise cosmic web environments?

2. What is the cosmological information contained in the cosmic web environments?

In the corpus of the manuscript, the context of the first question is made broader than the cosmological one and can be formulated, in a data science point of view, as "How can we efficiently extract meaningful representations of spatially structured point-cloud datasets?" Point-cloud data are indeed ubiquitous in many fields of science and even though cosmological datasets representing the large-scale matter distribution are the main object of study, we include the presented work in its larger context of data science when necessary.

Organisation of the manuscript

To distinguish the different contributions, the manuscript is divided into three parts. The first one is dedicated to the presentation of the cosmological context of the thesis by first focusing on the theory of structure formation in Chapter 1. Building-up on this theoretical introduction, we introduce in Chapter 2 the manifestation of cosmic structures in data and simulations, and we discuss the difficulty of linking observable quantities with theoretical predictions and cosmological models. This chapter also sets up the astrophysical and cosmological interests of identifying the different components of the cosmic web with a particular emphasis on the central role of filaments. Finally, it also discusses the difficulties of carrying an efficient extraction of structures from both simulated and observed data.

The second part of the manuscript exposes the methods developed to tackle the extraction of patterns and features from generic point-cloud datasets. Chapter 3 first analyses a statistical physics formulation of a machine learning procedure to gain a physical insight on the structure of point-cloud data. The associated developments and results led to a published article

(**T. Bonnaire**, A. Decelle & N. Aghanim, *Cascade of phase transitions for multiscale clustering*, Phys. Rev. E 103 012105 (2020) [arXiv:2010.07955]). Chapter 4 introduces a new formulation of the principal graph learning problem presented in **T. Bonnaire**, A. Decelle & N. Aghanim (*Regularisation of Mixture Models for Robust Principal Graph Learning*, under review in IEEE Tran. Pattern Anal. Mach. Intell. [arXiv:2106.09035]). This formulation provides the statistical formalism of the actual method allowing the extraction of a continuous one-dimensional structure from a given dataset. This method, named T-ReX, was also presented in **T. Bonnaire**, N. Aghanim, A. Decelle & M. Douspis (*T-ReX: a graph-based filament detection method*, Astron. Astrophys. 337 A18 (2020) [arXiv:1912.00732]).

In the third and last part of the manuscript, we propose to analyse the cosmic web pattern by first exploiting the previously-mentioned principal graph formulation to represent the filamentary part of the pattern in Chapter 5. In particular, we focus on the structural properties of filaments and we investigate in C. Gouin, **T. Bonnaire** & N. Aghanim (*Shape and connectivity of groups and clusters: Impact of dynamical state and accretion history*, Accepted for publication in Astron. Astrophys. [arXiv:2101.04686]) the impact of the filamentary pattern on the physical properties of astrophysical objects like galaxy clusters. Finally, Chapter 6 presents a comprehensive analysis aiming at theoretically quantifying the statistical information contained in the different cosmic web environments in order to constrain cosmological parameters. This work will be published in a forthcoming article **T. Bonnaire** et al., (in prep).

Part I

Emergence of large-scale structures

Chapter 1.

Structure formation in the Universe

"Ce week-end, je me suis rendu compte que la structure de l'univers est quand même vachement compliquée."
T. Perdereau

This first chapter briefly presents the theoretical context of large-scale structure formation in cosmology. It exposes the ideal setup of linear theory describing the evolution of matter perturbations at the largest scales of the Universe and the mathematical representations of common cosmological observables like the density field. Finally, it introduces the current favoured theoretical model of the Universe, the Lambda cold dark matter cosmological model, review its successes and main parameters. Of course, the chapter only introduces the headlines of these rich topics and do not constitute a thorough theoretical introduction. For a more detailed presentation of the cosmological context, the interested reader is referred to seminal text books such as Peebles [1980] or to articles such as Bernardeau et al. [2002].

1.1 The homogeneous universe

1.1.1 Distances in an expanding universe

As observed locally around us, the Universe appears highly inhomogeneous with small-scales structures made of stars and galaxies. However, when averaged over large scales, its structure is relatively "simple" and nearly homogeneous. Said differently, by smoothing the matter distribution at large enough scales, it appears flat with no strong dominant structures. This statement is partly at the foundation of theoretical cosmology and is expressed as the *cosmo-*

logical principle describing the Universe as statistically homogeneous and isotropic at large scales. This principle was initially introduced as a philosophical concept by arguing that our position in the Universe has nothing special and that there should not be preferred directions for the distribution of matter. The incredible advances in observing the sky, and particularly the tiny anisotropies detected in the cosmic microwave background [CMB, Smoot et al., 1992], however brought evidences supporting this principle.

Under these assumptions, the mathematical framework offered by general relativity, allowing to describe the evolution of the matter, energy and geometry of the Universe, lead to solutions of Einstein's field equations known as the Friedmann-Lemaître-Robertson-Walker metric [FLRW, Friedmann, 1922; Lemaitre, 1931; Robertson, 1935; Walker, 1937]. This metric describes the way to compute distances in a Universe that is statistically isotropic and homogeneous, and leads to

$$\mathrm{d}s^2 = c^2\mathrm{d}t^2 - a(t)^2\mathrm{d}\chi^2, \tag{1.1}$$

where c is the speed of light in vacuum, $\mathrm{d}\chi$ the comoving distance independent of how the Universe expands or contracts, and $a(t)$ is the scale factor[1] normalised such that today, $a(t_0) = a_0 = 1$. In our expanding universe [Hubble, 1929], it is indeed convenient to rely on distances between objects that are independent of the Universe global flow. The proper distance $d(t)$ of a source and its comoving counterpart are linked by the scale factor as $d(t) = a(t)\int \mathrm{d}\chi$. In Eq. (1.1), the squared distance element $\mathrm{d}\chi^2$ can take different forms depending on the geometry and curvature of the Universe. For instance, in Cartesian coordinates and assuming a flat universe, it simply reduces to $\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$. More generally expressed in terms of spherical polar coordinates (noting r the radial distance and Ω the solid angle) and with a curvature term, we have $\mathrm{d}\chi^2 = \mathrm{d}r^2/(1-kr^2) + r^2\mathrm{d}\Omega^2$, with $k = \{-1, 0, 1\}$ representing the sign of the spatial curvature. Considering the light emitted by a given source at time $t_{\rm em}$ following a null geodesic with $\mathrm{d}s^2 = 0$, we have from Eq. (1.1)

$$\chi(t_{\rm em}) = \int_0^{t_{\rm em}} \frac{c\,\mathrm{d}t}{a(t)}. \tag{1.2}$$

Due to the expansion, the observed wavelength $\lambda_{\rm obs}$ appears however larger than the emitted one $\lambda_{\rm em}$, which defines the redshift z as

$$\frac{\lambda_{\rm obs}}{\lambda_{\rm em}} = \frac{1}{a(t_{\rm em})} = 1 + z, \tag{1.3}$$

which also allows us to exhibit the relation between the redshift and the scale factor as $a = (1+z)^{-1}$.

1.1.2 The dynamics of the homogeneous Universe

Now equipped with a means of computing distances by the metric (1.1), the dynamics of the Universe can be described by the Friedmann equations linking its expansion with its energy content. By assuming that the Universe is filled with perfect fluid, the first Friedmann equation reads

$$H^2 = \frac{8\pi G}{3}\rho + \frac{\Lambda c^2}{3} - \frac{kc^2}{a^2}, \tag{1.4}$$

where G is the gravitational constant, Λ is the cosmological constant, ρ is the mass density of considered fluid and $H(t) \coloneqq \dot{a}(t)/a(t)$ is the Hubble parameter characterising the expansion

[1] That we sometimes unrigorously write a for convenience.

rate of the Universe at time t. Assuming energy conservation and taking the time derivative of Eq. (1.4), we obtain the second Friedmann equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda c^2}{3}, \tag{1.5}$$

where p is the pressure. The Universe is in fact assumed to filled with a mixture of perfect fluids made of matter and radiation leading to $\rho = \rho_{\rm m} + \rho_{\rm r}$. Similarly, Λ can be interpreted as a fluid with density $\rho_\Lambda = \Lambda c^2/8\pi G$ and k as a fluid of curvature with density $\rho_{\rm k} = -3kc^2/8\pi G a^2$. Assuming equations of state of the form $\rho = wp$ for fluids, we get $w_{\rm r} = 1/3$ for the radiation component, $w_{\rm m} = 0$ for the pressureless matter, $w_\Lambda = -1$ for the cosmological constant and $w_{\rm k} = -1/3$ for the curvature. Introducing further the critical density defined as $\rho_{\rm crit}(t) = 3H(t)^2/8\pi G$, the fluid densities can be expressed in terms of the critical one, also known as the density parameters $\Omega_i(t) = \rho_i(t)/\rho_{\rm crit}(t)$, where i designates either radiation, matter, Λ or curvature. Note that, in the rest of the thesis, when referring to an energy density Ω without specifying a precise time, we refer to the one at present time, t_0. Under these notations, the first Friedmann equation (1.4) reads $\Omega_{\rm r}(t) + \Omega_{\rm m}(t) + \Omega_\Lambda(t) + \Omega_{\rm k}(t) = 1$ and explicitly expresses the conservation of the matter-energy content in the Universe. By deducing a time evolution of the matter densities for each component using the first law of thermodynamics and coupling it back with the Friedmann equations, it yields

$$H(t) = H_0\sqrt{\frac{\Omega_{\rm r,0}}{a(t)^4} + \frac{\Omega_{\rm m,0}}{a(t)^3} + \frac{\Omega_{\rm k,0}}{a(t)^2} + \Omega_{\Lambda,0}}, \tag{1.6}$$

with $\Omega_{i,0} = \Omega_i(t_0)$ denoting the energy density of component i at the present time t_0 and H_0 the value of the Hubble parameter at present time. This simple paradigm of a homogeneous Universe already highlights the importance of the cosmological density parameters and gives an idea of the interplay between the dynamics of the Universe and its content. By knowing the energy density values at the present time, one gets crucial information on the evolution of the Universe, and on how it expanded through $H(t)$.

1.2 The birth of large-scale structures

The homogeneous model presented in the previous section brought tremendous knowledge on the evolution the Universe but also allows the computation of distances from the redshift of a source or the notion of horizon. Despite these achievements, the assumption of a homogeneous and isotropic Universe is only valid at large scales, above roughly 100 Mpc[2] [Yadav et al., 2005] and, by definition, does not predict the formation of structures at smaller ones. Astronomers, striving to map objects in the sky however depict a Universe made of a multitude of structures at different scales, with stars, galaxies, but also clusters of galaxies at larger scales. At even larger one is drawn an interconnected network of galaxies forming oriented structures in space that we call filaments found at the border of gigantic quasi-empty regions called voids. This non-uniform distribution is shown in the first mappings of galaxies for instance reported in Fig. 1.1. All these hierarchical structures suggest the presence of inhomogeneities in the matter distribution. The current leading theory of structure formation assumes that the seeds of these density perturbations have a quantum origin which led to the small inhomogeneities observed in the initial matter distribution. These small fluctuations of densities evolved

[2]The parsec (pc) is a unit of distance used in astronomy and 1 pc = 3.0857×10^{16} m.

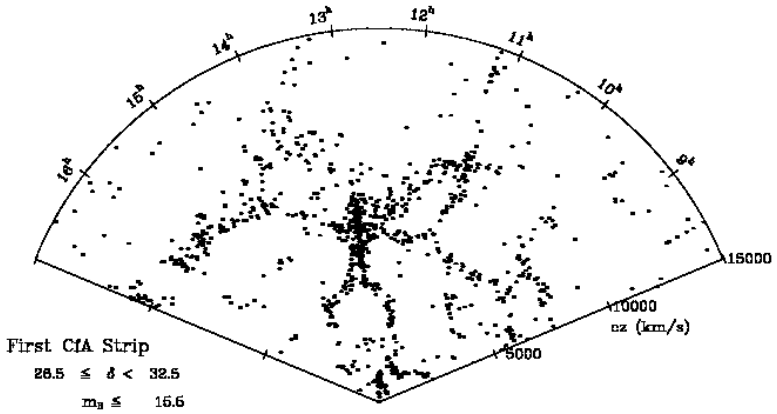

Fig. 1.1. Galaxy distribution from the Center for Astrophysics Redshift Survey [CfA, de Lapparent et al., 1987]. A galaxy cluster at the center is connected by elongated filaments.

through time and gravity to the rich structures that are observed today. In this section, we briefly expose[3] this scenario of gravitational instability.

1.2.1 Linear perturbation theory

In general, the introduction of fluctuations in the matter distribution makes the equations governing the dynamical evolution of the fluid not analytically tractable. Restricting the analysis to small amplitude fluctuations is already instructive and known as the linear perturbation theory leading, as we will see, to their growth. The perturbations can be characterised by the density contrast defined as

$$\delta(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x}) - \bar{\rho}}{\bar{\rho}}, \tag{1.7}$$

where $\boldsymbol{x}$ is a spatial location and $\bar{\rho}$ is the mean density. We focus in this section on perturbations that have small amplitudes $|\delta| \ll 1$. In addition, we restrict the analysis on a Universe made of matter only[4] where the matter component is the dominant one. modeled as a perfect fluid and making use of Newtonian dynamics. Naturally, this single-flow fluid description is expected to break at small-scales where the interactions and crossings between matter particles occur, that is, at scales where cosmic environments like filaments or dark matter halos are formed.

The three equations governing the motion of the fluid are given by the continuity, the Euler and the Poisson equations that we can respectively write, in comoving coordinates $\boldsymbol{x}$ and linearised considering at the same time small perturbations in the density and small velocities with respect to the Hubble parameter $H(t)$,

$$\frac{\partial \delta(\boldsymbol{x},t)}{\partial t} + \frac{1}{a}\nabla_{\boldsymbol{x}} \cdot \boldsymbol{v}(\boldsymbol{x},t) = 0, \tag{1.8}$$

$$\frac{\partial \boldsymbol{v}(\boldsymbol{x},t)}{\partial t} + \frac{\dot{a}}{a}\boldsymbol{v}(\boldsymbol{x},t) = -\frac{1}{a}\nabla\Phi(\boldsymbol{x},t), \tag{1.9}$$

$$\Delta_{\boldsymbol{x}}\Phi(\boldsymbol{x},t) = 4\pi\bar{\rho}G\delta(\boldsymbol{x},t), \tag{1.10}$$

[3]For a complete review of the topic, we refer to Bernardeau et al. [2002].

[4]Which is a valid assumption for a wide range of timescales from the epoch of recombination (3.8×10^5 years after the Big Bang, $z \sim 1100$), to $z \sim 2$

where $\boldsymbol{v}$ is the fluid peculiar velocity (independent from the expansion of the Universe) and Φ is the gravitational potential. The first equation is the conservation of matter, the second one is the Euler equation of motion stating that changes in the velocity flow are caused by the gravitational acceleration and the last one is the Poisson equation linking the gravitational potential Φ to the overdensity δ through the Laplacian Δ. The equations leads to the linear homogeneous second order differential equation describing the time evolution of the density perturbation δ,

$$\frac{\partial^2\delta(\boldsymbol{x},t)}{\partial t^2} + 2H\frac{\partial\delta(\boldsymbol{x},t)}{\partial t} - \frac{3}{2}H^2\Omega_{\rm m}(t)\delta(\boldsymbol{x},t) = 0, \tag{1.11}$$

in which we note the absence of spatial derivatives with respect to $\boldsymbol{x}$. The solution can hence be expressed as the linear combination of two independent solutions, one leading to a growing and the other a decaying perturbation, $\delta(\boldsymbol{x},t) = D_+(t)\delta_+(\boldsymbol{x},0) + D_-\delta_-(\boldsymbol{x},0)$ in which we decoupled the spatial and time evolution of the perturbation. At the considered linear scales, where interactions between matter particles can be neglected, the matter distribution at time t can thus be described by multiplying the initial distribution by the growth factor $D_\pm(t)$.

1.2.2 Zel'dovich formalism

We have seen that an initial small perturbation in the matter distribution can either grow or vanish at linear scales. This already powerful description of the matter distribution at any time is useful for theoretical predictions but, for scales of few tenth of Mpc, we need to resort to the non-linear versions of the continuity and Euler equations (1.8) and (1.9). Zel'dovich [1970] proposes to describe the evolution of perturbations using a Lagrangian description of the fluid. This leads, in particular, to re-write Eq. (1.11) in terms of a displacement field encoding the information of how a particle moved from its initial position. Focusing only the growing mode, Zel'dovich finds that the perturbation grows such that

$$1 + \delta(\boldsymbol{x},t) = \frac{1}{\left[1-\lambda_1 D_+(t)\right]\left[1-\lambda_2 D_+(t)\right]\left[1-\lambda_3 D_+(t)\right]}, \tag{1.12}$$

where $\lambda_1 > \lambda_2 > \lambda_3$ are the eigenvalues of a deformation tensor of the displacement field. In contrast to the linear evolution of perturbations from Sect. 1.2.1, the overdensity in the Zel'dovich approximation is evolving depending on the local properties of the matter distribution. It hence allows to describe the evolution of a perturbation in the mildly non-linear regime and predict the collapse of matter forming the locally anisotropic structures that are observed today and that form the cosmic web. If all eigenvalues are negatives in Eq. (1.12), then the density field is deformed to create a locally underdense region called a *void*, while if λ_1 is positive, then the collapse occurs in a preferential direction forming a locally 2D structure named a *wall*. Following this reasoning, if λ_1 and λ_2 are positive, we end up with a two-dimensional contraction creating tubular-like structures called *filaments* while a quasi-isotropic collapse ($\lambda_1 \simeq \lambda_2 \simeq \lambda_3$ and all positive) forms the *nodes* of the cosmic web.

This Lagrangian description is very powerful to understand how initial seeds of inhomogeneities in the matter distribution give birth to the diversity of structures that are observed in data and simulations (that we will more precisely discuss in the Chapter 2). The collapse takes place at different scales, along different directions and at different times as cadenced by the amplitude of the eigenvalues. Even though very powerful, this first order Lagrangian approximation also assumes that particles are not interacting which breaks at small scales and cannot represent accurately the formation of bounded structures. Several ways to improve the model have been proposed, such as the introduction of a viscosity term to the Euler equation (1.9) which led to the adhesion model [Kofman & Shandarin, 1988; Kofman et al., 1992].

Alternatively, the study of non-linear scales can be performed through numerical simulations that we discuss in the forthcoming Sect. 2.1.

1.3 Statistical descriptions of the matter distribution

One way to treat mathematically observables like the matter distribution in cosmology is to consider the fluctuations $\delta(\boldsymbol{x})$ as a random variable and the observed overdensity field as a realisation of a spatial process that is called a random field.

1.3.1 Discrete random fields

Putting ourselves in a discrete setup where the space can be regularly gridded such that a value of the studied quantity δ can be measured at a location $\boldsymbol{x} \in \mathbb{R}^D$, a random field is defined as the collection of random variables $\{\delta(\boldsymbol{x}_i)\}_{i\in\{1,\dots,n\}}$ and associates to a given realisation a probability to occur. Taking the simple example of an image, one can consider that an $\mathbb{R}$-valued random variable is associated to each pixel, making the entire image the set of random variables constituting one realisation of the random field. In the finite case, the random field is described by the joint probability density function (pdf) of the n discrete random variables $p(\delta(\boldsymbol{x}_1), \dots, \delta(\boldsymbol{x}_1))$.

The cosmological principle, stating that the distribution of matter in the Universe is, at large scales, homogeneous and isotropic (see Sect. 1.1), has a direct implication on the mathematical properties of the considered fields. The first is statistically expressed as the *stationarity* of the associated random process yielding that the joint pdf is invariant under translations (i.e. not a function of the spatial index set). The isotropy in turn induces the field to be invariant under rotation relieving the statistics from any preferential direction. Such invariant fields, widely used in cosmology are also a key element in many mathematical formulations in other fields of physics and applied mathematics. When equipped with additional local properties, they are for instance called Markovian and are at the basis of many developments in image processing such as texture recognition, classification and synthesis [see e.g. Efros & Leung, 1999; Varma & Zisserman, 2009].

In statistics, a common one-point summary of a probability distribution is given by the ensemble average over many realisations. One thing that makes cosmology a special science is that the object of study, the Universe, is the only realisation[5] we have access to. By invoking ergodicity, cosmologists however are able to assimilate ensemble and volume averages to extend the statistical properties of the uniquely observed Universe. As an example, the mean matter density in the Universe that we denoted $\bar{\rho}$ in Sect. 1.2.1 can be seen as the average over a sufficiently large volume of the observed density, without requiring other samples of the Universe.

1.3.2 Correlation functions and poly-spectra

The natural way to describe centred fields with $\langle\delta\rangle = 0$ is through the next non-vanishing moment, the covariance function, also called the two-point correlation function, defined at a given time t as

$$\xi(r,t) \coloneqq \langle\delta(\boldsymbol{x},t)\delta(\boldsymbol{x}+\boldsymbol{r},t)\rangle, \tag{1.13}$$

[5] Although simulations can help in computing ensemble averages by providing many realisations of a same universe, as we will see in the Chapter 6.

where $r = \|\boldsymbol{r}\|$. The dependency of this function on the norm of the separation only comes from the assumed statistical isotropy and homogeneity of the Universe.

By the Wiener-Khinchin theorem [Wiener, 1930; Khintchine, 1934], the Fourier transform of the two-point correlation function defines the power spectrum, as

$$P(k,t) = \int \xi(r,t) \exp\left(i\boldsymbol{k}\cdot\boldsymbol{r}\right) \mathrm{d}^3\boldsymbol{r}, \tag{1.14}$$

also defined as the covariance of Fourier modes, which can be expressed, for a statistically homogeneous and isotropic field, as

$$P(k,t)\delta_{\mathrm{D}}(\boldsymbol{k}_1+\boldsymbol{k}_2) = \frac{1}{(2\pi)^3}\langle\tilde{\delta}(\boldsymbol{k}_1,t)\tilde{\delta}(\boldsymbol{k}_2,t)\rangle, \tag{1.15}$$

where $\tilde{\delta}(\boldsymbol{k},t)$ denotes the Fourier transform of the field $\delta(\boldsymbol{x},t)$ and δ_{D} the Dirac delta distribution. In this latter expression, the assumed statistical properties of the overdensity field δ are expressed in the dependence on the norm k only in $P(k)$ for the isotropy and in the Dirac delta for the translation invariance.

One case of particular importance is the Gaussian random field [for a review, see Adler & Taylor, 2007] which has the particularity of being fully defined by its two first moments that are the average and the correlation function. Mathematically, a random field is said Gaussian if the joint probability of any subset of field points $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ follows a multivariate Gaussian distribution

$$p(\boldsymbol{\delta}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|}\exp\left(-\frac{1}{2}\left[\boldsymbol{\delta}-\boldsymbol{\mu}\right]^{\intercal}\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{\delta}-\boldsymbol{\mu}\right]\right), \tag{1.16}$$

with $\boldsymbol{\delta} \coloneqq \{\delta(\boldsymbol{x}_i)\}_{i=1}^n$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the sets of the n averages and covariances respectively. For centred fields[6] like the matter overdensity δ, the Gaussian random field is characterised by the sole knowledge of the covariance function. Indeed, by Wick's theorem, all the higher-order moments can be expressed as products of two-point functions summed over all possible pairings. Therefore, all the information is indeed contained in the covariance of the field $\boldsymbol{\Sigma}$. In cosmology, the Gaussian random field, beyond its appealing mathematical tractability, is also representing to great accuracy the density perturbations arising after the cosmic inflation as observed by the CMB [Planck Collaboration VII et al., 2020; Planck Collaboration IX et al., 2020]. The corresponding power spectrum of these primordial fluctuations, noted $P_0(k)$ is known to be scale invariant, usually expressed as $P(k,0) = A_{\mathrm{s}}(k/k_*)^{n_{\mathrm{s}}-1}$, with k_* a pivot scale. In the context of linear theory exposed in Sect. 1.2.1, the matter power spectrum of the density fluctuations can be simply expressed as the product of $P(\boldsymbol{k},0)$ and the growth factor $D_+(t)$ responsible for the growth of perturbations. This is particularly powerful since we are hence able, at least at large enough scales, to describe the matter power spectrum based only on the knowledge of the initial one. The gravitational evolution of the instabilities however leads to a distribution that is not well-represented by a Gaussian anymore with a pdf that is mostly skewed towards higher density contrast and better represented by a log-normal distribution [Peacock, 1998]. In such cases, there is a leak of information into higher-order moments that can be expressed similarly to Eq. (1.13). For instance, the three-point correlation function can be written

$$\zeta(r_{12},r_{13},r_{23},t) = \langle\delta(\boldsymbol{x}_1,t)\delta(\boldsymbol{x}_2,t)\delta(\boldsymbol{x}_3,t)\rangle, \tag{1.17}$$

[6] Even if the field is not centred, one can always study the random variable defined subtracting the mean which is now centred.

where r_{ij} is the norm of the vector $\boldsymbol{x}_j - \boldsymbol{x}_i$. The three-point correlation function can equivalently be expressed in Fourier space and is known as the bispectrum. Such summary statistics are at the basis of statistical analyses of large-scale structure datasets. Using higher order estimators of course bring more information about the underlying cosmology but also comes with a higher computational cost that innovative works try to circumvent [see e.g. Philcox, 2021].

1.4 The ΛCDM model

The current favoured theory for modelling the Universe and its evolution is called the Lambda Cold Dark Matter (ΛCDM) model. In this section, we review its main parameters, its main successes explaining why it stands as the "standard" model and introduce in a general manner the way cosmological parameters are constrained.

1.4.1 Presentation of the model

In its simplest form, the ΛCDM rely on a set of six cosmological parameters among which two are related to the physical density of baryons and dark matter, obtained by multiplying the density parameters and the reduced Hubble parameter defined as $h := H_0/100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, namely $\Omega_\mathrm{b}h^2$ and $\Omega_\mathrm{m}h^2$. Two parameters are then describing the amplitude and tilt of the power spectrum of the primordial matter fluctuations, A_s and n_s. Finally, one parameter accounts for the geometry of the Universe, and the last one is linked with the reionisation era of the Universe at which the first stars are formed. This minimalist set of six cosmological parameters together with theoretical equations governing the dynamics of the Universe assuming the cosmological principle partly exposed in Sect. 1.1 is the most simple one agreeing remarkably well with observations that made its supremacy. This is the reason why the ΛCDM model is also sometimes referred to as the *concordance model of cosmology*. Among all the observational evidences of the ΛCDM paradigm, perhaps the most important is the direct detection of the CMB [Smoot et al., 1992; Komatsu et al., 2011; Planck Collaboration I et al., 2016] constituting the first emitted light in the Universe that probes the existence of hot photons emitted at the recombination ($t \sim 3.8 \times 10^5$) that cooled down while travelling to us. The apparent isotropy of the emission argue in favour of the cosmological principle and the small temperature fluctuations advocates for small inhomogoneities in the initial matter distribution at the origin of the large-scale structures, as exposed in Sect. 1.2.1.

In the ΛCDM model, the Universe is born nearly 13.8 billion years ago from a singularity and, as indicated by its name, is made of three main components that are: (i) the cold dark matter (CDM) thought to be made of unrelativistic particles interacting only by gravity; (ii) the dark energy coming from the Λ cosmological parameters and responsible for the late-time acceleration of the Universe expansion; and (iii) the baryons composing the observable part of the matter like dust, stars and galaxies. The refined measurements provided by Planck Collaboration XIII et al. [2016] report that 25.8% of the total mass/energy inventory of the Universe is brought by the CDM and 69.4% which are actually associated to the dark energy. Even though filling most of the energy content of the Universe, the nature of the dark energy component remains unknown nowadays and is at the heart of many cosmological experiments trying to improve our understanding of this mysterious ingredient like Euclid [Laureijs et al., 2011] or the Dark Energy Survey [Abbott et al., 2016]. More importantly, this leaves us with only less than 5% of baryonic, visible matter which are distributed to form galaxies and over different gas phases [see for instance Fig. 8 from de Graaf et al., 2019]. Of course, several

Table 1.1. Proportion of the three main components of the ΛCDM model as measured by Planck Collaboration XIII et al. [2016].

Component	Mass-Energy proportion
Baryons	4.9
Cold dark matter	26.8
Dark energy	68.3

alternatives have been proposed for instance introducing dark energy in a different way, to extend the model accounting for additional free parameters like the summed neutrino mass M_ν or to modify general relativity [see e.g. Milgrom, 1983; Di Valentino et al., 2012, 2017; Maeder, 2017; Capparelli et al., 2018].

1.4.2 Cosmological parameters and matter power spectrum

Note that the six previous cosmological parameters were the fundamental ones describing the ΛCDM but some others are fixed or can be derived, including the Hubble constant H_0 and the density parameters Ω_m, Ω_b, Ω_Λ and ρ_crit already encountered in Sect. 1.1.2. Another common parameter of interest, equivalent to fixing the amplitude of the primordial spectrum A_s, is σ_8 which corresponds to the variance of the late-time matter fluctuations smoothed at a scale of 8 Mpc/h. Among the goals of observational cosmology are to test the validity of the ΛCDM model but also to provide accurate measurements its parameters. To do so, cosmologists rely on statistical representations such as the two-point correlation function of cosmological observables like the galaxy distribution, as discussed in Sect. 1.3. The cosmological parameter values can then be estimated by fitting parameters of the model to the observed statistics. By combining the information provided by different observables, the community was for instance able to draw a picture of the 3D matter power spectrum matching with precision the one from the linear matter prediction. This is shown in Fig. 1.2 where the solid line represents the best fit of ΛCDM linear theory and the coloured points are measurement provided by several probes, like the CMB to constrain the largest scales (low values of k), the galaxy distribution for the intermediate scales [such as data from Reid et al., 2016] and Lyman-α clustering [from quasar surveys like Abolfathi et al., 2018] for the smallest scales. These data agree remarkably well with the linear matter power spectrum obtained from the *Planck18* cosmology [Planck Collaboration VI et al., 2020], showing again the good agreement between ΛCDM and observations. Some parameters can however have similar impacts on the two-point matter clustering, like variations in σ_8 and Ω_m, leading to strong degeneracies and invoking the need of combining multiple information to break them. This can be done by using different probes like supernovae [see e.g. Abbott et al., 2019] or cluster analysis in Sunyaev Zel'dovich effect [see e.g. Salvati et al., 2018]. In Chapter 6, we shall also see how cosmic web environments can be used to constrain cosmological parameters of the ΛCDM model.

In this chapter, we have built the theoretical context of the large-scale structures formation. Starting from weak initial matter fluctuations and in the framework of first-order approximations, we were able to draw a picture of the late-time distribution of matter in which perturbations grow in an anisotropic way. This gives rise to the *cosmic web* [Bond et al., 1996] made of gravitationally collapsed structures in one, two or three spatial directions forming respectively the walls, filaments and nodes in the matter distribution. We have seen that such fields of fluctuations around the average noted δ can be represented by means of the random

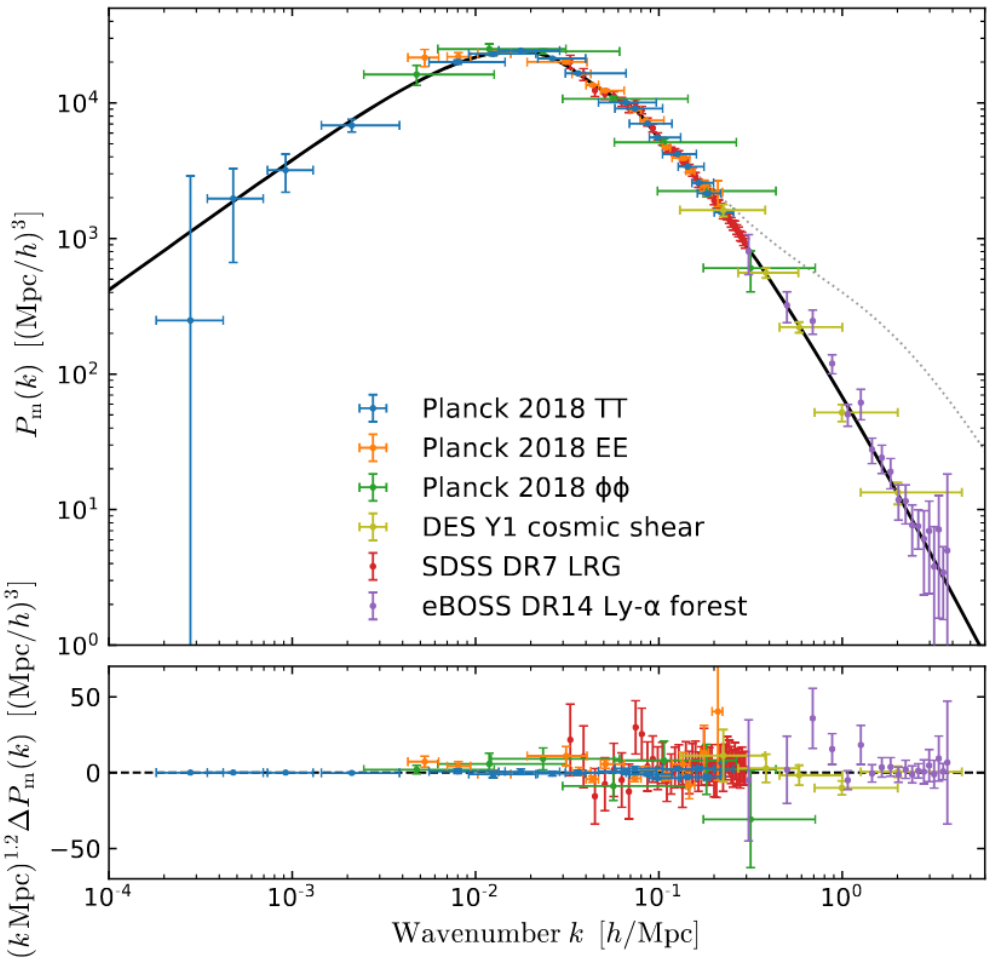

Fig. 1.2. The 3D linear matter power spectrum at $z = 0$ drawn from different observables. The solid black line is the theoretical expectation given the best-fit Planck Collaboration VI et al. [2020] cosmology while the dotted line is the non-linear power spectrum. Figure reproduced from Chabanier et al. [2019].

field theory and derived statistics like the n-point correlation functions are keystones that cosmologists seek to measure in order to enrich their knowledge about the Universe. Indeed, the power spectrum of the matter distribution provides a wealth of information on the cosmological model and its parameters that are themselves describing the dynamics of the observed Universe.

Chapter 2.

Large-scale structures manifestation

"Je préfère vivre dans le monde des simulations."
C. Gouin

Building upon the theoretical prescriptions of structure formation presented in Chapter 1, the current one supplies the complementary view of the large-scale structures provided by simulations and observations. We first introduce and motivate the central place of simulations for the study of non-linear physical processes and then focus on the difficulties induced by real-world observations of the cosmic web pattern. We finally emphasise through many previous works the importance of detecting cosmic structures and carry on an extended review of the current literature for doing so.

2.1 Large-scale structures in simulations

2.1.1 First exhibitions

Even with the limited computational resources available in the 70s, numerical simulations quickly appeared as an indispensable tool to study the evolution of the density field in the non-linear regime where $|\delta| \gg 0$. By resorting to approximations allowing the inclusion of the growth of non-linear structures through a first order approximation of perturbations, theory predicts the birth of structures depending on the eigenvalues of the local deformation tensor as exposed in Sect. 1.2.2 [Zel'dovich, 1970; Doroshkevich & Shandarin, 1978; Klypin &

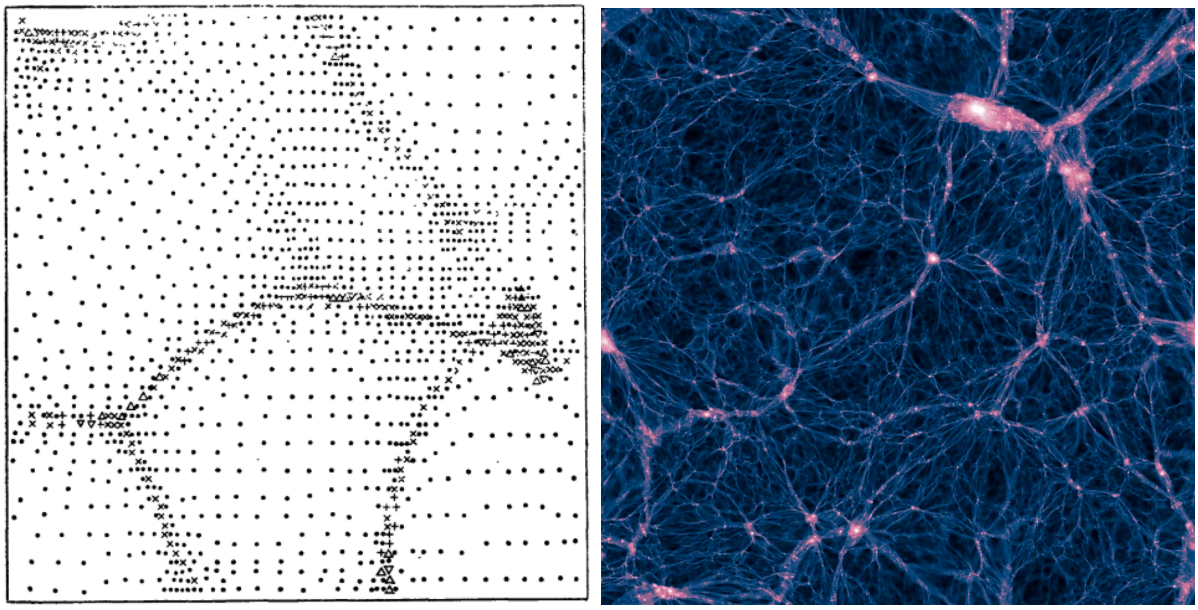

Fig. 2.1. The sophistication of simulations from the seventies to nowadays. *(left)* 2D distribution of points moved according first non-linear approximations in cosmological simulations. Image from [Doroshkevich & Shandarin, 1978]. *(right)* A 2D slice of the dark matter density field from the Illustris simulation [Vogelsberger et al., 2014].

Shandarin, 1983]. The left panel of Fig. 2.1 shows these early simulations of the formation of "dense pancake-shaped" structures [Doroshkevich & Shandarin, 1978] drawing for the first time what will be called later the "Cosmic Web" [Bond et al., 1996].

2.1.2 Dark matter only and hydrodynamical simulations

Early developments of simulations were based on the dark matter only evolution of the density field including solely the effect of gravity (also called N-body simulations). Starting from a set of initial conditions at very high redshift, usually taken from a Gaussian random field[1], the main idea is to dynamically evolve a set of N particles by solving the Vlasov-Poisson system of equations and iteratively move particles. Solving equations of motion for large N is a computationally heavy task and require N^2 operations at each timestep for which several sophisticated techniques emerged allowing a more efficient processing [Efstathiou et al., 1985; Springel, 2005]. The right panel of Fig. 2.1 shows the overdensity field computed in 2014 from a set of $N = 1820^3$ particles by the Illustris collaboration[2] showing the evolution made in this domain the past decades.

One of the main achievements of these N-body simulations is to allow the analysis of the density field at both large and small scales, reproducing accurately its statistics. Figure 2.2 shows for instance the power spectrum computed from one box of the Quijote simulation [Villaescusa-Navarro et al., 2020] and the one predicted by linear theory. We can see the deviation between the two at $k \sim 0.15\,h$/Mpc showing the incapacity of using such theoretical modelling at small scales. In principle, elements can be added to the linear theory to allow the description of smaller scales. Yet they are still limited, reaching the percent accuracy

[1] Some "constrained simulations" propose to tune the initial conditions of the simulation to reproduce the observed local Universe in a Bayesian framework and studies the properties of the time evolving reconstructed matter density field [see e.g., Kolatt et al., 1996; Jasche & Wandelt, 2013; Sorce et al., 2016].

[2] *https://www.illustris-project.org/*

at a mildly-non linear scales of $k \sim 0.3\ h$/Mpc [Carrasco et al., 2012] which is already an achievement but is not sufficient to carry out precise cosmological analyses and to understand the physical processes occurring at smaller scales (galaxy evolution, baryonic physics, etc.). Simulations are thus indispensable for the accurate assessment and analysis of gravitational dynamics at small scales, but also for the development of statistical tools used for the study of future large-volume surveys like Euclid [Laureijs et al., 2011] or Vera Rubin Observatory [Collaboration et al., 2009]. As an example, simulations are particularly interesting to assess how a statistics derived from an observable varies with cosmological parameters, or to build accurate covariance matrices, problems encountered in Chapter 6. For all these reasons, dark matter only simulations strive pushing further the scale limit by using always larger volume and finer resolutions, as for instance Millenium [Angulo et al., 2012], MultiDark [Klypin et al., 2016] or Quijote [Villaescusa-Navarro et al., 2020]. Beyond statistical analysis of the matter distribution, N-body simulations are also particularly interesting to study collapsed objects like halos identified by means of post-processing to refine predictions of their number counts or density profiles [Tinker et al., 2008; More et al., 2011].

However, the dark matter is not directly observable and, with the interest of reaching smaller and smaller scales, grew that of including baryonic matter in the simulations. Modelling the complex non-linear interactions between baryons, gas, stars, black holes and dark matter happening at all scales in cosmological volumes is one of the goals of hydrodynamical simulations. For that precise purpose, the inclusion of additional equations of motion is required which increases even more the complexity of the computation, hence requiring trade-offs between mass/volume resolutions and computational time of such simulations. Even with those difficulties, many large-scale hydrodynamical simulations were developed to study the role of baryonic physics in the evolution of large-scale structures, like Horizon-AGN [Dubois et al., 2014], EAGLE [Schaye et al., 2015], Illustris [Vogelsberger et al., 2014] or Illustris-TNG [Nelson et al., 2019]. As such, hydrodynamical simulations enlighten our understanding of structure formation and evolution of individual objects like galaxies or stars that can then be tested against astrophysical observations to support or refute the proposed models [Pearce et al., 2001; Dubois et al., 2014; Schaye et al., 2015; Crain et al., 2015; Nelson et al., 2019; Donnari et al., 2019].

2.2 From darkness to light: cosmic web and galaxies

2.2.1 Galaxy surveys

The first observations of the large-scale distribution of galaxies were made soon after the depiction of the filamentary pattern in simulations using the first available mappings of galaxies such as the Palomar Observatory Sky Survey [Joeveer et al., 1978] or the CfA survey show in Fig. 1.1. Since then, astronomers carried out extensive surveys to map as precisely as possible the galaxies observed in the sky and trace the filamentary pattern of the cosmic web. For that purpose, the past twenty years witnessed the succession of numerous surveys like the Sloan Digital Sky Survey [SDSS, York et al., 2000], the 2 Degree Field Galaxy Redshift Survey [2dFGRS, Colless et al., 2001], the Galaxy and Mass Assembly survey [GAMA, Driver et al., 2009] or the Dark Energy Survey [DES, Abbott et al., 2016]. This flood of spectroscopic and photometric surveys aim at both covering the largest portion of the sky and have the deepest possible observations. This race to the most accurate mapping is being still pursued nowadays and for the forthcoming years with surveys like the Vera Rubin Observatory's Large Synoptic Survey Telescope [LSST, Collaboration et al., 2009], Euclid [Laureijs et al., 2011] or the Dark

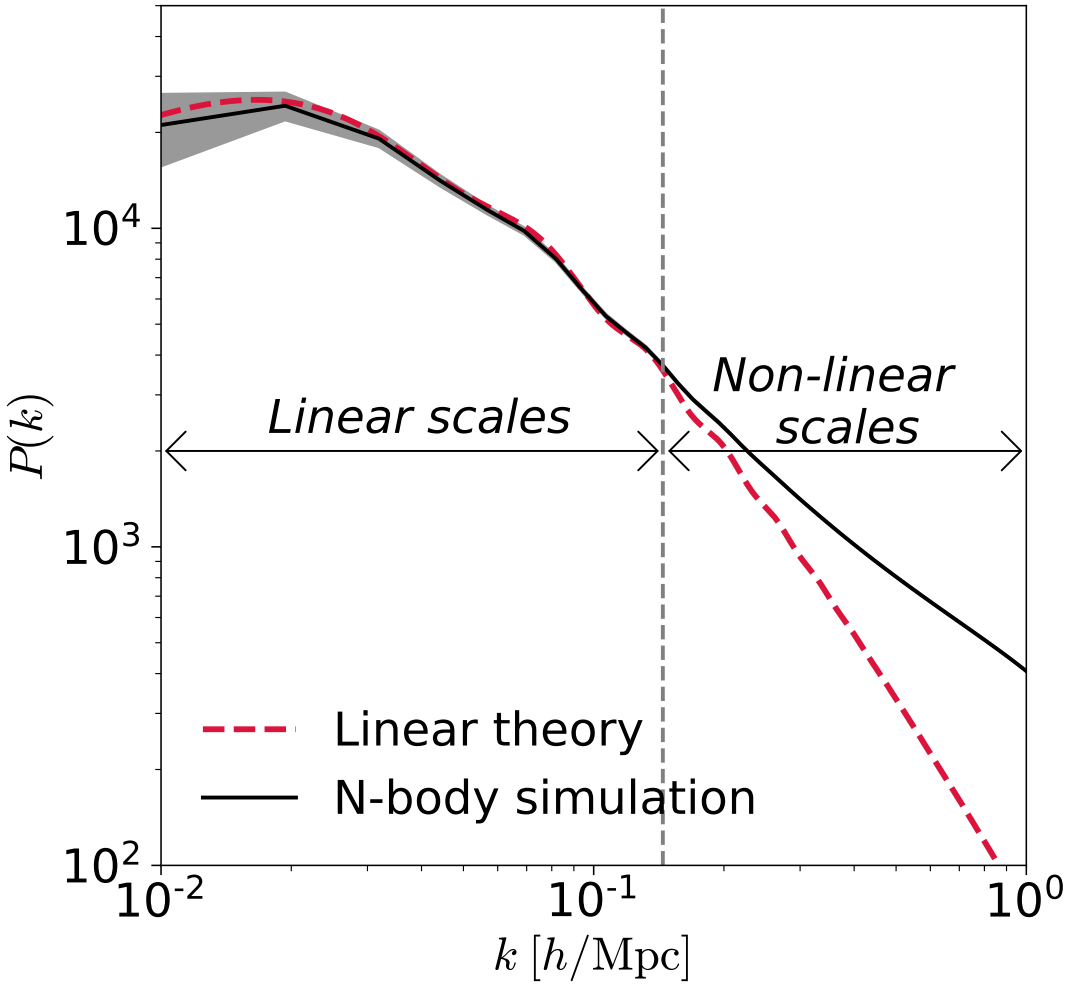

Fig. 2.2. Illustration of the crucial role of simulations for statistical analyses at small scales. The red dashed line shows the power spectrum computed from linear theory while the black one is the average computed from $7,000$ realisations of the Quijote simulation. The grey shaded area shows the 1-σ interval around the mean. Both agree at the percent level below the grey vertical dashed line at $k = 0.1455\ h/\mathrm{Mpc}$.

Energy Spectroscopic Instrument [DESI, Levi et al., 2013].

2.2.2 Observational effects

In addition to the general effects that one can expect from observations, like uncertainty on the estimated quantities, outliers and Poisson noise emerging from the discrete sampling of galaxies, cosmology is subject to additional effects related to the physics of the Universe and of the observed objects. In this section, we review the main observational effects and discuss their practical impact on the observation of large-scale structures.

Redshift-space distortions

When dealing with real-world observations, the redshift is used as a measure of the distance. However, this quantity is the linear combination of two contributions: the first one corresponding to the motion of the source (such as a galaxy) due to its peculiar velocity and the

another one due to the expanding universe. Considering $\boldsymbol{x}$ the source position in the comoving space, $\boldsymbol{r}$ its position in the redshift space, $\boldsymbol{v}$ its peculiar velocity and $\hat{\boldsymbol{n}}$ a unit vector in the line of sight (LoS) direction, it holds the mapping relation

$$\boldsymbol{r} = \boldsymbol{x} + \left(\frac{\boldsymbol{v}\cdot\hat{\boldsymbol{n}}}{aH(z)}\right)\hat{\boldsymbol{n}}, \tag{2.1}$$

with a the scale factor and $H(z)$ the Hubble parameter. It is worth mentioning that this effect, even though corrupting the position of observed objects along the LoS, also carries cosmological and astrophysical information about both the dynamics of galaxies and the expansion of the Universe. In particular, it is nowadays well established that redshift-space distortions (RSDs) provide key information on the growth rate of structures, denoted f, which scales with the matter density $\Omega_{\rm m}(z)^{\gamma}$ and can be used to study dark energy as well as alternative gravity models [Hamilton, 1998; Linder, 2005].

The fact that observations are carried out in redshift space has a direct implication on the collected data. More specifically, the spatial distribution of matter is distorted with overdense clustered regions appearing elongated in the LoS direction due to the high velocity of sources describing the overdensity. This effect is called "Finger-of-god" [Jackson, 1972] and is illustrated schematically in the right part of Fig. 2.3 where the peculiar velocities of the two sources having a non-zero velocity component with respect to the LoS elongate the shape of the spherical overdensity in this direction. At larger scales is observed a squashing effect of dense regions along the LoS [Kaiser, 1987] as shown in the left part of Fig 2.3. All these effects can be visually appreciated in the right panel of Fig. 2.4 where RSDs have been simulated by assuming the Z axis as the LoS and moving dark matter particles using Eq. (2.1).

One of the keystone of cosmology is the assumption of statistical homogeneity and isotropy of the Universe at large scales which makes the overdensity field δ a homogeneous random field invariant to rotations and translations. As depicted in Sect. 1.3, cosmological fields are statistically described by quantities derived from their correlations functions, or Fourier-equivalent poly-spectra. By breaking the isotropy, RSDs are breaking the rotation invariance of δ used for the establishment of some estimated statistics such as the matter power spectrum. This is more precisely described in Sect. 6.3.2.

Alcock-Paczynski effect

When working with observed objects, the translation of the estimated redshift $z_{\rm s}$ of a source into a physical distance $d_{\rm s}$ is expressed as

$$d_{\rm s} = \int_0^{z_{\rm s}} \frac{c}{H(z)}{\rm d}z, \tag{2.2}$$

and requires the assumption of a cosmological model for $H(z)$, the evolution of the Hubble expansion with redshift. Models of $H(z)$ vary depending on the mass $\Omega_{\rm m}$, curvature $\Omega_{\rm k}$ and dark energy Ω_{Λ} densities in the Universe. The difference between the assumed model and *the truth* induces some geometrical effects in the observed matter distribution known as Alcock-Paczynski distortions [Alcock & Paczynski, 1979]. In particular, it modifies the ratio between the radial extent and the angular size of an observed object (like clusters) making them either elongated or squashed along the LoS depending on the mismodelling of cosmology. As for RSDs, the Alcock-Paczynski effect turns out to be a way to assess cosmological models, and more precisely the expansion and geometry of the Universe by measuring the induced distortions when disentangled from the dynamical ones [López-Corredoira, 2014].

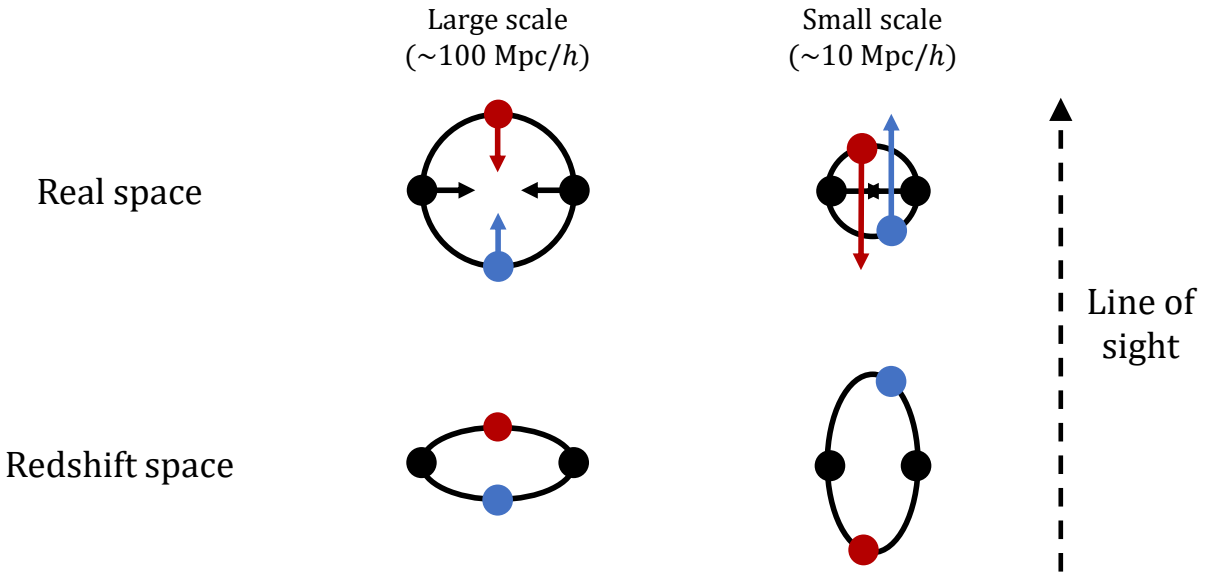

Fig. 2.3. Schematic illustration of the effect of redshift space distortions for spherical overdensities (represented as the black circles) at large and small scales. RSDs only alter the position of objects that have a non-zero velocity component along the line of sight (coloured points only).

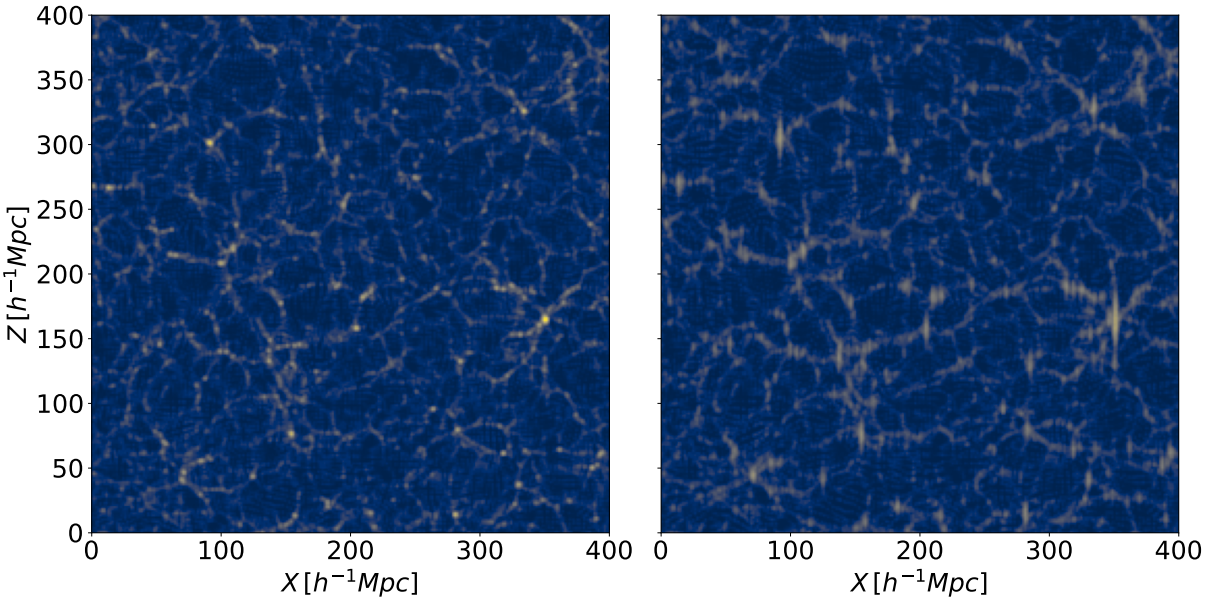

Fig. 2.4. Effect of redshift-space distortions in N-body simulations. *(left)* Representation of $\log_{10}(2+\delta)$ in real space on a 2D slice of the Quijote simulation. *(right)* Same in redshift space, assuming Z as the line of sight and displacing particles according Eq. (2.1) with $a = 1$ and $H = H_0$.

2.2.3 Galaxy bias

As presented in Chapter 1, the dark matter is the dominating matter component in the Universe which consequently rules the formation and evolution of gravitational potential. It is also on the dark matter overdensity field δ that most of the theoretical predictions are based in cosmology, such as the commonly used power spectrum of today's matter. The real-life observation of the large-scale structure is however performed through the luminous matter made, for instance, of galaxies. In theory, galaxies are the result of the non-linear evolution of dark matter creating potential wells in which baryons fall to form them and consequently trace preferentially overdense regions of the dark matter [Bardeen et al., 1986; Mo & White, 1996]. When one wants to link observations made from galaxies and the extracted statistics with theory, it is hence crucial to know how this statistics is actually linked with the one that would be derived directly from dark matter. This effect, called "bias", appears whenever a tracer different from the dark matter itself is used. The current literature hence provides a rich amount of information on the handling of bias induced by galaxies but also by other tracers of the matter distribution like quasars, voids or galaxy clusters, to name only a few. Accounting for this effect requires the understanding of how the spatial distribution of tracers is linked with the one of dark matter and can be studied either theoretically or numerically, with the above-mentioned observational effects complicating even more the picture. The simplest bias form implies a linear relation [Kaiser, 1984] between the overdensity field of tracers δ_{tracer} and of dark matter δ

$$\delta_{\text{tracer}}(k, z) = b(z)\,\delta(k, z), \tag{2.3}$$

with $b(z)$ the linear bias factor. In the case of the power spectrum, this translates into a constant $b(z)^2$ factor between the two spectra [Mo & White, 1996], only impacting its amplitude and inducing no shape dependencies. This simple linear bias model has shown great concordance for large and linear scales but more sophisticated bias models were investigated, accounting, for instance, for the tidal and environmental effects [Yang et al., 2017; Paranjape et al., 2018b]. In practical applications to real surveys, this bias is one of the biggest contributions to errors and is simply due to the lack of understanding of the tracers which appear as a set of "nuisance parameters" in the analyses. For a complete review on this rich topic, we refer the reader to Desjacques et al. [2019].

2.3 Motivations for cosmic web classification

2.3.1 The limitations of statistical analyses

Statistical representations of random fields based on the first orders of poly-spectra decomposition are limited in their representation of non-Gaussian patterns. In particular, most cosmological analyses of the matter distribution are based on the evaluation of the power spectrum which is completely insensitive to the texture of the cosmic web. Since $P(k)$ is only taking into account the modulus of the random field in the Fourier space, it omits the information contained in the phase. This is illustrated in Fig. 2.5 where the two fields have the exact same power spectrum, yet showing very different structural information, easily captured by eye. The inclusion of a sensitivity to this pattern in the analyses requires the evaluation of higher order statistics becoming already computationally hardly tractable at the order three with the bispectrum and come with a theoretical expression of the uncertainty which involves the fourth moment. The measure of such first high-order statistics also require many datapoints to be accurate.

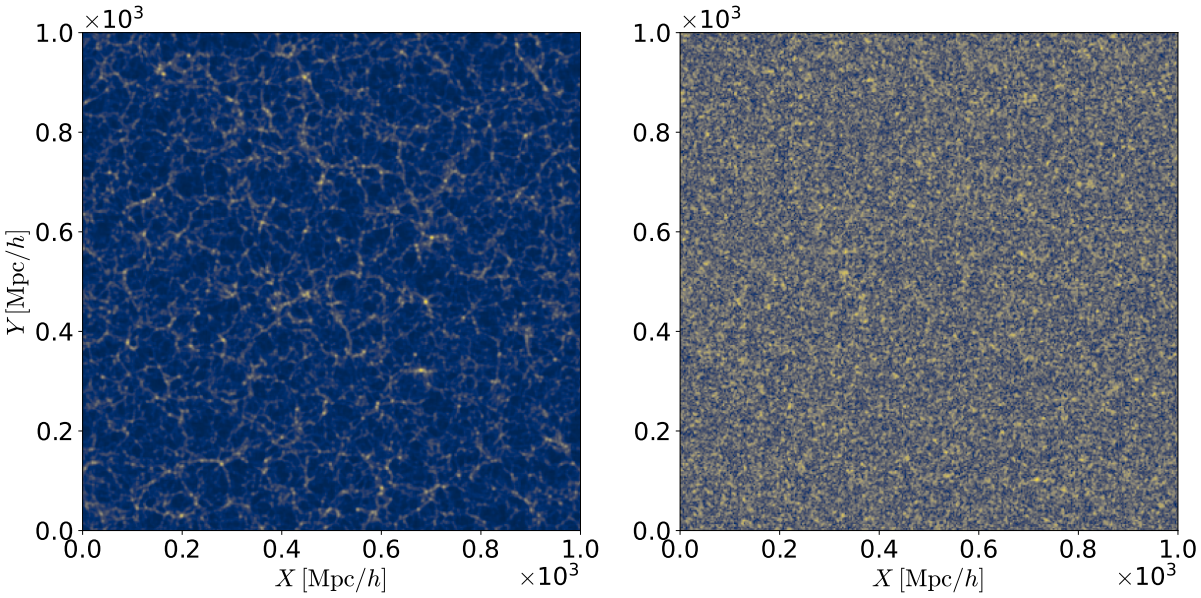

Fig. 2.5. The limitation of power spectrum analyses. *(left)* Overdensity field $\log_{10}(2+\delta)$ of a 2.77 Mpc/h depth slice from a N-body simulation. *(right)* Random phase reshuffling of the left panel. Both fields have the exact same evaluation of the power spectrum since $P(k)$ is blind from phase information.

As a consequence, many works try to identify the *best* way to reduce the information contained in these non-Gaussian fields, which remains one of the biggest challenge of modern cosmology. The fair representation of such fields should ideally satisfy three constraints: i) include, in some ways, higher-than-two orders information to go beyond the Gaussian representations; ii) be as compressed as possible such that only few coefficients are needed to store most of the information; iii) be physically interpretable, at least giving, in an empirical way, an idea of the information stored. To extract partial information from higher-order statistics, solutions based on topological criteria were for instance proposed back in the eighties [Gott et al., 1986; Mecke et al., 1994]. Using a set of topological invariants relying on Euler characteristics, Betti numbers and Minkowski functionals to describe non-Gaussian fields like the late-time matter distribution or weak lensing convergence maps has indeed successfully shown to encode higher-order information [Kratochvil et al., 2012; Shirasaki & Yoshida, 2014]. For the task of representing and analysing astronomical data, wavelet decomposition has also been extensively used. The first surveys were for instance quantitatively analysed by means of the wavelet transform [Martinez et al., 1993] which was also used to detect voids in the first galaxy catalogues [Slezak et al., 1993]. Later were formulated wavelets specifically sensitive to the different cosmic environments that are filaments and walls with the 3D ridgelets and beamlets [Starck et al., 2005; Woiselle et al., 2010]. This interest of statistically representing the matter or galaxy distributions never vanished and, still nowadays, recent developments improve the efficiency of the computation and analysis of the third order statistics at non-linear scales [Philcox & Eisenstein, 2020; Philcox, 2021; Hahn et al., 2020; Hahn & Villaescusa-Navarro, 2021] or exploit the additional information contained in velocities [Kuruvilla & Aghanim, 2021]. The rise of data science also led some older descriptions of non-Gaussian fields to be redesigned based on machine-learning derived tools. In that sense, the wavelet scattering transform [Mallat, 2012] mixes elements from convolutional neural networks [LeCun et al., 1999] and wavelet decomposition to compute a set of representative coefficients exploiting the invariance under

rotation and translation of the field at hand. The resulting summary is shown to be sensitive to higher-than-two points features in 2D cosmological fields in Cheng et al. [2020] and Allys et al. [2020]. Similarly, topological data analysis and persistent homology [see Edelsbrunner & Harer, 2008; Bubenik, 2015; Chazal & Michel, 2017] are extending the previous topological definitions by studying the properties of the topological features over a range of different spatial scales (which can be seen as a continuous analogous to the discrete scale-space analysis). The analysis of such topological features is also a promising way to encode non-Gaussian cosmic information, as shown in Wilding et al. [2020]. Even though interesting for the building of summary statistics taking into account the textural information, most of these statistical representation do not directly allow the possibility of a spatial identification nor extraction of cosmic structures.

2.3.2 The cosmological sensitivity of environments

Since the resulting pattern of the cosmic web is mainly driven by gravitational dynamics, the extraction of quantitative information from the observed structures provides key insights on the underlying cosmological model and enlighten our understanding of dark matter and dark energy. The first extensive and quantitative analyses of the multi-scale cosmic web in simulations suggests that each individual environment span a broad range of densities [see the right panel from Fig. 2.6, reproduced from Cautun et al., 2014] which in turn advocates for different cosmological histories. One could hence expect that individual environments inherit from different imprints and may show dissimilar behaviours with respect to cosmological models and parameters. As an example, voids are believed to be pristine environments, only little deformed by gravity and free from complex multi-streaming thus providing a perfect playground for the study of dark energy [Lee & Park, 2009; Lavaux & Wandelt, 2012; Hamaus et al., 2014, 2015; Pisani et al., 2015] or for constraining neutrino mass [Massara et al., 2015]. In the opposite way, clusters are highly non-linear objects with high over-density enclosing a large fraction of the mass for a small part of the volume. The statistics of these peaks (number, distribution with redshift) in the density field have been shown to be particularly sensitive to some cosmological parameters like the normalisation of the matter power spectrum or the matter density [Bahcall et al., 1997; Bahcall & Fan, 1998; Holder et al., 2001]. They are also unique laboratories to constrain the baryon gas fraction [White & Frenk, 1991; White et al., 1993] and to study the evolution of galaxies [Butcher & Oemler, A., 1984; Baldry et al., 2006]. This relationship between the different environments of the cosmic web and the cosmological parameters of the ΛCDM model is an aspect that we will develop in Chapter 6 using the two-point statistics of the different environments.

2.3.3 The role of the environment in shaping galaxies and clusters

At the astrophysics level, detecting cosmic structures may also help in proposing scenarios for the formation and evolution of galaxies. The first hints of environmental effects on galaxies were reported in [Oemler, 1974] showing that the densest regions of the Universe were more likely hosting elliptical than spiral galaxies. These observations were then refined with the recrudescence and availability of web finder algorithms. In particular, the most prominent structure, also traced in observations, is the filamentary part of the pattern. These massive bridges act like highways in the cosmic web, allowing the transport of the matter. In this picture, galaxies escape low-density regions and travel along the network being carried by the flow of matter in filaments towards the most massive parts of the web, the nodes [Aragon

Calvo et al., 2019; Cadiou et al., 2020]. This journey leaves an imprint on galaxy properties and revealing the filamentary pattern of the cosmic web in data and simulations hence offers the possibility to study the influence of the environment on the formation and evolution of galaxies. This topic has received a considerable interest these past years showing many correlations between the physical properties of galaxies (e.g. their mass, shape, luminosity, orientation or ability to form stars) or halos and those of the underlying web or related tidal anisotropies [Kauffmann et al., 2004; Hahn et al., 2007; Martinez et al., 2016; Kuutma et al., 2017; Malavasi et al., 2017; Laigle et al., 2018; Ganeshaiah Veena et al., 2018; Sarron et al., 2019]. For instance, it has been shown that galaxies closer to the spine of filament are more likely to be red and massive while it gets bluer and lighter when the radial distance is larger [e.g. Bonjean et al., 2020]. Some studies also draw a correlation between the orientation of galaxies and the direction of the filament they are hosted in [Ganeshaiah Veena et al., 2018, 2019; Kraljic et al., 2020].

The insightful analysis of the cosmic web in large-scale simulations carried out by Cautun et al. [2014] teaches us that the filamentary structure contains half of the dark matter mass of the Universe at the present time for only few percents of the total volume. By also hosting halos of various masses, typically from 10^{10} $\mathrm{M}_{\odot}/h$ to roughly 10^{13} $\mathrm{M}_{\odot}/h$, filaments are the ideal place to study collapsed objects like galaxies. Numerous works, based on simulations or observations, additionally show that a considerable fraction of baryons are hidden in the form of hot gas in filaments [Cen & Ostriker, 2006; Martizzi et al., 2019; Tanimura et al., 2020a; Galárraga-Espinosa et al., 2021] hence emphasising the crucial role this particular environment is playing in baryonic processes shaping the formation and evolution of galaxies. These findings highlight the importance of detecting the filamentary pattern both to improve the quality of the predictions in simulations and to discover new correlations with hosted tracers.

Galaxy clusters are massive biased tracers of the underlying matter observed both in simulations and surveys. It is now well-established that studying their properties like shapes, masses and redshift is a wealthy source of information on how they structure and evolve with time and on the underlying cosmological model [Yoshida et al., 2000; Peter et al., 2013; Sereno et al., 2018]. These properties have also been shown to be influenced by the local environment of halos and clusters and how they are locally embedded in the cosmic web [Poudel et al., 2017; Darragh Ford et al., 2019; Gouin et al., 2020]. In particular, Musso et al. [2018] expect that low-mass halos are more likely lying inside filaments while massive halos are found to be closer to nodes. cosmic web anisotropies are hence indicators of halo assembly bias and therefore strongly correlated with halo properties [Paranjape et al., 2018a; Ramakrishnan et al., 2019]. At a topological level, the number of filament a node, or massive cluster, is connected to, a quantity called the connectivity, is expected to depend on the growth factor hence allowing to put constraints on dark energy [Gay et al., 2012; Codis et al., 2018]. These relations between nodes and their local environments of the cosmic web will be investigated in more details in Chapter 5, Sect. 5.5.

2.4 Challenges in detecting cosmic filaments

Several difficulties make the detection of cosmic structures like filaments a challenging task. In this section, we discuss the main obstacles and review the several definitions adopted for cosmic filaments in the past and current literature.

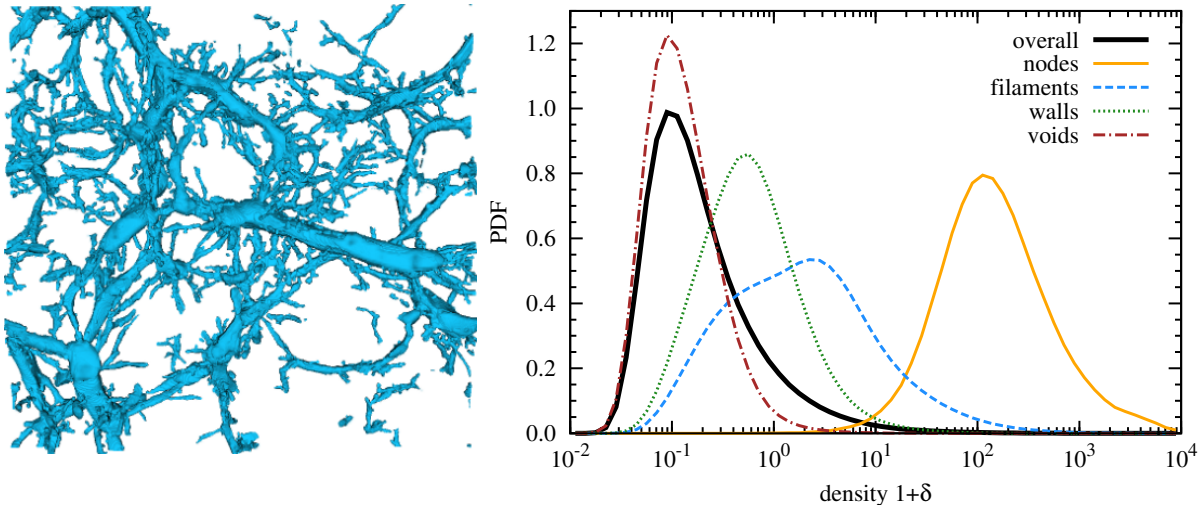

Fig. 2.6. Illustration of the complexity of cosmic web pattern. *(left)* The filamentary structure detected by NEXUS+ in the EAGLE [Schaye et al., 2015] simulation. Image from Fig. 1 of Cautun et al. [2014]. *(right)* Density PDF in each environment of the cosmic web as detected by NEXUS+. Plot from Fig. 13 of Cautun et al. [2014].

2.4.1 Structural complexity of the pattern

Cautun et al. [2014] draw a multi-scale picture of filaments probing several order of magnitudes in densities, from tenuous filaments of $\delta \sim 0.1$ to dense bridges of matter with overdensities of few hundreds, as shown in the right panel of Fig. 2.6. This large range implies overlapping in densities between the several environments and suggests the use of more refined methods than simple density thresholds. It is also shown that filaments are spreading over a wide range of widths with densities above the background from few to tenth of megaparsecs around the spine [Cautun et al., 2014; Galárraga-Espinosa et al., 2020], as illustrated in the filamentary structure depicted by the NEXUS+ algorithm [Cautun et al., 2013] in the left panel of Fig. 2.6. Designing methods applicable for data surveys also makes the picture more complex. Indeed, the incompleteness of the galaxy samples at all masses shows a pattern traced by the most luminous objects only and prevents an accurate sampling (illustrated in Fig. 2.7). Additionally, the redshift-space distortions induced by the peculiar velocities of galaxies alter the pattern drawn by the tracers. In particular, we have seen in Sect. 2.2.2, and illustrated by Fig. 2.4, that it makes clustered regions appear elongated along the line-of-sight and may hence create "fake filaments" [Malavasi et al., 2020a]. Finally, galaxy surveys are usually subject to holes in the observations with complex spatial selection functions creating absence of data in large portions of the sky. All these effects are, in addition to the natural complexity of the cosmic web pattern, creating further difficulties that the developed methods should take into account, either intrinsically or using appropriate pre- or post-processing.

2.4.2 Non-unicity of the definition

Section 2.3 emphasise how essential is the detection of cosmic environments for both improving our cosmological and astrophysical understanding of the Universe. However, there is no consensus on the definition of web elements and, these past decades have seen the emergence of many algorithms to identify structures with their own mathematical definition. Since filaments are of particular interest for astrophysics purposes and because these have been one of

the main topics of this thesis, we will discuss in detail the several definitions provided for this precise environment. A non-exhaustive list is reported in Table 2.1 together with the definition adopted for filaments in each case. For a more detailed view of classification schemes, we refer the interested reader to Libeskind et al. [2017] or to individual references mentioned in Table 2.1. Broadly speaking, there are six families of methods proposed for detecting cosmic web filaments:

Graph-based. The first representations of the filamentary pattern by Barrow et al. [1985] were carried out using a tool coming from graph theory called "the minimum spanning tree" that links galaxies together with the minimum total Euclidean distance to do so (discussed in more detailed in Sect. 4.2). In an attempt to extract quantitative information from it, Colberg [2007] and Alpaslan et al. [2014b,a] propose post-processing operations of the graph by cutting short edges or removing spurious ones to define filaments as branches of the tree structure. Thanks to its ability to easily capture the pattern with no parameter and by relying directly on the position of galaxies, the minimum spanning tree is still used today and extended to obtain smooth filaments, passing in the middle of the distribution of galaxies they host [Bonnaire et al., 2020, 2021b; Pereyra et al., 2020b,a]. At the interface between graph-based methods and point-based geometric ones, Chapters 4 and 5 will present ways to incorporate ideas of smoothness and robustness while learning a graph structure into a single formulation under the flag of principal graphs.

Point-based geometric. Also addressing the detection from the observational point of view by directly relying on discrete input, Stoica et al. [2005a, 2007] introduce a stochastic formalism based on the point-process theory defining filaments as a set of random connected and aligned cylinders paving the galaxy distribution and fulfilling some criteria based on sizes and local densities. The application of this latter, named the Bisous model in current surveys were later carried out by Tempel et al. [2014, 2016]. Alternatively, the FINE algorithm proposed in González & Padilla [2010], by considering physical principles based on the mass of halos or galaxies, is extracting filaments between two node tracers as the shortest line following the local highest density. In the same spirit, Genovese et al. [2014] and Chen et al. [2014] use a principal curve formulation to define filaments as the set of projected points standing on the ridge passing in the middle of galaxies in 2D datasets in the Subspace Constraint Mean-Shift algorithm [SCMS, Ozertem & Erdogmus, 2011]. Duque et al. [2021] recently proposed a way to overcome the hand-tuning of the fixed scale in the SCMS algorithm by using machine learning based combination of multiple scales.

Hessian-based. With the evolution of simulation accuracy and availability, a class of method based on a smooth estimate of the density field traced by dark matter particles emerged using the Hessian matrix of either the density [Aragon-Calvo et al., 2007; Cautun et al., 2013] or tidal [Hahn et al., 2007; Forero-Romero et al., 2009; Kitaura, 2013] field to classify cells according to their level of local anisotropy measured by the number of eigenvalues below a given threshold. The Multi-scale Morphology Filter [MMF, Aragon-Calvo et al., 2007] and NEXUS+ [Cautun et al., 2013] methods are carrying their analysis using the joint analysis of the curvature within a range of various Gaussian smoothing scales thus highlighting features of different sizes in the smooth field. Note that these classifiers provide a segmentation at the cell level but do not allow the identification of single objects like *a filament* without post-processing of the output.

Topological. In parallel to hessian-based methods were proposed descriptions of the density field invoking topological criteria [Sousbie et al., 2008; Pogosyan et al., 2009]. Topology is at

the heart of cosmological analyses of the large-scale galaxy distribution since its first observation [Gott et al., 1986; Mecke et al., 1994]. By describing the density field through the Discrete Morse Theory, the Discrete Persistent Structures Extractor [DisPerSE, Sousbie, 2011] proposes a mathematically elegant formulation of the cosmic network using prescriptions from computational geometry and persistent homology as the set of constant gradient lines connecting saddles and maxima. Such topological arguments coupled with watershed transform based methods, originally used for the detection of voids [Platen et al., 2007], led Aragón-Calvo et al. [2010] to define filaments as the intersections between three or more watershed basins in the SpineWeb algorithm.

Phase-space. Specifically dedicated to simulations, phase-space methods rely on the additional information of velocities to describe the structures using the full 6D information. By tracking the flow of matter escaping from empty regions to reach the cosmic network, Lavaux & Wandelt [2010] proposed a way to identify voids that Ramachandra & Shandarin [2015] extends to study all environments by counting the number of streams in the density field and, in particular, attribute to filaments those with more than 17 streams. A similar formalism, proposed by Falck et al. [2012], and followed-up by Falck & Neyrinck [2015] to identify voids, propose to count the number of orthogonal directions along which a shell-crossing occurs, which, for filaments, is two. Even though anchored in physical principles and useful to study the dynamics of the cosmic web, the intrinsic definition of this class of method makes them hardly applicable to real-world datasets since velocities of individual sources are only partly and difficulty measured.

Supervised Machine Learning. More recently, some attempts to apply supervised machine learning algorithms like the U-net architecture were initiated by Aragon Calvo et al. [2019]. By training the deep-learning model on a set of outputs obtained from the MMF method [Aragon-Calvo et al., 2007] or Voronoi mock dataset, it can then predict the class of a given overdensity cell. In the same vein, Buncher & Carrasco Kind [2020] train a Random Forest from local indicators of anisotropies and densities using k-nearest-neighbours and principal component analysis. However, the unavailability of ground truth training data strongly limits their application, even though the work of Aragon Calvo et al. [2019] suggests that it can generalise the results and identify structures like tenuous filaments beyond those of the training set.

2.5 Conclusions and perspectives for the thesis

The comparison of this landscape of methods is not easy and heavily depends on the targeted application and underlying goals. Eventually, there is no *best* definition for cosmic environments and each method come with its own caveats which point the user to a particular algorithm depending on the application. As a matter of fact, some algorithms, mainly those from the two first categories have been developed targeting point-based real-data applications by relying on matter tracers like galaxies. Other methods, usually requiring continuous input, have been mainly designed to be applied on simulations with dark matter particles as input to obtain an accurate estimate of the density field δ. These different inputs are illustrated in Fig. 2.7 showing in the bottom panel the galaxy distribution, visible counterpart of the dark matter density field tracing with much more details the cosmic web in the top panel. Even in the absence of other observational effects, the sparse sampling of the distribution complicates the task of extracting accurately the web-like pattern. Several applications of procedures based on continuous inputs on survey data have however been carried out with the DisPerSE algorithm

Table 2.1. A non-exhaustive list of the existing procedures to extract cosmic filaments. Table extended from Libeskind et al. [2017]. Algorithms whose names are tagged with an asterisk * are those requiring a continuous input, usually involving a pre-processing step to estimate the density field δ from the set of given particles.

Type	Method	Filament definition	References
Graphs	Adapted MST	Branches of a post-processed MST	Alpaslan et al. [2014a,b]
	Semita	Branches of a MST smoothed with splines	Pereyra et al. [2020b]
	T-ReX	Branches of a principal graph	Bonnaire et al. [2020], Chapter 4 and 5
Point-based geometric	Bisous	Connected and aligned cylinders	Stoica et al. [2007], Tempel et al. [2016]
	FINE	Lines linking two tracer nodes	González & Padilla [2010]
	SCMS	Ridges of the distribution	Chen et al. [2014, 2015] Duque et al. [2021]
Supervised Machine Learning	U-net*	Tag cells with a U-net trained on MMF outputs	Aragon-Calvo [2019]
	Random Forest	Tag particles with RF based on knn and PCA	Buncher & Carrasco Kind [2020]
Hessian	T-web*, CLASSIC*	Cells with local tidal anisotropies	Hahn et al. [2007], Kitaura & Angulo [2012]
	MMF*, NEXUS+*	Cells with multi-scale local anisotropies of δ	Aragon-Calvo et al. [2007], Cautun et al. [2013]
Topology	Spineweb*	Intersections between 3 watershed basins	Aragón-Calvo et al. [2010]
	DisPerSE*	Constant gradient lines between critical points of δ	Sousbie [2011] Sousbie et al. [2011]
Phase-space	ORIGAMI	Particles that have undergone shell-crossings along 2 axes	Falck et al. [2012]
	MSWA	Regions with more than 17 streams	Ramachandra & Shandarin [2015]

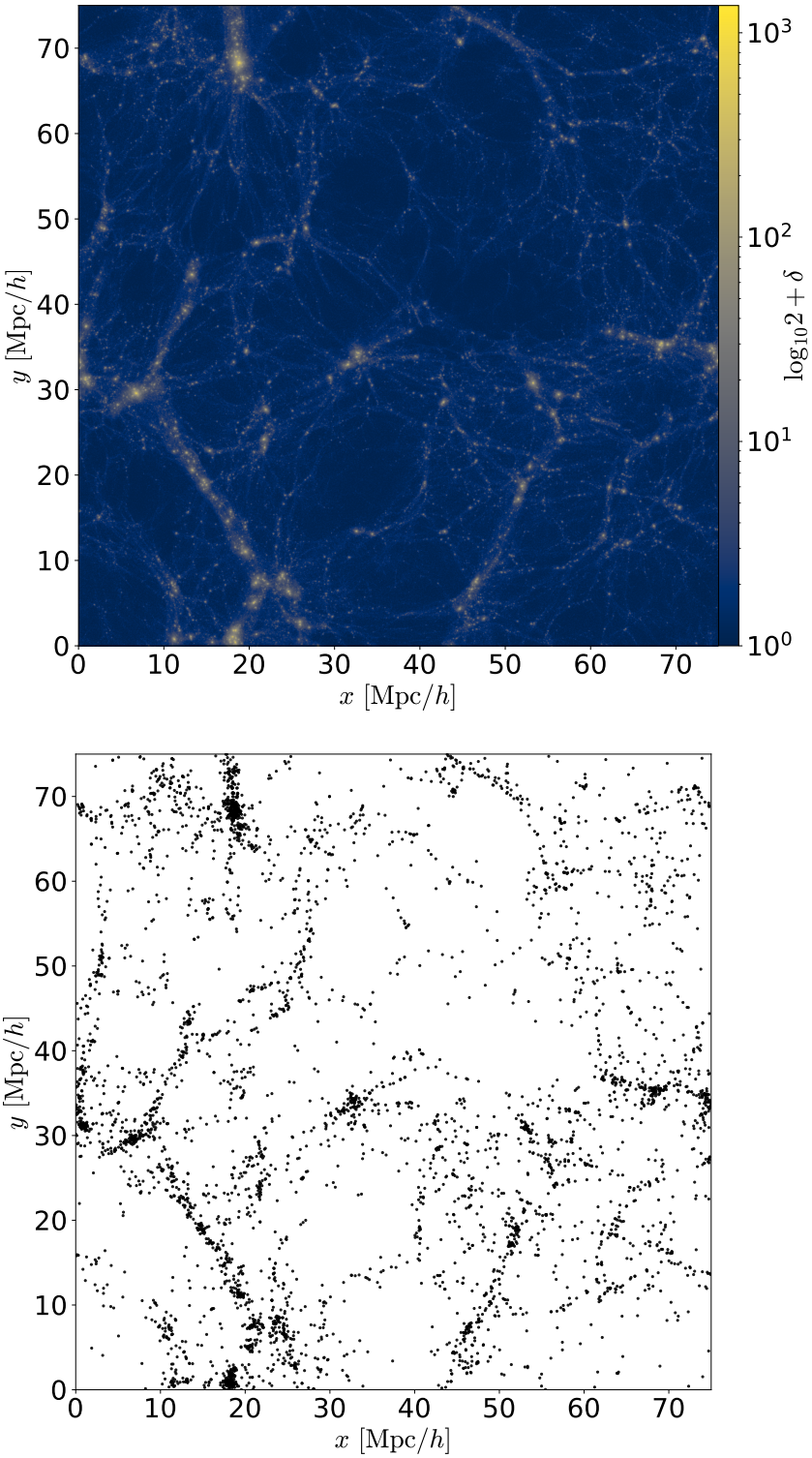

Fig. 2.7. *(top)* A slice from the Illustris simulation showing the estimated density field δ drawn from dark matter particles. *(bottom)* The galaxy counterpart. The sparse sampling of the pattern by the visible matter makes the extraction of the pattern more complicated in observations than in simulations where the full information is available.

[e.g. Malavasi et al., 2017; Laigle et al., 2018; Kraljic et al., 2020], and imply the computation of δ directly from the observed galaxy distribution. The accurate estimate of δ is hence an additional difficulty that needs to be tackled and for which several algorithms have been proposed, like the Delaunay Tessellation Field Estimator [DTFE, Schaap & Weygaert, 2000] or B-spline interpolation schemes [Hockney & Eastwood, 1981; Jing, 2005]. To alleviate the requirement of methods based on smooth field inputs, one could use jointly sophisticated reconstruction algorithms like BORG [Jasche & Wandelt, 2013; Leclercq et al., 2013] or Barcocde [Bos et al., 2014] providing a Bayesian estimate of the full underlying dark matter density field from its visible counterpart only (galaxies or clusters respectively). However, in direct applications, effects like masking or RSDs, are usually easier to handle in formalisms relying directly on discrete inputs like the MST [Alpaslan et al., 2014a]. They also generally allow the definition of filaments as continuous one-dimensional objects standing in the cloud of galaxies, which is not the case for other classifiers like NEXUS+ or T-web that need post-processing to obtain individual object detection.

Despite all these effects and the different definitions for the cosmic web that do not always agree one with each other [Libeskind et al., 2017], it is remarkable to see the many results obtained both in simulations and observations. In addition to the achievements listed in Sect. 2.3.3 on the interplay between the physical properties of filaments and galaxies, the detection of large-scale filaments in other wavelengths have been made possible with the recent advances in observational astronomy. This usually involves stacking methods to increase the signal after their detection via the distribution of galaxies. The observation of the filamentary pattern is hence currently performed and used in different observables like X-ray [Dietrich et al., 2012; Eckert et al., 2015; Nicastro et al., 2018; Tanimura et al., 2020b], weak lensing [Gouin et al., 2017; Epps & Hudson, 2017; Tanimura et al., 2020c] or through the Sunyaev-Zel'dovich effect [Bonjean et al., 2018; Tanimura et al., 2019; de Graaf et al., 2019; Tanimura et al., 2020c]. Yet, the observable that traces best the web-like pattern is the galaxy distribution from which filaments are first extracted to then study their properties in other wavelengths and this is why numerous works are providing catalogues of filaments to the community [e.g. Tempel et al., 2014; Chen et al., 2016; Malavasi et al., 2020b].

One of the aim of this manuscript is to propose ways to analyse such complex spatial patterns such as the one depicted by the spatial matter distribution. In particular, in Chapter 4, we present a method for the learning of a principal graph that extends some algorithms of the graph-based category. We then discuss the interest of the method for cosmological purposes and apply it to related datasets in Chapter 5.

Part II

Statistical methods for pattern extraction

Chapter 3.

Statistical physics for clustering

"De mon point de vue, si on arrive à comprendre,
c'est toujours bien..."
A. Decelle

This chapter is based on the results from Bonnaire et al. [2021a]. As the first chapter of the second part of the manuscript, it sets up the general context of machine learning and statistical methods of the thesis and discusses some of its caveats for applications in science, and more precisely in physics. It also introduces a generic formulation of the unsupervised optimisation problems encountered throughout the manuscript and the required knowledge about mixture models and the expectation-maximisation algorithm, two concepts exploited in the present and forthcoming chapters. Finally, the chapter exposes how a statistical physics formulation of an unsupervised machine learning task, the clustering, can be utilised to obtain an insight on the learning dynamics of the classes during the iterative procedure providing information about the structure of the data.

3.1 Context and motivations

3.1.1 Machine learning and physics

The incredible advances of Machine Learning (ML) we witnessed the past decade made it an indispensable tool embedded in our daily life through our smartphones, the software we use and even our cars. Beyond these industrial successes and thanks to its ability to extract information from huge amount of data, ML also sparked the interest of researchers in a large variety of domains ranging from physics to biology. Perhaps one of the most common applications to various fields is the image processing in which ML algorithms, and more precisely deep learning ones, have shown an unprecedented power in the learning of relevant features from large labelled databases [LeCun et al., 2015]. Astrophysics and cosmology were no exception and even showed themselves to be particularly interesting for the application of ML algorithms. The recent availability of data in cosmology, such as larger and larger simulations with various cosmologies (see Sect. 2.1) for instance paved the way for the estimation of cosmological parameters [Ravanbakhsh et al., 2017; Ribli et al., 2019; Cheng et al., 2020; Allys et al., 2020] or for the mapping of the dark matter and its visible counterpart [Zhang et al., 2019; Hong et al., 2021] using deep-learning-based architectures. The use of such methods also led to promising results in providing fast and accurate alternatives to the computationally costly N-body simulations [Rodríguez et al., 2018; He et al., 2019; Ullmo et al., 2021]. For observational tasks, ML-based algorithms also allowed the automatic classification of galaxies morphology based on optical data [see Barchi et al., 2020, for a review], the improvement of redshift estimation from photometric surveys [Carrasco Kind & Brunner, 2013] or the prediction of unknown galaxy properties like their masses, star formation rates or full spectra based on their photometry [Bonjean et al., 2019; Mucesh et al., 2021]. All these successes rely on different learning schemes that fall into three categories:

Supervised Learning in which is learnt a function mapping an input to an output based on paired examples. This goal can be achieved using various methods such as the random forest [Ho, 1995] or derived from neural-network architectures [Rosenblatt, 1959; Rumelhart et al., 1986], to cite only the most popular ones.

Unsupervised Learning methods that aim at identifying hidden structures or patterns in a given dataset without requiring labels. Usually based on the learning of the probability distribution (defined explicitly or not) that most probably generated a given dataset, famous algorithms of this category are generative models like Generative-Adversarial Networks [Goodfellow et al., 2014] and derived models [Mirza & Osindero, 2014; Radford et al., 2016; Arjovsky et al., 2017], Restricted Boltzmann Machines [Smolensky, 1986] or Variational Auto-Encoder [Kingma & Welling, 2013]. All the studied mathematical problems we will encounter in this part of the manuscript fall in this family, may it be the clustering or the principal graph learning presented in Chapter 4.

Reinforcement Learning [see Sutton & Barto, 2018, for an extensive introduction] is the task of learning multiple functions, called agents, to predict an output based on environmental information (for instance from sensors) in order to maximise a reward function. A famous example of algorithm in this category is the AlphaGo program [Silver et al., 2016] that learnt how to play Go against itself.

Despite their increasing popularity in physics applications and the performance they can achieve, ML approaches, in their fundamental idea of automatically "learning" a model from the sole information of the data remain very opaque. Particularly true in the case of deep learning, the learnt models and the features that were used to build them are generally not exploitable to *understand* the underlying phenomena, which is precisely what physics aims at. This opacity has led researchers to address questions like "Can we ourselves interpret the models learnt by the machine?" or "What are the important features in the data?". To an-

swer these questions but also more fundamental ones on the unusual generalisation abilities of deep learning models [Zhang et al., 2016; Hastie et al., 2019; Advani et al., 2020; D'Ascoli et al., 2020], the joint effort of various communities from computer science, mathematics and physics is being put together. Such answers are essential to provide prescriptions for the developed algorithms, their range of optimality but also increase their interpretability. For that purpose, leads can be found in the application of theoretical physics to ML algorithm. Back in the eighties already, the learning dynamics of the first neural networks [Little, 1974; Little & Shaw, 1978; Hopfield, 1982] were analysed by means of spin-glass models from statistical physics [Amit et al., 1985]. In a shameful attempt to give, in a few lines, an intuition on the connections between the two fields (and leave to Nishimori [2001] the task of giving the reader a more rigorous exposition), let us consider the case of the famous Convolutional Neural Networks [CNN, LeCun et al., 1999]. Taken individually, each component of the procedure is "simple", with linear operations (here, convolutions) interwined with somehow simplistic non-linear activation functions (like sigmoids). Parameters of the full resulting model (filter weights and biases) are estimated iteratively using gradient-descent-based algorithms minimising a quadratic cost functions. This composition of basic mathematical operations allows the modelling of complex functions and led to the best known results in numerous applications, as exposed above. On the other hand, statistical physics, in its most fundamental definition, is aiming at describing the microscopic interactions between simple elementary components of a system and understanding how complex behaviours can macroscopically emerge from it. In physical systems, statistical physics usually deals with a large number of particles and hence omits the individual characteristics to rather focus on the average properties of the entire system. In this analogy, the overall model learnt by the CNN is the macroscopic system while the many parameters are the particles composing it. Global characteristics of the model like its ability to generalise are analogous to macroscopic properties of the system. By setting themselves in idealised setups where the learning is "controlled", statistical physicists shown numerous analogies with ML – or related – algorithms allowing new theoretical insights and prescriptions [see Zdeborova & Krzakala, 2016, for a review]. In particular, many optimisation and inference problems have shown an equivalent formulation in statistical physics [Mezard & Montanari, 2009] that allowed a brand-new look at some long-standing problems and improved the understanding of complex algorithms [Mézard & Mora, 2009]. As an example, the identification of the phase diagram and phase transitions of a model can bring interesting insights such as knowing under what conditions a given algorithm is *optimal* or if a given information can be retrieved from the model and dataset at hand [such as Decelle et al., 2011; Lesieur et al., 2016; Tubiana & Monasson, 2017; Decelle & Furtlehner, 2021]. Such insights on the developed algorithms are precious, both in supervised and unsupervised contexts, since they provide key information on what matters during the learning process at different levels, from the topology of the cost function to be optimised [Choromanska et al., 2015; Spigler et al., 2019; D'Ascoli et al., 2020] to the impact of the data structure itself [Goldt et al., 2019].

In what follows, we give a general formulation of the optimisation problems that we will be focusing on in this part of the manuscript and introduce the concept of regularisation right before drawing the purpose and drawbacks of clustering algorithms this chapter is analysing.

3.1.2 Optimisation problems and regularisation

Inference (or learning) problems often come as the minimisation of an energy function (also called cost or loss function depending on the community) [see for instance Vapnik, 1998]. To keep the formulation general, let us consider two sets of points, a dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ with

$\boldsymbol{x}_i \in \mathcal{X}$ and $\{\boldsymbol{y}_k\}_{k=1}^K$ with $\boldsymbol{y}_k \in \mathcal{Y}$ such that $\{f(\boldsymbol{y}_k)\}$ builds a representation of the data, with $f\colon \mathcal{Y} \to \mathcal{X}$ a mapping function between the two spaces. Calling $\mathcal{F}$ the set of all such mappings, many mathematical problems in machine learning and inference contexts come formulated as

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{N} \min_{\boldsymbol{y}} E(\boldsymbol{x}_i, f(\boldsymbol{y})), \tag{3.1}$$

where $E(\boldsymbol{x}_i, f(\boldsymbol{y}))$ is the data fidelity term quantifying the energy cost of representing $\boldsymbol{x}_i$ by $f(\boldsymbol{y})$. As an example, one can express the empirical mean as the one-point representation of the dataset resulting from Eq. (3.1) under a quadratic energy cost with $K = 1$ and $\mathcal{F}$ the set of constant functions, leading to find $\boldsymbol{y}$ minimising $\sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{y}\|_2^2$. In the special context of statistics, one can see equivalently the maximisation of the log-likelihood function as the minimisation of an energy cost between the observed data and a model. The widely known and studied problematic of supervised machine learning can also be expressed in a similar manner. Supervised machine learning aims at finding the mapping f drawn from a set of tuples $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ encoding a signal and its label. A popular form of the cost function in these contexts is a quadratic energy cost $E(\boldsymbol{x}_i, f(\boldsymbol{y}_i)) = \|\boldsymbol{x}_i - f(\boldsymbol{y}_i)\|_2^2$. Note that this is also the cost emerging when assuming Gaussian additive noise in the error of the representation of $\boldsymbol{x}_i$ by $f(\boldsymbol{y}_i)$, i.e. a Gaussian likelihood in a probabilistic setting.

When the mathematical problem is inherently ill-posed and solutions exist only for a subset of $\mathcal{F}$, one can constrain the form of the solution, may it be to penalise its complexity or because of prior knowledge, by adding a regularisation term leading to

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \min_{\boldsymbol{y}} E(\boldsymbol{x}_i, f(\boldsymbol{y})) + \lambda R(f) \right], \tag{3.2}$$

where R is a functional penalising the solution and λ a parameter acting like a trade-off between the two terms. The introduction of such a penalty in the optimisation scheme is very popular in the machine learning community to avoid the over-fitting of the learnt models [Kukačka et al., 2017], but also in signal processing to relax some reconstruction problems [Jalalzai, 2012]. One of the common choices for R is an L_2 norm[1] hence penalising unsmooth functions leading to Tikhonov-like estimators [Tikhonov & Arsenin, 1977]. Under these assumptions of Gaussian distributed errors and quadratic regularisation, equivalent to a bounded L_2 norm constrained optimisation, the typical problem one has to solve is thus

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \left[\frac{1}{2} \sum_{i=1}^{N} \min_{\boldsymbol{y}} \|\boldsymbol{x}_i - f(\boldsymbol{y})\|_2^2 + \frac{\lambda}{2} \|f\|^2 \right], \tag{3.3}$$

which will be the generic form of most formulations we shall encounter in this part of the manuscript, like the clustering studied in this chapter or the principal graph in Chapter 4. Note that this choice of quadratic regularisation is not innocuous. From a Bayesian point of view, this is the form arising when imposing a Gaussian prior[2] distribution. Equation (3.3) thus corresponds to the maximisation of a log-posterior with a Gaussian log-likelihood and a Gaussian prior distribution. Thereby, λ can be seen as the ratio of variances between the

[1] L_1 regularisation is also very famous and known in the context of regression as the Lasso [Tibshirani, 1996]. This regularisation is used to build sparse representations of signals, forcing some coefficients to 0, a vast topic discussed in more details by Starck et al. [2010].

[2] Similarly, if we use an L_1 regularisation, optimisation problem (3.3) can be seen as the maximisation of a Gaussian likelihood with a Laplace prior distribution.

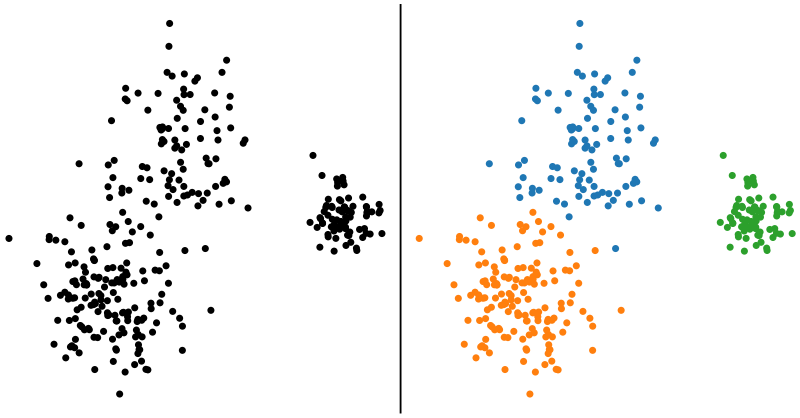

Fig. 3.1. Illustration of the purpose of clustering. Given a set of collected points in an arbitrary space (here the two-dimensional dataset in the left panel), one wants to identify $K = 3$ clusters (shown as coloured points in the right panel). Note that the choice of K is application- and user- dependent and that we could have chosen to identify only two clusters as well.

two Gaussian laws. Note also that, when maximising Eq. (3.3) with $\lambda = 0$, the variance of the likelihood does not play any role and is coupled with the variance of the prior distribution when $\lambda > 0$.

3.1.3 Clustering and its drawbacks

Clustering is a one of the most ancient unsupervised tasks of ML aiming at identifying a partition of a given dataset into multiple groups called "clusters". Figure 3.1 illustrates this objective in a 2D case where the aim is to go from the left panel to the right one, by attributing a class to each input datapoint. The common ground to many methods that have been proposed these past decades [MacKay, 2002] is that they all embed, in some ways, a measure of the "similarity" between datapoints living inside the same cluster. The simplest way of thinking this is a measure based on the Euclidean distance leading to representations in which "close" datapoints more likely fall into the same cluster while distant ones should reside in different clusters. The prototypical method to partition a D-dimensional dataset into K clusters is the K-means algorithm [MacQueen, 1967]. Taking back the general formulation of optimisation problems of Sect 3.1.2, we can write the K-means goal as finding the set of points minimising the sum of squared distances between datapoints and cluster centres. We end up in a setting seeking to find the set of points $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}^\mathsf{T}$ minimising

$$\sum_{i=1}^{N} \min_{k \in \{1,\dots,K\}} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2. \tag{3.4}$$

The K-means algorithm solves this optimisation problem in a simple way. Starting from a set of initial positions $\boldsymbol{\mu}^{(0)}$, it assigns to each datapoint $\boldsymbol{x}_i$ the closest (in the sense of the Euclidean distance) centroid among $\boldsymbol{\mu}$, noted $\boldsymbol{\mu}(\boldsymbol{x}_i)$, and moves $\boldsymbol{\mu}_k$ accordingly to the average of all datapoints projecting on it, namely $\boldsymbol{\mu}_k^{(t+1)} = \mathbb{E}(\boldsymbol{x}_i \,|\, \boldsymbol{\mu}(\boldsymbol{x}_i) = \boldsymbol{\mu}_k^{(t)})$. As we can see, this formulation assigns a single centroid to a datapoint without any level of uncertainty, making it a "hard" version of clustering. In case of overlapping clusters, such as the two on the left of Fig. 3.1, it is however natural to think that datapoints living at the border could be either in the blue or the orange cluster. Also, because the association energy appearing in Eq. (3.4) is

based on the Euclidean distance, it tends to generate spherical groups. These restrictions of the K-means algorithm later led to the Gaussian mixture model formulation of the clustering that we will present in this chapter. This latter introduces a fuzziness by assigning a probability to each datapoint of being generated by one the K clusters but also extends the definition to anisotropic clusters of various densities.

Thanks to their ability to describe clustered patterns by grouping datapoints without prior information, clustering algorithms naturally found many applications in science [see Saxena et al., 2017, for a review]. In the particular context of astrophysics and cosmology, they found applications in the identification of astronomical sources (stars, quasars, galaxies, etc.) together with their morphological properties to complete and improve survey catalogues [see e.g. Aghanim et al., 2015; Barchi et al., 2016; Logan & Fotopoulou, 2020]. Despite their attractive data-driven and unsupervised foundations[3], clustering algorithms are not free of flaws and, in addition to specific drawbacks proper to each method, most of them come with the requirement of inputting a number of cluster K to identify in the dataset. They also usually act as opaque boxes and the clustering procedure only outputs a partition of the dataset without any information on what has been "learnt" from their separation. For a successful application of most clustering procedures, one hence needs to have strong priors on the wanted clusters, such as their number, and a broad idea of the family of datapoints that need to be grouped together, even if not labelled, to assess the result of the clustering.

The aim of the rest of the chapter is to expose precepts originating from statistical physics to obtain a physical insight on what happens during the clustering procedure when carried out using the Gaussian mixture model and the Expectation-Maximisation procedure to optimise the log-likelihood. After having introduced both concepts, we will show that, in the context of deterministic simulated annealing, we are witnessing a cascade of phase transitions and that we can study the linear stability of the iterative scheme of the Expectation-Maximisation algorithm to derive theoretical thresholds at which these transitions occur. More importantly, we show that, by tracking empirically some physically relevant quantities related to the size of the clusters being learnt by the model, we end up with crucial information on the number of structures at a given scale, their hierarchy (i.e., how they are embedded in space with respect to each other), and their size. This is done without relying on prior knowledge, such as the number of components usually required by clustering methods. All the physical transitions are visible in a two-dimensional diagram which allows the extraction of the structural information, even in high dimensions when visualisation is not possible, hence opening the path for the exploration of complex high-dimensional datasets.

3.2 Mixture models

3.2.1 General formalism

Real-life data often come as drawn from probability distributions with complex shapes and multiple modes that cannot be satisfactorily described by a single well-known probability distribution. Mixture models [see McLachlan & Krishnan, 1997; Bishop, 2006] propose to model this complexity by using a linear combination of K known laws. Figure 3.2 illustrates this

[3] In fact, some clustering approaches are not fully unsupervised and exploit partial pre-classified data to guide the overall clustering, making them semi-supervised [see e.g. Grira et al., 2004; Bair, 2013].

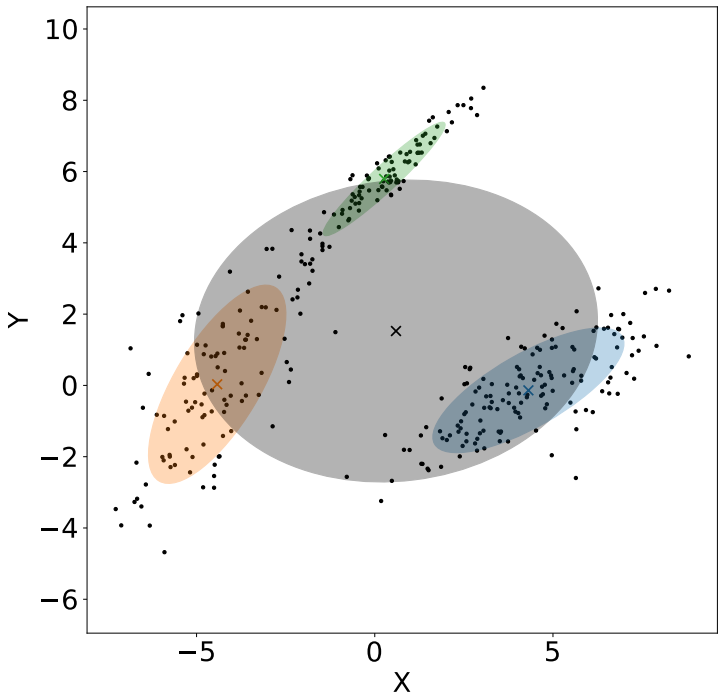

Fig. 3.2. Illustration of the interest of mixture models. Black dots are datapoints generated by three Gaussian clusters. Green, orange and blue ellipses are 1-σ confidence ellipses assuming a Gaussian mixture model with $K = 3$ while the black ellipse is obtained by maximising a Gaussian likelihood from the data. Discussed in Sect. 3.2.

need on a multi-modal 2D distribution of points. Although easy to study, a single Gaussian component (grey ellipse) do not explain the non-Gaussian dataset while a linear combination of three Gaussian (coloured ellipses) leads to a better representation. This idea of combination of known laws, despite its conceptual simplicity, can lead to accurate representations of highly complex density distributions. This is in particular why mixture distributions are nowadays at the basis of many mathematical tools like kernel density estimation [Parzen, 1962; Li & Barron, 2000], clustering [Jain et al., 2000] or mixture density networks [Bishop, 1994] in machine learning.

Keeping the formulation as general as possible, a mixture distribution is the linear combination of K probability distributions $\{f_k\}$ with parameters $\boldsymbol{\theta}_k$. The probability of a datapoint $\boldsymbol{x}$ being generated by the model is hence

$$p(\boldsymbol{x} \mid \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}, \boldsymbol{\theta}_k), \tag{3.5}$$

with $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ the set of model parameters, π_k the mixing coefficient (also called amplitude) of component k. Note that, given the properties of probability distributions, amplitudes are normalised and positive by definition, namely $\sum_{k=1}^{K} \pi_k = 1$ and $\forall k \in \{1, \ldots, K\}, \pi_k \geq 0$.

3.2.2 The Gaussian case

A particular class of mixture model is the Gaussian case, where $\forall k \in \{1, \dots, K\}$, $f_k(\boldsymbol{x}, \boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{x}, \boldsymbol{\theta}_k)$, with

$$\mathcal{N}(\boldsymbol{x}, \boldsymbol{\theta}_k) = \frac{\exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)^\mathsf{T} \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}_k\right)\right\}}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_k|^{1/2}}, \tag{3.6}$$

the Gaussian probability distribution with parameter $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, respectively corresponding to the average of the k^{th} Gaussian component and its covariance matrix. The particular Gaussian formulation of mixture models has been extensively used in machine learning, mainly for density estimation and clustering purpose.

3.3 Expectation-Maximisation algorithm

3.3.1 Introduction through Mixture Models

In a parametric setup, one can estimate the optimal set of parameters $\boldsymbol{\Theta}$ of a model by maximising the log-likelihood function $\log p(\boldsymbol{X} \mid \boldsymbol{\Theta})$. Assuming a mixture model as defined in Eq. (3.5), we get

$$\log p(\boldsymbol{X} \mid \boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i, \boldsymbol{\theta}_k) \right). \tag{3.7}$$

This function cannot be analytically maximised because of the sum inside the logarithm function. To circumvent this issue, the Expectation-Maximisation algorithm [EM, Dempster et al., 1977] allows the optimisation of the log-likelihood of models with latent variables. To reformulate the mixture model in terms of latent variables, we first introduce a set of random variables $\boldsymbol{Z} = \{z_i\}_{i=1}^N$ with $z_i \in \{1, \dots, K\}$. $\boldsymbol{Z}$ represents the partition of the dataset and z_i denotes by which of the K component the datapoint $\boldsymbol{x}_i$ has been generated from. By doing so, the log-likelihood from Eq. (3.7) can be written as the marginal distribution

$$\log p(\boldsymbol{X} \mid \boldsymbol{\Theta}) = \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}), \tag{3.8}$$

where the joint probability distribution $\log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta})$ is often referred to as the "completed log-likelihood" of the model in the sense that the joint knowledge of datapoints and latent variables $\{\boldsymbol{X}, \boldsymbol{Z}\}$ is the "completed dataset". It can be expressed as

$$\log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \left(\pi_{z_i} f(\boldsymbol{x}_i, \boldsymbol{\theta}_{z_i}) \right). \tag{3.9}$$

The key idea of EM is that the log-likelihood of $\{\boldsymbol{X}, \boldsymbol{Z}\}$ is easier to maximise than the one of Eq. (3.7): if we knew the vector $\boldsymbol{Z}$, we could easily maximise Eq. (3.9) and find parameters $\boldsymbol{\Theta}$ of the model. Unfortunately, $\boldsymbol{Z}$ are unknown hidden variables that we need to estimate, in addition to $\boldsymbol{\Theta}$. For this precise purpose, EM provides a procedure with two alternating maximisation steps to iteratively estimate both quantities. The presented formalism is more general than the context of mixture models and Eq. (3.8) remains true for any model with latent variables. In this section, if not clearly stated, we consider a general log-likelihood, without referring to the particular one of Eq. (3.7).

3.3.2 Iterative scheme

From the observation that the completed log-likelihood $\log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta})$ is easier to maximise, the goal is thus to obtain an estimate of the probability distribution over the latent variables. Jensen's inequality[4], from the concavity of the logarithm function, together with Eq. (3.8) allow us to write, for any normalised distribution $q(\boldsymbol{Z})$ over the latent variables,

$$\log p(\boldsymbol{X} \mid \boldsymbol{\Theta}) \geq L(q, \boldsymbol{\Theta}) \coloneqq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta})}{q(\boldsymbol{Z})} \right]. \tag{3.10}$$

Hence, the log-likelihood of the model is bounded below by $L(q, \boldsymbol{\Theta})$. If easier to handle, maximising this quantity could help maximising $\log p(\boldsymbol{X} \mid \boldsymbol{\Theta})$. We can re-write this lower-bound, using the normalisation of $q(\boldsymbol{Z})$ and the decomposition $p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}) = p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}) p(\boldsymbol{X} \mid \boldsymbol{\Theta})$, as

$$L(q, \boldsymbol{\Theta}) = -D_{\mathrm{KL}}(q, p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})) + \log p(\boldsymbol{X} \mid \boldsymbol{\Theta}), \tag{3.11}$$

where $D_{\mathrm{KL}}(q, p) \coloneqq \sum q \log q/p \geq 0$ is the Kullback-Leibler divergence [Kullback & Leibler, 1951]. To maximise the lower-bound $L(q, \boldsymbol{\Theta})$, the Kullback-Leibler term, at the origin of the inequality in Eq. (3.10), has to cancel out, hence leading to estimate the probability distribution of the latent variables. To this end, the current values of the parameters, $\boldsymbol{\Theta}^{(t)}$, are used to compute $\hat{q}(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)})$ that is then exploited to obtain the lower-bound $L(\hat{q}, \boldsymbol{\Theta}^{(t)})$. This step is called the "E-step" as a reference to "Expectation" because the lower-bound can be written, using Eq. (3.10), as

$$\begin{aligned} L(\hat{q}, \boldsymbol{\Theta}) &= \mathbb{E}_{\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}} \{ \log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\Theta}) \} - \sum_{\boldsymbol{Z}} \hat{q}(\boldsymbol{Z}) \log \hat{q}(\boldsymbol{Z}), \\ &\coloneqq Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) + H(\boldsymbol{\Theta}^{(t)}), \end{aligned} \tag{3.12}$$

where Q is the expectation of the complete log-likelihood over the latent variables and H is the negative log-entropy of the latent variables distribution $\hat{q}(\boldsymbol{Z})$.

In a second step, the values of the parameters are updated by maximising the lower-bound over $\boldsymbol{\Theta}$,

$$\begin{aligned} \boldsymbol{\Theta}^{(t+1)} &= \operatorname*{argmax}_{\boldsymbol{\Theta}} L(\hat{q}, \boldsymbol{\Theta}), \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}). \end{aligned} \tag{3.13}$$

This step is the M-step, standing for "Maximisation". The full EM algorithm, by alternating between E and M steps, iteratively maximises the log-likelihood of the model through the double maximisation of a lower-bound, first over the distribution of latent variables and then over the parameters of the model, $\boldsymbol{\Theta}$.

One of the key advantages of EM, beyond its tractability, is its guaranteed convergence towards a local maximum of the log-likelihood through successive monotonic increases [Wu, 1983; McLachlan & Krishnan, 1997]. To give an intuition of that, it is straightforward to write $L(\hat{q}, \boldsymbol{\Theta}^{(t)}) = \log p(\boldsymbol{X} \mid \boldsymbol{\Theta}^{(t)})$ from Eq. (3.11). Then, $\boldsymbol{\Theta}^{(t+1)}$ is chosen such that maximising $L(\hat{q}, \boldsymbol{\Theta})$ which yields $L(\hat{q}, \boldsymbol{\Theta}^{(t+1)}) \geq L(\hat{q}, \boldsymbol{\Theta}^{(t)})$. Since $L(\hat{q}, \boldsymbol{\Theta})$ is a lower-bound of the log-likelihood, increasing it guarantees the increase of the log-likelihood, which in fact increases by an amount of $D_{\mathrm{KL}}(q, p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}))$, as suggested by Eq. (3.11). To better visualise this property, Fig. 3.3 illustrates one iteration of the algorithm.

[4] Stating that, for any random variable X and convex function f, $f(\mathbb{E}\{X\}) \leq \mathbb{E}\{f(X)\}$ [Jensen, 1906].

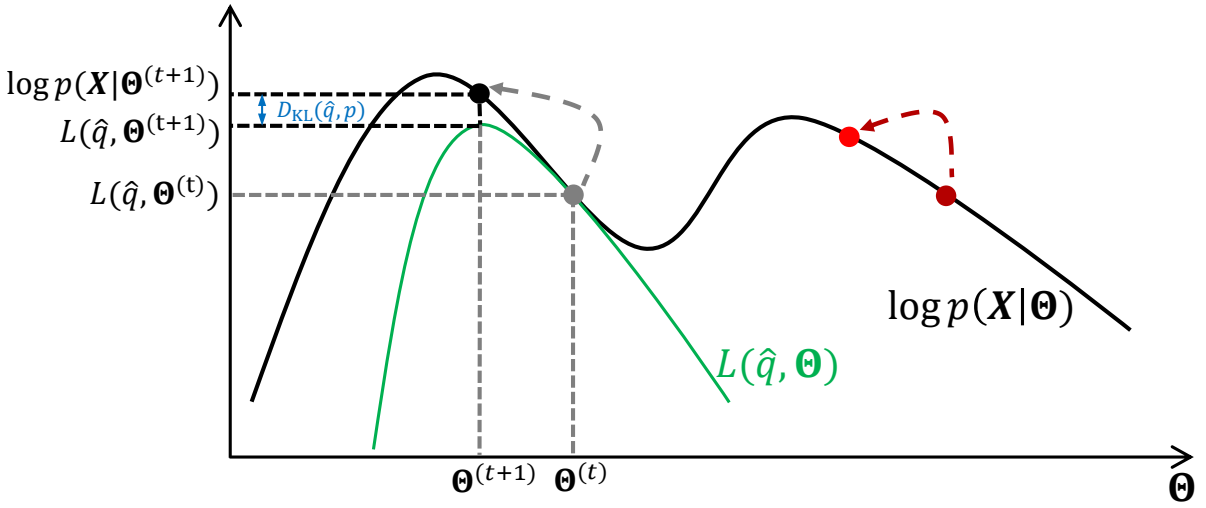

Fig. 3.3. Schematic view of one iteration of the Expectation-Maximisation procedure. The grey dot represents an iteration t. By applying the E and M steps, one moves to the black dot, climbing the log-likelihood towards the local maximum. Starting from another point (red one) would however lead the optimisation to converge to a lower value of the log-likelihood. Details can be found in Sect. 3.3.2.

Despite these mathematically appealing characteristics, some of the main drawbacks of the EM algorithm are found in its convergence properties, mainly because of its slow convergence rate [Wu, 1983] and its dependency on the initialisation [Kloppenburg & Tavan, 1997; Ueda & Nakano, 1998]. Indeed, the direct application of the procedure is known to be easily trapped in local maxima of multi-modal likelihoods leading to variability in the provided results depending on the initialisation of the algorithm. This can be intuited from Fig. 3.3 where starting from different points (grey or red) will irremediably lead to one of the two modes of the log-likelihood. When used in practice, it is hence natural to start from several random locations and keep the realisation with the highest value of the final log-likelihood (as implemented in Pedregosa et al. [2011] for instance).

3.3.3 The particular case of Gaussian mixtures

When considering a mixture model (see Sect. 3.2), the log-likelihood and the completed log-likelihood are respectively given by Eq. (3.7) and Eq. (3.9). In this scenario, we can derive, in the E-step, the probability of the latent variables given the observed data and the current parameters of the model, also called responsibilities, $p_{ik} \coloneqq p(z_i = k \,|\, \boldsymbol{x}_i, \boldsymbol{\Theta}^{(t)})$ using Bayes' theorem,

$$p_{ik} = \frac{\pi_k \, f(\boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t)})}{\sum_{j=1}^{K} \pi_j \, f(\boldsymbol{x}_i, \boldsymbol{\theta}_j^{(t)})}. \tag{3.14}$$

This estimation is then used in the M-step to evaluate the first term of the lower-bound through Eq. (3.12) (the other one being independant of $\boldsymbol{\Theta}$) as

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \log \left[\pi_k \, f(\boldsymbol{x}_i, \boldsymbol{\theta}_k)\right], \tag{3.15}$$

and update the parameters through Eq. (3.13) by maximising over $\boldsymbol{\Theta}$. Note that, in this equation, parameter values referring to time t are hidden in the responsibilities p_{ik} of Eq. (3.14).

In the case of a Gaussian mixture model (GMM, see Sect. 3.2.2), we have $f_k(\boldsymbol{x}_i, \boldsymbol{\theta}_k) = \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\theta}_k)$, with $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. The M-step consists then in solving

$$\underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik} \left[\log \pi_k - \frac{1}{2} \log \boldsymbol{\Sigma}_k - \frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k \right)^{\mathsf{T}} \boldsymbol{\Sigma}_k \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k \right) \right], \tag{3.16}$$

where we can recognise a relaxed version of the general form of Eq. (3.2) with $\lambda = 0$ (no prior). Indeed, compared to the rigid K-means formulation of Eq. (3.4) in which a datapoint is associated to a unique cluster, the GMM method not only includes a parameter for the shape of the cluster through $\boldsymbol{\Sigma}_k$ but also quantifies the probability of a datapoint $\boldsymbol{x}_i$ to be represented by a given cluster k through the responsibility p_{ik}.

From Eq. (3.16), it is possible to derive an update equation for each parameter of $\boldsymbol{\Theta}^{(t+1)}$ as

$$\begin{aligned} \pi_k^{(t+1)} &= \frac{1}{N} \sum_{i=1}^{N} p_{ik}, \\ \boldsymbol{\mu}_k^{(t+1)} &= \frac{\sum_{i=1}^{N} \boldsymbol{x}_i \, p_{ik}}{\sum_{i=1}^{N} p_{ik}}, \\ \boldsymbol{\Sigma}_k^{(t+1)} &= \frac{\sum_{i=1}^{N} p_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathsf{T}}}{\sum_{i=1}^{N} p_{ik}}. \end{aligned} \tag{3.17}$$

Note that, if we consider spherical Gaussian clusters only, meaning that $\forall k \in \{1, \dots, K\}, \boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}_D$, the sum of weighted covariances to update $\boldsymbol{\Sigma}_k$ simply reduces to the one of L_2 norms $\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$.

3.4 Phase transitions in Gaussian mixtures

3.4.1 Statistical physics formulation of clustering

In the context of clustering, it has been demonstrated that the GMM can be formulated as a statistical mechanics problem [Rose et al., 1990; Akaho & Kappen, 2000] where the negative log-likelihood can be interpreted as a free energy. We review here the main steps establishing this analogy that we will exploit in the next sections. Assuming that the clustering aims at identifying a partition $\boldsymbol{Z}$ of the data into K clusters with centres $\{\boldsymbol{\mu}_k\}$, the average energy cost of a given configuration can be written

$$E(\boldsymbol{Z}) = \sum_{i=1}^{N} \sum_{k=1}^{K} p(z_i = k \,|\, \boldsymbol{x}_i, \boldsymbol{\mu}_k) E_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k), \tag{3.18}$$

where $E_k(\boldsymbol{x}_i, \boldsymbol{\theta}_k)$ is the energy cost of the association of $\boldsymbol{x}_i$ to the cluster k. Under a quadratic energy cost $E_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k) = \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$, the direct minimisation of the total energy $E(\boldsymbol{Z})$ leads to the K-means clustering of Eq. (3.4) and associates hard probabilities for the associations, i.e. 0 or 1 by simply attributing to datapoints the closest cluster in the sense of the distance measurement provided by $E_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$. Instead, we can introduce a level of uncertainty through a temperature in the associations by maximising the free energy $F = H - \beta E(\boldsymbol{Z})$ where H is the entropy of the system [Shannon, 1948] and β acts like the inverse temperature of the

analogue physical system. For low temperatures $\beta \to \infty$, we retrieve the hard solution while finite values lead to a fuzzy association of datapoints to clusters.

The distribution $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\mu})$ maximising the entropy under the constraint of Eq. (3.18) is, by invoking the principle of maximum entropy [Jaynes & Rosenkrantz, 1983], the Boltzmann distribution, hence reading

$$p(z_i = k \mid \boldsymbol{x}_i, \boldsymbol{\mu}_k) = \frac{\exp\{-\beta E_k(\boldsymbol{x}_i, \boldsymbol{\theta}_k)\}}{\xi(\boldsymbol{x}_i)}, \tag{3.19}$$

where $\xi(\boldsymbol{x}_i) = \sum_{k=1}^{K} \exp\{-\beta E_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\}$ is the individual partition function. For non-interacting systems, the partition function is hence $\Xi = \prod_{i=1}^{N} \xi(\boldsymbol{x}_i)$ (usually denoted Z but that we modified to avoid confusion with the partition of the dataset). The free energy can hence be equivalently written

$$\begin{aligned} F &= -\frac{1}{\beta} \log \Xi, \\ &= -\frac{1}{\beta} \sum_{i=1}^{N} \log \sum_{k=1}^{K} \exp\{-\beta E_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\}. \end{aligned} \tag{3.20}$$

Assuming a quadratic energy cost for $E_k(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$, we recognise the negative log-likelihood from Eq. (3.7) under the spherical and equal variances and uniform amplitudes assumptions, meaning that $\forall k \in \{1, \dots, K\}, \boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_D$ and $\pi_k = 1/K$. As such, this formulation of the clustering is equivalent to the one of a mixture model with a set of parameters reduced to $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$. In the model, the clusters variance is thus playing the role of the temperature of the system $1/\beta = 2\sigma^2$. EM update equations, thanks to this analogy, can be used to minimise the free energy F. The Boltzmann distribution of Eq. (3.19) is exactly the responsibility obtained in the E-step

$$p_{ik} = p(z_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_k) = \frac{\exp\{-\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2 / 2\sigma^2\}}{\sum_{j=1}^{K} \exp\{-\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_2^2 / 2\sigma^2\}}. \tag{3.21}$$

The M-step then refines the position of cluster's centres based on the current values of p_{ik} to minimise F as

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} p_{ik} \boldsymbol{x}_i}{N_k} \tag{3.22}$$

where $N_k = \sum_{i=1}^{N} p_{ik}$ stands for the number of datapoints associated with the cluster k and with parameters not indexed by iteration are relative to time (t). By applying iteratively Eq. (3.21) and Eq. (3.22), one gets the set of positions maximising the log-likelihood, or negative free energy, of the model. In the context of clustering, it means that one hence gets a set of K positions $\{\boldsymbol{\mu}_k\}$ and variances $\{\sigma_k\}$ corresponding to the center and extensions of the K clusters. If a new datapoint is given and one wants to predict its belonging, Eq. (3.21) provides exactly the probability of it having been generated by the several detected clusters under the assumptions of the model.

3.4.2 From paramagnetic to condensation phase

At very high values of σ^2 (high temperatures), the responsibilities (3.21) are uniformly distributed with $\forall i \in \{1, \dots, N\}, p_{ik} = 1/K$. A given datapoint $\boldsymbol{x}_i$ has hence an equal probability to be attributed to any of the K clusters. Injecting this result in the update equation of cluster

positions (3.22) yields $\boldsymbol{\mu}_k = \sum_{i=1}^N \boldsymbol{x}_i/N$ meaning that they are all collapsed at the center of mass of the dataset. This phase is called the paramagnetic phase as an analogy to the Ising model phases, and is, for instance, illustrated in the top left panel of Fig. 3.4.

We are now interested in deriving the transition value of the temperature σ_c^2 at which the centres split into several subgroups. Without any loss of generality, the dataset is considered centred, with $\sum_{i=1}^N \boldsymbol{x}_i = \mathbf{0}_D$ where $\mathbf{0}_D$ is the D-dimensional zero vector. Linearly Taylor expanding the expression of p_{ik} given by Eq. (3.21) for small perturbations $\boldsymbol{\mu}_k \approx \mathbf{0}_D$ gives

$$p_{ik} = \frac{1}{K}\left[1 + \frac{1}{\sigma_k^2}\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\mu}_k\right] + o(\boldsymbol{\mu}_k). \tag{3.23}$$

Re-writing the update equation with this expansion leads to study the stability of a dynamical linear system whose evolution is governed by the matrix [Rose et al., 1990]

$$\boldsymbol{M} = \frac{1}{\sigma_c^2}\left[\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}\right], \tag{3.24}$$

where we recognise $\boldsymbol{C} = 1/N\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T}$, the data covariance matrix. Consequently, the system leaves stability and enters the condensation phase when the spectral radius of $\boldsymbol{M}$ is strictly greater than 1, meaning that the stability is governed by the maximum eigenvalue of the $\boldsymbol{M}$. In that phase the position of the means starts to be correlated with the position of the clusters. More specifically, $\{\boldsymbol{\mu}_k\}_{k=1}^K$ move away from the centre of mass when $\sigma_c^2 < \Gamma$, where Γ is the largest eigenvalue of $\boldsymbol{C}$.

Since we are focusing on datasets supposed to be clustered, with multiple groups spanning a D-dimensional space, we can expect to see successive transitions when a Gaussian component of the model is describing the local size of a subgroup in the data. This idea is the one explored in the forthcoming sections in the context of simulated annealing where the temperature is decreased progressively.

3.4.3 Hard annealing

Cascade of phase transitions

The combination of the previously exposed drawback of the EM algorithm, namely the trapping in local maxima of multi-modal likelihoods and the one of clustering with the choice of the number of components to model the data, makes parametric mixture models very sensitive to the initialisation and the choice of hyper-parameters [Ueda & Nakano, 1998]. In that regard, the statistical physics formulation of clustering established in Sect. 3.4.1 helped to overcome these issues by making use of deterministic simulated annealing. In particular, it allows the relaxation of the non-convex optimisation problem by solving it iteratively while the temperature, or equivalently in our our case, the variance of all components σ^2, is controlled and slowly reduced [Kirkpatrick et al., 1983]. The idea behind these approaches is to smooth out the likelihood by starting with a very high variance leading to a concave function. Decreasing it slowly leads to a finer and finer description of the dataset hence resulting in a more complex likelihood function with multiple modes appearing. In cosmological analyses, simulated annealing is for instance used successfully for the optimisation of the energy function in the Bisous model used to extract cosmic filaments from galaxies in Stoica et al. [2004, 2005a,b] and Tempel et al. [2016].

In Sect. 3.4.2, we focused on the range of temperatures σ^2 for which the likelihood is concave and hence for which all of the K components are collapsed into a single location

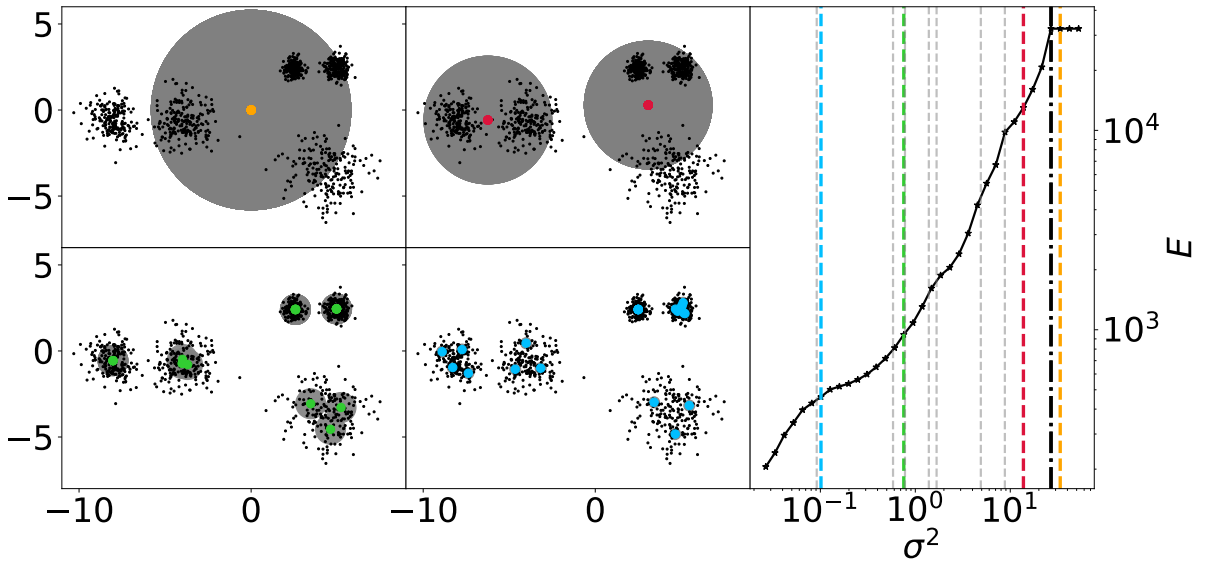

Fig. 3.4. *(left)* Four configurations of the system for different values of the temperature during the annealing procedure. Black points are datapoints, coloured crosses are positions of cluster centres. From top left to bottom right, $\sigma^2 = \{33.59, 13.76, 0.76, 0.10\}$. Grey shaded areas correspond to 1-σ circles. *(right)* Evolution of the average energy $E(\boldsymbol{Z})$ of Eq. (3.18) of the system as a function of the temperature σ^2. The black dashed dotted line corresponds to the critical temperature $T_c^{\text{hard}} = 26.87$ while the grey ones illustrate the temperatures of the successive further transitions.

centred at $1/N \sum_{i=1}^{N} \boldsymbol{x}_i$. We shown that the critical quantity above which this behaviour is observed, that we note T_c^{hard}, is the maximum eigenvalue Γ of the data covariance matrix $\boldsymbol{C}$ [see also Rose et al., 1990]. Figure 3.4 displays the centres position for an artificial dataset with five Gaussian clusters at different temperatures. When $\sigma^2 > T_c^{\text{hard}}$ (top left panel), even though $K = 25$ components are used in the model, they are all collapsed as $K_r = 1$ physical cluster at the center of mass of the dataset (here chosen to be 0). When σ^2 becomes slightly smaller than T_c^{hard} (top right panel), the likelihood is deformed [see Ueda & Nakano, 1998, for illustrations and discussions about the likelihood point of view] and centres get aligned with the first principal direction given by the data covariance matrix $\boldsymbol{C}$. When σ^2 continues to decrease, the dataset description becomes more and more detailed and K_r takes increasing values (bottom panels of left part of Fig. 3.4).

These transitions are physically relevant of the underlying structure of the data by occurring at variances related to the local size of the sub-system being represented. We hence argue that it is possible to extract information on the structure of a dataset from the clustering during the annealing procedure. This can be achieved by tracking the evolution of the physical size represented by a given component k through the maximum eigenvalue Γ_k of its weighted covariance matrix, namely $\boldsymbol{\Sigma}_k = 1/N_k \sum_{i=1}^{N} p_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$, that is the quantity governing the successive transitions.

The right panel of Fig. 3.5 illustrates the evolution of the ratio Γ_k/σ^2 during the annealing for the same artificial dataset as Fig. 3.4 coloured by their end-point cluster. In the left panel of Fig. 3.5, we see the cascade of transitions and successive splitting of centres when σ^2 decreases. When two or more centres collapse, they share similar values of $\boldsymbol{\mu}_k$ and p_{ik} leading to similar evaluations of Γ_k. This is why all lines are superimposed for $\sigma^2 > T_c^{\text{hard}}$ in the right panel. Each time a curve reaches the horizontal unit line, one of the K_r sub-systems made of collapsed

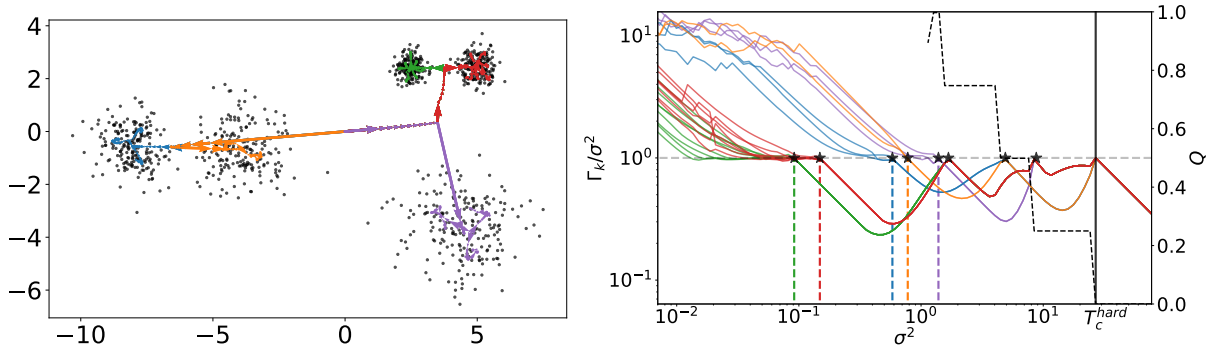

Fig. 3.5. *(left)* Displacement of $K = 25$ centres during the annealing procedure for a dataset with five spherical Gaussian clusters. Colours indicate in which final cluster the center ends. *(right)* Evolution of the ratio Γ_k/σ^2 as a function of σ^2. Black stars correspond to the scales of successive transitions, the black vertical line to $T_{\rm c}^{\rm hard}$ and coloured ones indicate the size of the clusters as defined by the maximum eigenvalue of the empirical covariance. The black dashed curve shows to the evolution of Q as defined in Eq. (3.25) that we identify as an order parameter. This quantity is not represented for $\sigma \leq 1$ since the number of physical clusters K_r begins to be higher than the number of generated clusters q.

components reached the temperature of the sub-dataset it represents, namely $\sigma^2 \simeq \Gamma_k$. From there, we observe either a bounce or a cross of the line. When bouncing, there is a split between two populations of centres that were representing the same part of the dataset but that will, from now on, take different paths. Centres thus move towards a smaller cluster and the value of Γ_k decreases. Crossing the line occurs when centres split inside an individual cluster due to its inner random structure. In that case, the imposed variance gets smaller than the physical one. Since the transitions are ruled by the maximum eigenvalues of the empirical covariance matrices, it is this quantity that is used to measure the size of spherical clusters and plotted as vertical lines on the figures. Note however that if we use instead the empirical variance $\hat{\sigma}_k^2$ as a measure of the cluster sizes, transitions would be shifted since $\hat{\sigma}_k^2 = \mathrm{Tr}\{\mathbf{\Sigma}_k\}/D$ is the average of all eigenvalues of $\mathbf{\Sigma}_k$. In this case, transitions would thereby occur at a different temperature in the annealing than the estimated variances of the cluster because they are driven by Γ_k.

Following *a posteriori* the several curves and the successive transitions in the right panel of Fig. 3.5 provides an informative insight on the structure of the dataset. In particular, it allows to visualise the evolution of the local size representation of the data and the interactions between centres. For instance, we clearly see that the $\{\mathrm{purple}, \mathrm{red}, \mathrm{green}\}$ sets of centres represent the same information when $\sigma^2 > 9$ and then split into $\{\mathrm{purple}\}$ and $\{\mathrm{red}, \mathrm{green}\}$. This indicates the presence of a sub-system made of two clusters. Later, we observe a crossing of the horizontal line for the $\{\mathrm{purple}\}$ centres before splitting again after crossing. This indicates that the effective variance of the cluster is larger than the one fixed by the annealing and, therefore, that these centres now describe fluctuations within a "true" cluster. The $\{\mathrm{red}, \mathrm{green}\}$ sets of centres split at lower σ^2 followed as well by a crossing of the line at the scale of individual cluster sizes (indicated as coloured dashed vertical lines on the figure).

Successive transitions can be computed in two steps: first by identifying the K_r macro-components resulting from the collapse of centres based on their positions and then by assigning to each datapoint (black dots in the left panel of Fig. 3.5) the label of the macro-component that most probably generated it. We are thus assuming, at a given iteration, a GMM with K_r components to compute responsibilities from Eq. (3.21). Thereby, we can group together

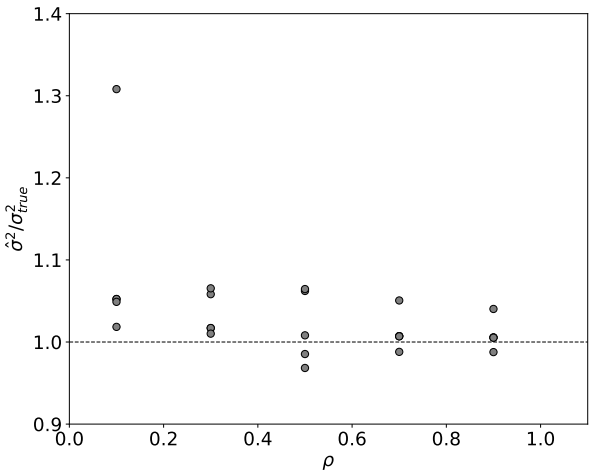

Fig. 3.6. Ratio between the estimated variances $\hat{\sigma}^2$ obtained when freezing the $K = 25$ centres when $\Gamma_k/\sigma^2 \simeq 1$ and the empirical ones σ^2_{true} from data of Fig. 3.5 for several values of ρ, the fraction of datapoints kept for the computation.

datapoints with identical labels and compute the next transition as the maximum eigenvalue of the covariance matrix for each sub-system. These quantities, basically corresponding to successive evaluations of the critical temperature in sub-systems, can be approximated during the annealing and are shown as black stars in the right panel of Fig. 3.5.

We identify the overlap Q, defined as the quality of the data classification at each temperature, as an order parameter whose value changes throughout the several phases during the annealing. Formally,

$$Q\left(\{\hat{z}_i\},\{z_i\}\right) = \frac{\max_{\pi\in\Pi}\sum_i \delta^{\mathrm{K}}_{\hat{z}_i,\pi(z_i)}/N - 1/q}{1-1/q}, \tag{3.25}$$

where δ^{K} is the Kronecker delta, Π denotes all the possible permutations of the set $\{1,\dots,q\}$ with q the true number of clusters used to generate the data and $\hat{z}_i = \mathrm{argmax}_k\, p_{ik}$ the estimated latent variable for the affiliation of the datapoint $\boldsymbol{x}_i$. Q is based on the responsibilities p_{ik}, all being $1/K$ for a random assignation giving $Q = 0$ and taking value 1 when $\forall i, \hat{z}_i = z_i$. By doing so, Q is zero when $\sigma^2 > T_{\mathrm{c}}^{\mathrm{hard}}$ and undergoes successive transitions as the system is cooled down, as illustrated by the black dashed curve of the left panel in Fig. 3.5. During the annealing, Q remains at 1 for the range of σ^2 between the last split of centres between two true clusters and before the first split due to the inner random structure of the largest of them. Note that this metric is meaningless when $K_r > q$ since the dataset is partitioned into more clusters than actually used for the generation, and this is why the curve is not shown for $\sigma^2 \leq 1$.

To further assess the robustness and accuracy of the transitions with the density of input points, we use the dataset from Fig. 3.5 where only a fraction ρ of the datapoints is randomly kept for the computation. During the annealing, we freeze all the $K = 25$ centres at the last split before reaching $\Gamma_k/\sigma^2 = 1$ and then let variances evolve freely hence providing an estimate for each detected cluster that we note $\hat{\sigma}^2$. Figure 3.6 shows that the retrieved variances are, even in highly sparse sampling settings, with $\rho \leq 30\%$, close to the true ones of the clusters. It is worth emphasising that all five clusters are always correctly identified and that the value of Q is always close to 1 at the end of the process, showing the ability of the method to highlight structures, even in sparse configurations.

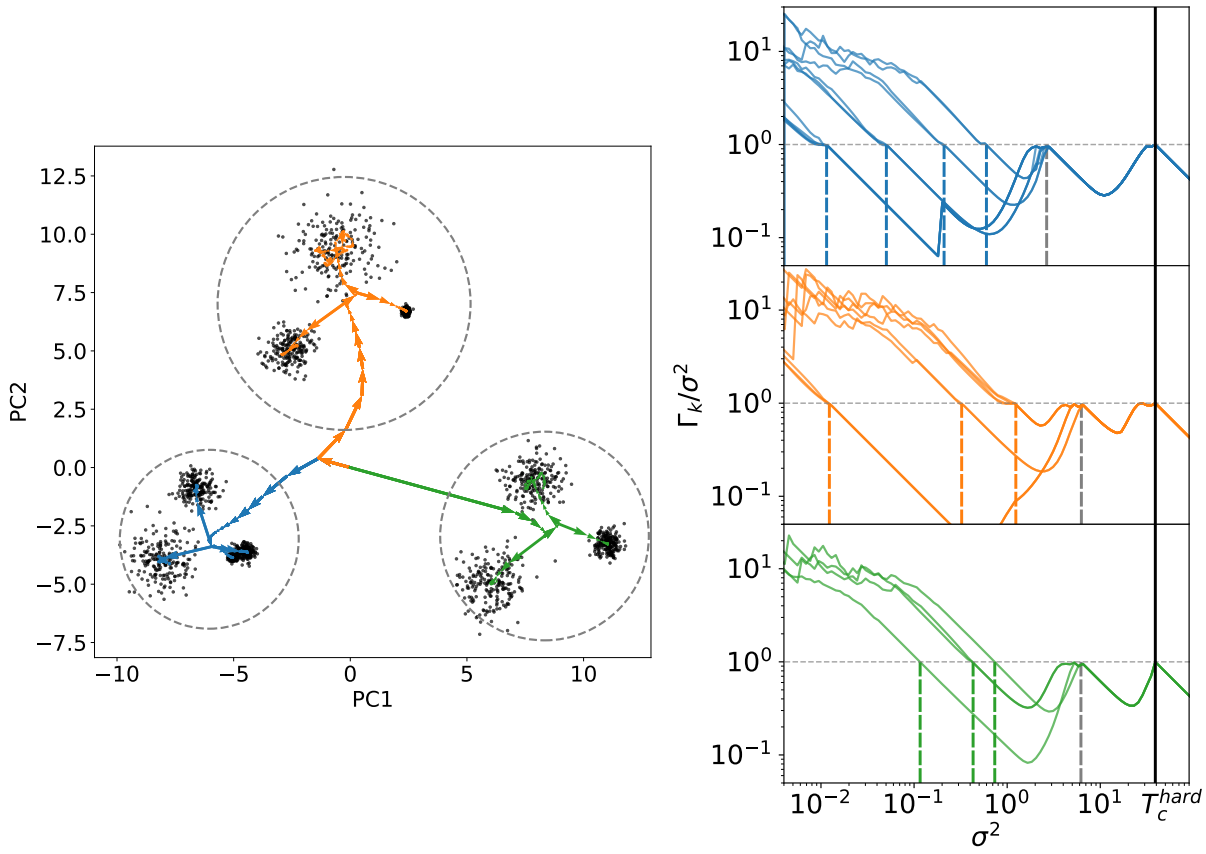

Fig. 3.7. *(left)* Displacement of $K = 25$ centres during the annealing procedure for a dataset made of ten 5D spherical Gaussian clusters visualised in the plane of the two first principal components. Colours depend on the macro-cluster the component stands in at the last iteration. *(right)* Evolution of the ratio Γ_k/σ^2 as a function of σ^2, the hard annealing parameter. Coloured vertical lines indicate the actual size of the corresponding Gaussian cluster or macro-cluster (grey dashed lines, first from the right in each panel) as defined by the maximum eigenvalue of the empirical covariance.

Higher dimensions and nested representations

As exposed in Sect. 3.1.3, usual applications of GMMs for clustering are performed blindly by inputting the desired K to obtain a classification making use of all components. Some criteria, based on information theory [Akaike, 1974; Oliver et al., 1996] or Bayesian approaches [Schwarz, 1978; Roeder & Wasserman, 1997], were proposed to overcome this major drawback of unsupervised clustering. Here, we propose an approach to avoid such a K-dependent unique solution. A key aspect of the annealing is the collapse of $\{\boldsymbol{\mu}_k\}_{k=1}^K$ at the center of mass of successive sub-datasets providing a hierarchical view of clustering with an increasing number of physical clusters. The proposed 2D diagram enables to capture this set of nested representations by catching the several transitions, even if the input space is of high dimensions. Figure 3.7 represents an application for a 5D artificial dataset made of ten Gaussian clusters, spatially appearing as three clusters at larger scale. Transitions occurring at large scales in each panel (grey vertical dashed lines) clearly indicate that the dataset is described as three different physical clusters. Pursuing the decrease of variances leads to a finer description where each of the three macro-clusters splits into smaller ones that still have physical

interpretations[5]. Such information on the spatial organisation of the dataset are of crucial importance when having no prior idea of its structure nor the number of underlying components. The method hence allows a hierarchical view of clustering at different scales as proposed by hierarchical clustering methods [Murtagh & Contreras, 2012]. Because sometimes there is not a unique solution for the number of clusters depending on what they physically represent for the application at hand, the proposed way to explore the data can be of interest before running any kind of blind clustering algorithm.

3.4.4 Soft annealing

When the dataset is more complex with nested structures or overlapping clusters of different sizes, the previously presented analysis is not suitable. Multiple scales cannot be represented at the same time, hence biasing those of embedded structures towards higher values. To overcome this, we relax the equal variance assumption considered until now for the GMM, the parameter set including thereby both the positions and variances of components $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$, with $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \sigma_k)$. Under these circumstances, EM update equations are

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^N p_{ik} \boldsymbol{x}_i}{N_k}, \tag{3.26}$$

$$\sigma_k^{(t+1)} = \left[\frac{\sum_{i=1}^N p_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2}{D N_k} \right]^{1/2}. \tag{3.27}$$

Keeping the idea of a temperature being slowly reduced, we propose a modified annealing acting on the mode of an *a priori* distribution on each σ_k^2 allowing the representation of multiple size in the same time. In particular and to obtain a closed-form expression of the posterior distribution, we use the conjugate prior for variances, namely an inverse-Gamma distribution, with shape parameter $1 + \lambda_\sigma$ and scale parameter $\lambda_\sigma \sigma^2$ so that the distribution has a mode at the position σ^2. Formally, it reads

$$\ln p(\sigma_k^2) = -\lambda_\sigma \left[\ln \sigma_k^2 + \frac{\sigma^2}{\sigma_k^2} \right] + \text{cst}, \tag{3.28}$$

where the constant comes from the normalisation of the probability distribution. Introducing such a prior modifies the update Eq. (3.27) for variances as

$$\sigma_k^2 = \frac{\sum_{i=1}^N p_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2 + 4\lambda_\sigma \sigma^2}{D \sum_{i=1}^N p_{ik} + 4\lambda_\sigma}. \tag{3.29}$$

Consequently, when $\lambda_\sigma \to 0$, the prior, and hence the annealing, has no effect and components update their variances as Eq. (3.27). Inversely, for a large enough value of λ_σ, σ_k^2 will be close to σ^2 resulting in the classical hard annealing procedure discussed in Sect. 3.4.3. Choosing intermediate values for λ_σ imposes a broad trend for all components but lets each of them correct the prior by the actual value of the neighbouring covariance. In what follows, we refer to this procedure as "soft annealing" that we distinguish from the "hard annealing" to describe the classical procedure acting directly on the variance parameter. There is no general rule to fix the hyper-parameter λ_σ and, in this work, we adopt[6] $\lambda_\sigma = 2$.

[5] The separation in three panels with different colours is for visualisation purposes. No prior knowledge was used in the analysis.

[6] In our experiments, keeping $\lambda_\sigma \approx O(1)$ did not change the results quantitatively.

Similarly as in the hard annealing case (see Sect. 3.4.2), we can compute the threshold value $T_{\mathrm{c}}^{\mathrm{soft}}$ such that all components are collapsed at the centre of mass when $\sigma^2 > T_{\mathrm{c}}^{\mathrm{soft}}$. In the context of soft annealing, we however deal with two sets of parameters that are the position of Gaussian components and their associated variances which leads us to study the stability of the joint perturbations over the two sets. Still considering a centred dataset, we propose in a first place to derive the fixed-point variance σ_0^2 of all centres when $\sigma^2 \gg T_{\mathrm{c}}^{\mathrm{soft}}$ assuming that $\sigma^2 > T_{\mathrm{c}}^{\mathrm{soft}} \implies \forall k \in \{1,\dots,K\}, \sigma_k^2 = \sigma_0^2$. Injecting the linearly Taylor expanded expression of responsibilities (3.23) in the update equation (3.29) gives

$$\sigma_0^2 = \frac{4\lambda_\sigma K\sigma^2 + \sum_{i=1}^N \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_i}{ND + 4\lambda_\sigma K}. \tag{3.30}$$

This equation links σ_0^2, the actual variance attributed to all Gaussian components, to σ^2, the inverse of the annealing temperature and is valid in the large σ^2 limit. Taking now into account both parameters hence considering perturbations $\boldsymbol{\epsilon}_k$ and δ_k, respectively around the fixed points $\boldsymbol{\mu}_k = \mathbf{0}_D$ and $\sigma_k^2 = \sigma_0^2$, we can derive the set of equations for the vectorised perturbations $\underline{\underline{\boldsymbol{\epsilon}}} = \left(\boldsymbol{\epsilon}_{\mathbf{0}}^\mathsf{T},\dots,\boldsymbol{\epsilon}_{\boldsymbol{K}}^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{KD\times 1}$ and $\boldsymbol{\delta} = (\delta_0,\dots,\delta_K) \in \mathbb{R}^{K\times 1}$. Responsibilities can be Taylor expanded as

$$p_{ik} \simeq \frac{1}{K}\left[1 + \frac{1}{\sigma_0^2}\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\epsilon}_k + \frac{\delta_k}{2\sigma_0^4}\|\boldsymbol{x}_i\|_2^2 - \frac{1}{K\sigma_0^2}\boldsymbol{x}_i^\mathsf{T}\sum_l \boldsymbol{\epsilon}_{\boldsymbol{l}} - \frac{\|\boldsymbol{x}_i\|_2^2}{2K\sigma_0^4}\sum_l \delta_l\right], \tag{3.31}$$

which, when combined with the update equations (3.26) and (3.27) lead to the system of perturbations in positions and variances

$$\begin{cases} \underline{\underline{\boldsymbol{\epsilon}}}^{(t+1)} = \dfrac{1}{\sigma_0^2}\left(\boldsymbol{U}\otimes\boldsymbol{C}\right)\underline{\underline{\boldsymbol{\epsilon}}}^{(t)} + \left(\boldsymbol{U}\otimes\boldsymbol{a}\right)^\mathsf{T}\boldsymbol{\delta}^{(t)}, \\ \boldsymbol{\delta}^{(t+1)} = \left(\boldsymbol{U}\otimes\boldsymbol{b}\right)\underline{\underline{\boldsymbol{\epsilon}}}^{(t)} + c\,\boldsymbol{U}\boldsymbol{\delta}^{(t)}, \end{cases} \tag{3.32}$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{U} = \left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{J}_K\right)$ with $\boldsymbol{I}_K$ is the $K\times K$ identity matrix, $\boldsymbol{J}_K$ the $K\times K$ all-ones matrix, and

$$\boldsymbol{a} = \sum_i \frac{\|\boldsymbol{x}_i\|_2^2\boldsymbol{x}_i^\mathsf{T}}{2N\sigma_0^4}, \tag{3.33}$$

$$\boldsymbol{b} = \sum_i \frac{\|\boldsymbol{x}_i\|_2^2\boldsymbol{x}_i^\mathsf{T}}{m\sigma_0^2}, \tag{3.34}$$

$$c = \frac{1}{2\sigma_0^2 m}\left(\sum_i \frac{\|\boldsymbol{x}_i\|_2^4}{\sigma_0^2} - D\sum_i \|\boldsymbol{x}_i\|_2^2\right), \tag{3.35}$$

with $m = ND + 4\lambda_\sigma K$. Putting it all together leads to the matrix representation of the system perturbations $\underline{\underline{\boldsymbol{\eta}}} = \left(\underline{\underline{\boldsymbol{\epsilon}}}, \boldsymbol{\delta}\right) \in \mathbb{R}^{K(D+1)\times 1}$

$$\underline{\underline{\boldsymbol{\eta}}}^{(t+1)} = \left(\boldsymbol{U}\otimes\boldsymbol{M}\right)\underline{\underline{\boldsymbol{\eta}}}^{(t)}, \tag{3.36}$$

with $\boldsymbol{M}$ the squared block matrix of order $D+1$

$$\boldsymbol{M} = \left(\begin{array}{c|c} \boldsymbol{C}/\sigma_0^2 & \boldsymbol{a}^\mathsf{T} \\ \hline \boldsymbol{b} & c \end{array}\right), \tag{3.37}$$

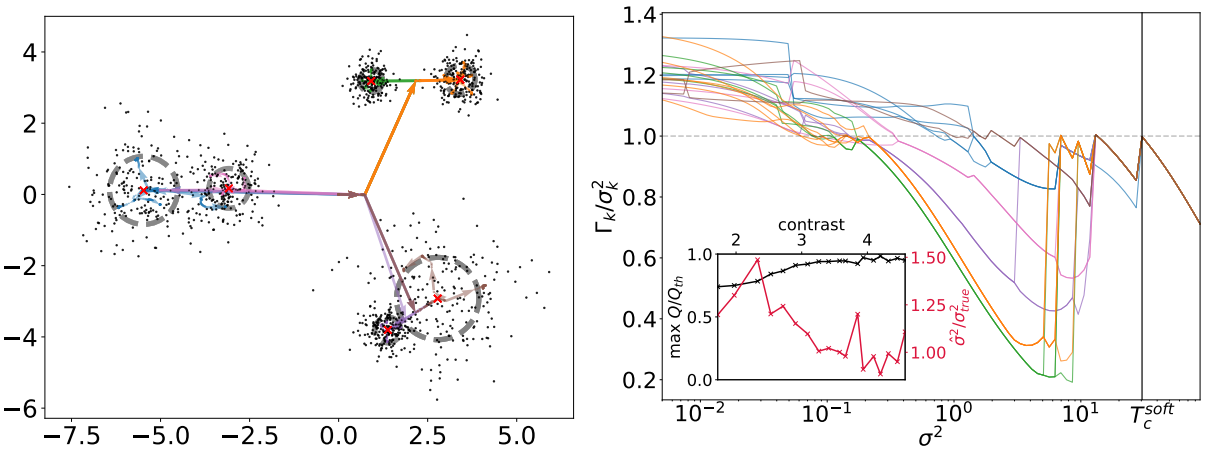

Fig. 3.8. *(left)* Arrows indicate the displacement of $K = 25$ centres during the soft annealing procedure for a dataset made of six spherical Gaussian clusters (black points). Colours relate to the cluster in which the component ends. Red crosses and grey dashed circles respectively indicate the positions and variances fixed *a posteriori* when the center undergoes its last split before remaining above the $\Gamma_k/\sigma_k^2 = 1$ line. *(right)* Evolution of the ratio Γ_k/σ_k^2 as a function of σ^2. The vertical black line corresponds to T_c^{soft}. The inset figure shows the evolution of the ratios $\max Q/Q_{\text{th}}$ and $\hat{\sigma}^2/\sigma^2_{\text{true}}$, when varying the contrast between the two nested clusters.

where $\boldsymbol{C}$ is the data covariance matrix. Since the eigenvalues of the Kronecker product are given by the product of all individual eigenvalues of the two matrices involved and that $\boldsymbol{U}$ has eigenvalues 0 or 1, we can only restrict the analysis to those of $\boldsymbol{M}$. Therefore, the value of σ^2 at which the first transition occurs, namely T_c^{soft}, can be derived as the value of σ^2 such that the spectral radius of $\boldsymbol{M}$ is 1, leading to instabilities in the dynamic of the system (3.32).

The left panel of Fig. 3.8 illustrates the result of the soft annealing procedure on an artificial dataset made of six clusters similar to Fig. 3.5 but with more complexity such as overlapping (the two clusters on the left) and nested clusters (the two bottom right clusters). The right panel focuses on the evolution of the ratio between the size of the represented sub-system by a given component and its actual variance, namely Γ_k/σ_k^2. This is the same physical quantity as in the hard annealing case, except that we relax the constraint on σ_k^2 which is now varying for each component. In this soft configuration, all the $\boldsymbol{\mu}_k$ are collapsed for $\sigma^2 > T_c^{\text{soft}}$ followed by steep transitions when σ^2 decreases. This relaxed annealing is especially useful for the representation of the two nested clusters. Even though we would learn that those structures are encapsulated, reaching an accurate size description for the smallest nested component would not be possible in hard annealing because its variance would be boosted by neighbouring datapoints of the surrounding cluster. The left panel of Fig. 3.8 illustrates positions (red crosses) and variances (grey dashed circles) of all $K = 25$ components when fixing parameters *a posteriori* at the value they had during the annealing at their last transition point just before crossing the line $\Gamma_k/\sigma_k^2 = 1$. Although $K = 25$ components are used, we correctly identify $K_r = 7$ physical clusters with their variances and means.

To assess the robustness of the soft annealing procedure in clustering complex datasets, we focus on a setup restricted to the two nested clusters of Fig. 3.8 in the bottom right part of the left panel. We dilute the small one by varying its number of sampling points N. This translates into a decreasing contrast between the signal to noise ratios $\sigma/\sqrt{N}$ of the two clusters. The inset of the right panel of Fig. 3.8 shows the evolution of the ratio between the maximum overlap value Q obtained during the annealing and Q_{th}, the theoretical overlap computed us-

ing the ground truth parameters as a function of the contrast. It can be seen that both clusters are recovered when the contrast is sufficiently high (above 1.5 in practice) while below, there is not the necessary information for the model to retrieve it. The ratio between the estimated variances is also shown to and we observe an overestimated variance at lower and lower contrast which explains the decreasing $\max Q/Q_{\text{th}}$ ratio. This effect can partly be explained by the uniform weights hypothesis being less and less true when the contrast decreases.

3.4.5 Graph-regularised mixture model

In Chapter 4, we shall introduce a method to learn a smooth graph representation from point-cloud distributions. For the purpose of studying the transitions of the model, we set ourselves in a simpler context than the one presented in Chapter 4. We hence consider that the learning of the graph structure is done through a regularised mixture model with fixed-variance. The regularisation term acts on component averages and constrains the graph smoothness through the Laplacian defined as

$$\|\boldsymbol{\mu}\|_{\mathcal{G}}^2 = \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ij}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_k\|_2^2, \tag{3.38}$$

where a_{ij} is an element of $\boldsymbol{A}$, the adjacency matrix taking value 1 when centres i and j are linked and 0 otherwise. This added term on the log-likelihood acts as an attractive quadratic interactions of centres connected on the graph $\boldsymbol{A}$. As for the soft annealing case, the introduction of a prior in EM equations only impacts the M-step update of center positions of Eq. (3.26) as

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p_{ik}/\sigma^2 + 2\lambda_\mu \sum_{j=1}^{K} a_{kj}\boldsymbol{\mu}_j^{(t+1)}}{\sum_{i=1}^{N} p_{ik}/\sigma^2 + 2\lambda_\mu \sum_{j=1}^{K} a_{kj}}. \tag{3.39}$$

As done for previous iterative learning models, it is possible to compute the value of σ^2 for which the high-temperature system becomes unstable, noted $T_{\text{c}}^{\text{graph}}$ considering perturbations around the fixed point $\boldsymbol{\mu}_k = \mathbf{0}_D$. Analogous derivations as in Sect. 3.4.4 shows that the system is unstable when the maximum eigenvalue of $\boldsymbol{M}$ is greater than 1, with

$$\boldsymbol{M} = \left[\left(\boldsymbol{I}_K - \frac{1}{K}\boldsymbol{J}_K\right) \otimes \boldsymbol{C}\right] \left[\sigma^2 \boldsymbol{I}_{KD} + \frac{2\lambda_\mu K \sigma^4}{N} \boldsymbol{L} \otimes \boldsymbol{I}_D\right]^{-1}, \tag{3.40}$$

where $\boldsymbol{L}$ the Laplacian matrix defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ with $\boldsymbol{D}$ the diagonal $K \times K$ degree matrix with $d_{kk} = \sum_i a_{ik}$.

The graph prior in Eq. (3.40) is expressed through $\boldsymbol{L}$ and, in a first place, we propose to solve analytically the simple case of a complete graph prior in which a node is connected to all other nodes. For more general cases, the threshold depends on the form of the Laplacian and is computed numerically depending on the graph and the data at hand. Using a complete graph prior, the temperature can be computed with $\boldsymbol{A} = \boldsymbol{J}_K$ and $\boldsymbol{D} = (K-1)\,\boldsymbol{I}_K$, leading to solve a second degree equation in $T_{\text{c}}^{\text{graph}}$ with a unique non-negative solution for $\lambda_\mu > 0$ given by

$$T_{\text{c}}^{\text{graph}} = \frac{-1 + \sqrt{1 + 8\Gamma\lambda_\mu K^2/N}}{4\lambda_\mu K^2/N}. \tag{3.41}$$

This simple case shows how the critical temperature varies with λ_μ. The limit $\lambda_\mu \to 0$ and Eq. (3.40) with $\lambda_\mu = 0$ lead to $T_{\text{c}}^{\text{graph}} = \Gamma$, the absence of regularisation being equivalent to the hard annealing case. When λ_μ increases, the threshold is shifted towards lower temperatures,

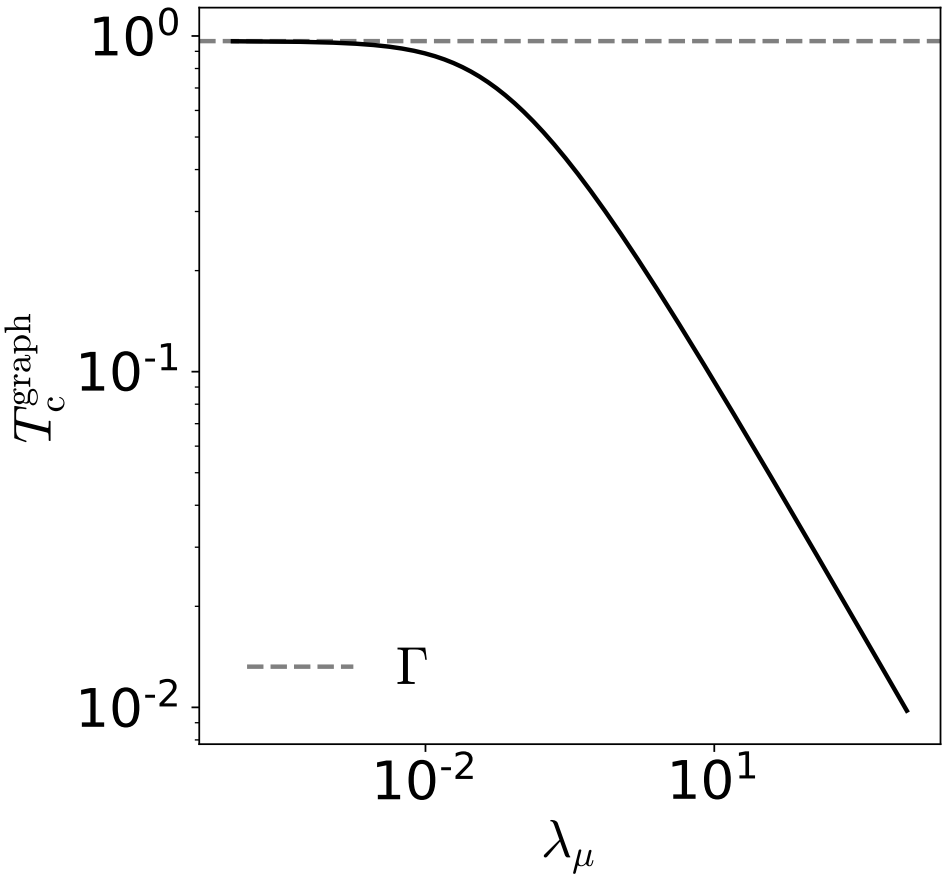

Fig. 3.9. Evolution of T_c^{graph} as a function of λ_μ in case of a complete graph prior given by Eq. (3.41). The grey dashed horizontal line represents Γ, the maximum eigenvalue of the data covariance matrix $\boldsymbol{C}$.

as illustrated in Fig. 3.9 showing the evolution of T_c^{graph} given by the relation (3.41). Physically, since the prior is basically pulling adjacent centres, increasing λ_μ increases the strength of the bonds between nodes of the graph.

In Fig. 3.10 is shown the result of an annealing procedure for $K = 100$ components, $\lambda_\mu = 300$ and using a graph prior given by the minimum spanning tree construction [Borůvka, 1926] assuming that centres are all linked together with the minimum total length (see Sect. 4.2.3 for a more detailed presentation of graph constructions). In this case, the graph is computed from the random set of initial nodes and is updated at each iteration of the EM procedure when the temperature of the annealing is below the critical one, $\sigma^2 < T_c^{\text{graph}}$. When paving a continuously structured data distribution with Gaussian clusters standing on a prior graph structure as proposed by the regularised mixture model, the scale of interest is the local width of the elongated structure. This size is given locally, in our 2D case, by the minimum eigenvalue γ_k of the weighted covariance $\boldsymbol{\Sigma}_k$ that we now follow during the transitions As predicted by linear stability, centres are first aligned with the principal axis of the dataset at the beginning of the annealing and then spread over the structure to pave it more precisely, as shown on the top left panel of Fig. 3.10. By tracking the evolution of the ratio γ_k/σ^2 in the bottom panel, we clearly distinguish four types of behaviours, signature of four distinct scales for structures in the dataset. It is also interesting to observe the absence of sharp phase transitions or splits within this continuous dataset tending to smooth out the evolution of the energy when the temperature is decreasing. By imposing, for each component, the variance σ^2 during

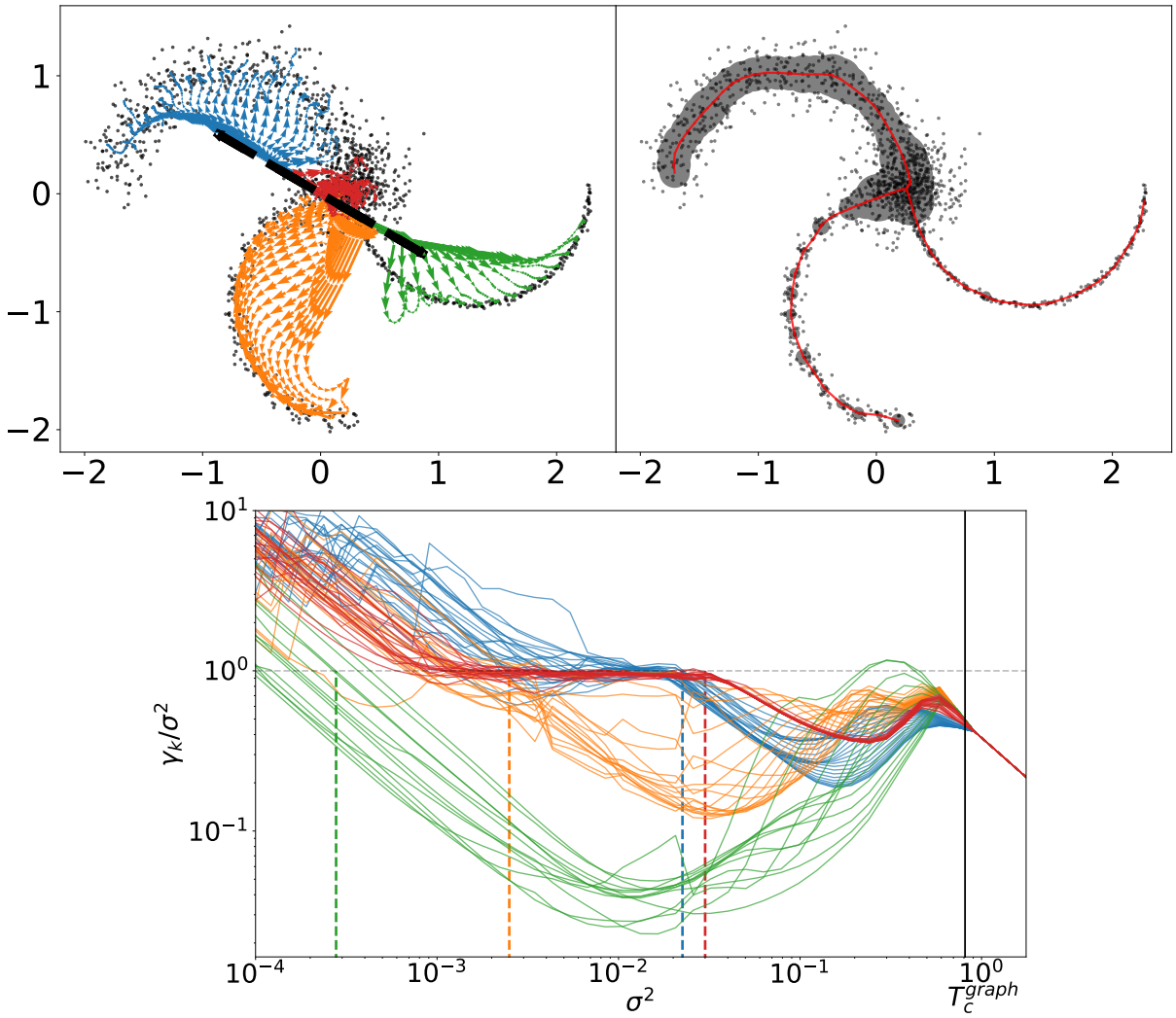

Fig. 3.10. *(top left)* Displacement of $K = 100$ components during the hard annealing of a tree branches dataset with different sampling standard deviations. Black dashed line corresponds to the first principal direction. Colours refer to branches in which centres end. *(top right)* Learned structure when stopping the annealing for components reaching the temperature $\sigma^2 \simeq \gamma_k$. Red lines are edges of the graph and grey shaded areas are 1-σ_k circles. *(bottom)* Evolution of the ratio γ_k/σ^2 as a function of σ^2. Vertical lines indicate the used variance for the generation of branches. The black vertical line corresponds to the value of T_c^{graph} predicted by the linear stability analysis of the model in Eq. (3.40).

the annealing at the moment $\gamma_k \simeq \sigma^2$, we obtain the graph of the top right panel of Fig. 3.10, showing multiple adaptive scales, even though branches have one order of magnitude difference in sampling standard deviation.

3.5 Summary and prospects

In this chapter, we introduced the rich context of machine learning and discussed its applicability and interpretability for physical sciences. We saw that various communities were addressing such questions and that statistical physics was one of the promising lead to do so. After exhibiting some connections between the two fields, we introduced the clustering aim-

ing at identifying, in an unsupervised way, some subgroups in the data. For that purpose, we saw that mixture models can be used together with the Expectation-Maximisation algorithm to estimate parameters of the model with latent variables.

We then established a statistical physics formulation of the clustering procedure and showed that it was equivalent to a particular case of mixture models which can be solved by the EM algorithm. We used this physical analogy to study the learning dynamics of the clustering scheme in the context of deterministic simulated annealing. In particular, we observed and characterised the cascade of phase transitions occurring when tracking the evolution of eigenvalues of the successive covariance matrices of mixture components in multiple cases. We showed that the thresholds at which the first transition occurs can be computed analytically by studying the linear stability of the fixed-point iterative scheme and that we can approximate the next ones during the annealing based on the current estimate of the responsibilities. The proposed way to use these transitions to explore the data and build a hierarchical description independent from K provides a qualitative and quantitative insight on the structure of the dataset at different scales and without requiring any prior knowledge. The 2D diagram representing the evolution of the physical size of the represented clusters with the annealing temperature allows an analysis independent from the data dimensionality and highlights characteristic scales at which physical transitions occur. This diagram carries information about the number of components, their scale and hierarchy that can be used *a posteriori* for data exploration before running blind clustering methods.

Interestingly, we saw that the latent variables of the GMM are directly related to the values taken by the order parameter in Eq. (3.25). Since the GMM can be recast as a particular case of Restricted Boltzmann Machines [RBM, Smolensky, 1986] using a soft-max prior on the hidden nodes, making it a particular type of auto-encoder [Bourlard & Kamp, 1998], the presented work can be seen as the learning of a latent representation of the phase transitions, as is discussed in Van Nieuwenburg et al. [2017] and Wetzel [2017]. Intriguingly, we also witnessed a complex dynamical behaviour in the learning of a simple algorithm of clustering with a restricted number of parameters and putting ourselves in a simplistic setup with uniform and spherical Gaussian mixtures. This says long on the difficulty of handling overparametrised deep neural models in such setups, a vast current topic of research [Bahri et al., 2020]. One of the aspects I personally would be interested to explore is the relation between the data structure and the learning dynamics in more generic distributional learning approaches. As a starting point, the RBM is the perfect candidate since it extends the GMM formulation but it is also a simple version of a neural network with a single hidden layer that aims at learning a statistical representation of the data. The statistical physics analysis of the RBM through spin-glass analogies and mean-field approximations already led to the characterisation of the phase diagram of the model in many cases [see Decelle & Furtlehner, 2021, for a review]. Mixing this more general version of the GMM and the hidden manifold model proposed in Goldt et al. [2020], meant to study the learning of two-layer networks when high-dimensional data are embedded onto a submanifold of smaller dimension, is a promising way to build a framework enabling the understanding of relationship between the data structure and the learning dynamics in unsupervised setups.

Among the different and numerous applications of clustering, the task of segregating datapoints in different classes is a long-standing problem in astrophysics with for instance the galaxy morphological classification (elliptic, spiral, etc.) based on optical data [Weinmann

et al., 2006; Shamir, 2009] or the classification of gas phases based on gas properties (temperature, density, etc.) [see e.g. Martizzi et al., 2019; Galárraga-Espinosa et al., 2021]. By including the identification of hierarchical structures, it may help in defining the classes in a physically motivated way. It could also tackle the problem of detecting subhalos in simulations based on discrete particle positions as an alternative to the SubFind algorithm [Springel et al., 2001; Dolag et al., 2009]. By proposing a non-unique solution to clustering, it also offers the possibility to explore the different relevant substructures. Applications of such a clustering procedure are hence of interest in these contexts to study the link between the different possible physical and spatial structural information of a cluster (may it be through its dark or baryonic matter composition) and how it correlates with those of the neighbouring filamentary structure.

Chapter 4.

Principal graph learning

"Graph theory is a natural language to describe the Cosmic Web."
M. Aragón-Calvo

This chapter is presenting the results from Bonnaire et al. [2021b] and partially those from Bonnaire et al. [2020]. In Chapter 3, we mainly studied some datasets appearing as split into multiple groups that we wished to identify. With the same interest for the spatial structure of datasets, we aim at learning non-linear representations of continuously structured data assuming as standing on one-dimensional underlying manifolds. The goal of this chapter is to expose an alternative formulation of the principal curve problem that overcome some of the drawbacks of the existing definitions. In particular, we focus on establishing a procedure that allows the handling of outliers and variations in the local size of the sampling noise in view of applying it to the detection of the filamentary pattern depicted by matter tracers (halos or galaxies) in Chapter 5. To do so, we first review some basic elements of graph theory and mix them with the mixture model (see Sect. 3.2) framework to propose a formulation in which the

graph is modelling the underlying 1D manifold. After a comparison to other ridge finders, we discuss several technical aspects of the algorithm, like its convergence properties and time complexity before using the detection of roads from noisy GPS measurements as a showcase.

4.1 Context

4.1.1 Spatially structured point-cloud data

Many datasets come as a set of discrete D-dimensional datapoints $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=0,\dots,N}$ with $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^D$ sampled from an unknown probability distribution. These datapoints are usually not spreading uniformly over the entire $\mathbb{R}^D$ space but often result from the sampling of a lower dimensional manifold whose topology and characteristics are linked with the process that generated the data. Taking the example of the famous MNIST dataset, in which datapoints correspond to images of 28×28 pixels digits, making it a 784 dimensional dataset, there is only a very restricted volume in this space that would generate images that actually *look like* a digit [Hein & Audibert, 2005; Facco et al., 2017]. Capturing this information, may it be for visual, geometrical or topological analyses of the dataset requires the application of non-linear methods that are parts of the manifold learning field [Roweis & Saul, 2000; Belkin & Niyogi, 2003; van der Maaten & Geoffrey, 2008; Van Der Maaten et al., 2009]. In some applications, data even appear as standing on a continuous one-dimensional structure. It is for instance the case for GPS measurements collected by vehicles standing on the road network [Ahmed et al., 2015], vessel networks transporting blood through the human body [Moccia et al., 2018] but also for the large-scale matter distribution describing the filamentary structure of the cosmic web, as seen in Chapter 2 illustrated for instance by the left panel of Fig. 2.6. As part of unsupervised machine learning methods (see Sect. 3.1.1 for a discussion on categories of machine learning approaches), one of the key aspects of pattern analysis is to extract from such inputs, with the least prior knowledge, the sufficient information to build a meaningful representation of the data to understand the underlying structure that generated it, build models and make predictions.

4.1.2 Principal curves

The problem of estimating an one-dimensional manifold approximating the underlying distribution of $\boldsymbol{X}$ is a particular case of dimensionality reduction also known as *ridge detection* or *principal curve extraction*. The seminal work of Hastie & Stuetzle [1989] provides an intuitive definition of a principal curve as the line passing in "the middle" of the point cloud distribution hence providing a non-linear generalisation of principal components. More formally, in this early formulation, a principal curve is a self-consistent smooth non-intersecting line with finite length. Self-consistency is maybe the most important property and implies that each point of the curve corresponds to the average of the datapoints projecting right on it, i.e. $f(y) = \mathbb{E}(\boldsymbol{X} \mid y_f(\boldsymbol{X}) = y)$, with $y_f(\boldsymbol{x})$ the projection index of a datapoint $\boldsymbol{x}$ on the curve f (see Fig. 4.1 for an illustration). The self-consistency property is also shared by principal components to which principal curves are a relaxation of the straight line condition and can be seen as what local regression is to linear regression. Exactly as principal components, principal curves can be written as the minimisation of the quadratic sum of projected distances.

In that regard, it can be formulated in the general form (3.1)

$$\underset{f\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{N} \|\boldsymbol{x}_i - f(y_f(\boldsymbol{x}_i))\|_2^2. \tag{4.1}$$

which, under the constraint that $\mathcal{F}$ is the set of linear relations $f(y) = f_0 y$, is solved by the straight line in the direction of the first principal component. If $\mathcal{F}$ is more generally the set of continuous functions, optimisation problem (4.1) defines the principal curve formulation proposed by Hastie & Stuetzle [1989]. Figure 4.1 illustrates the first principal component of a 2D non-linear dataset and a principal curve. Although well-defined for the principal component case, the minimisation of problem (4.1) in its most general form do not lead to "principal curves". A simple way to see that is to imagine the curve passing exactly on every datapoints, hence minimising the total cost but which is not a principal curve. One way to take into account that prior knowledge is to include regularisation terms (see Sect. 3.1.2) to penalise the set $\mathcal{F}$ and constrain the solution f. In Kegl et al. [2000] is for instance proposed an algorithm fixing the overall length of the curve while Smola et al. [2001] propose more general smoothness constraints. Since the work of Hastie & Stuetzle [1989], several studies followed to extend these definitions to higher dimensions (principal surface or volume) or to give more suitable definitions of principal curves as ridges of a probability density [e.g. Tibshirani, 1992; Ozertem & Erdogmus, 2011].

To allow the description of more flexible structures than those imposed by curves and to bypass their inability to represent self-intersecting or cycling topologies, a formulation relying on graph theory to model the one-dimensional structure was introduced in Gorban & Zinovyev [2005] and extended in Gorban & Zinovyev [2009]. This latter is based on predefined rules for growing the graph and includes regularization terms to limit its complexity. However, this model come with a large number of parameters and with no guaranteed convergence to which the double optimization scheme of Mao et al. [2015] offers an alternative for the learning of a tree structure. In this landscape of methods, only a few address the problem of estimating a principal graph with a proper handling of outliers, which considerably complicates the learning of the graph structure. A built-in robustness to outliers is proposed in Gorban et al. [2016] and Albergante et al. [2020] by discarding from the update of a node position all datapoints beyond a robustness radius R_0. However, the choice of R_0 is not trivial, scale-dependent and require a careful tuning to deal with uniform background noise [Albergante et al., 2020].

The aim of the chapter is to combine mixture models presented in Sect. 3.2 to approximate the underlying data distribution and regularise it over a graph structure to constrain Gaussian centroids to pave the approximation of the manifold given by the graph. The method extends the original presentation of Tibshirani [1992] which is making use of smooth differentiable curves to a more general representation given by a graph structure that acts like a topological prior in the Bayesian model turning the problem into a maximum *a posterior* estimation. We will see that the proposed formulation through mixture models naturally allows the learning of the local width of the represented one-dimensional structure with robustness to outliers hence freeing it from heavy pre-processing, as opposed to the previous principal curve or graph algorithms. It also comes with guaranteed convergence to a local maximum of the log-posterior inherited from the Expectation-Maximisation algorithm.

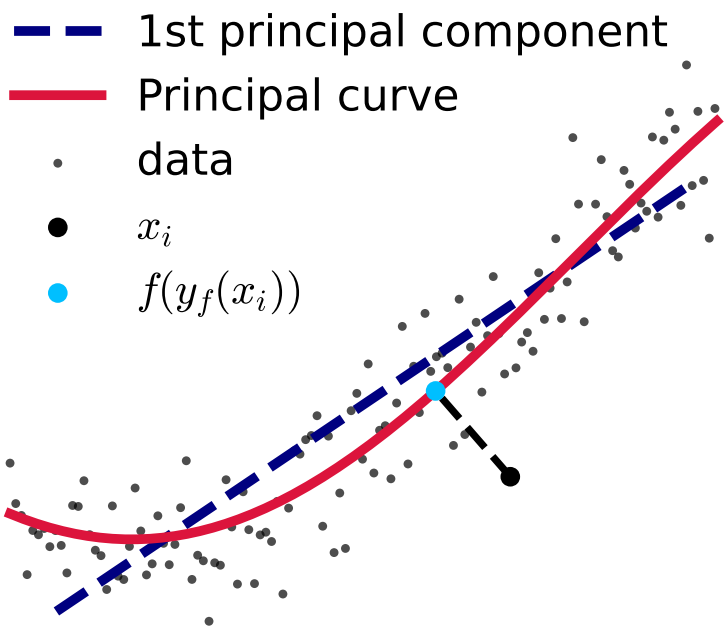

Fig. 4.1. Illustration of a principal curve for a non-linear dataset. The principal curve (in red), minimising the total sum of squared projection distances of all datapoints under some smoothness constraints generalise the straight line principal component analysis (dashed blue line).

4.2 Elements of graph theory

Graph theory is a branch of mathematics that became popular in many fields of science such as social science [Borgatti et al., 2009], biology [Koutrouli et al., 2020] and physics [Estrada, 2013]. For this latter use, it has been exploited in fields ranging from quantum physics for the propagation of waves [Kuchment, 2008; Berkolaiko, 2017] to cosmology for the study of the galaxy distribution [Barrow et al., 1985; Colberg, 2007; Coutinho et al., 2016]. Widely known for their great representative power of relationships between objects, graphs became also a keystone of computational geometry by allowing the representation of non-Euclidean geometries. This section hence aims at providing the required formal background of graph theory that is of importance to establish and understand the method proposed in this manuscript. For a more complete introduction to graph theory, we refer the reader to the book of Bondy & Murty [2008].

4.2.1 Introduction and definitions

In simple terms, a graph is a collection of nodes with edges linking them together. Nodes can represent physical observables, like galaxies in astrophysics, towns and cities in the road network, human beings in a social network or URL addresses in the internet. Edges indicate a physical or conceptual relation between two vertices, linking them if they share some similarities. In the case of a social network, they can be thought as a connection between two individuals if they know each other while, in the road network, two cities can be linked if there is a route to travel from one to the other. Edges usually come with a weight that encodes the similarity measurement or the cost of associating two nodes and depend on the application. Taking back the example of roads, they could be a measure of the geodesic distance between two linked cities. Mathematically, we write $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ is the set of $K \coloneqq |\mathcal{V}|$

nodes, $\mathcal{E} \subset \{1, \dots, K\} \times \{1, \dots, K\}$ is a set of tuples of two vertices that are linked and $\mathcal{W} = \{w_{ij}\}_{(i,j)\in\mathcal{V}^2}$ is the set of edge weights such that $\forall (i,j) \in \mathcal{V}^2, w_{ij} \geq 0$ being non-zero when $(i,j) \in \mathcal{E}$. If $\mathcal{E}$ is an unordered (resp. ordered) set, the graph is said undirected (resp. directed), meaning that the link between two nodes i and j is reciprocal and $w_{ij} = w_{ji}$ (resp. not reciprocal and, in general, $w_{ij} \neq w_{ji}$). In the example of the internet network, a URL address can refer to another one while it might not be the case in the other way, making it a directed network.

All of the graphs discussed in this manuscript are parts of the undirected and simple graphs family, meaning that they exhibit no self-loop (i.e., no node i such that $(i,i) \in \mathcal{E}$) and no multiple edges (i.e., each association $(i,j) \in \mathcal{E}$ is unique in $\mathcal{E}$). Although one can define many characteristics for nodes in a graph, we only restrict ourselves to the introduction of the degree $\deg(i)$, which is the number of individual edges incident with the node indexed i.

Moreover, in all the applications presented in this thesis, graphs are considered as objects embedded in $\mathbb{R}^D$, whose nodes have a spatial position, making them "spatial graphs". As for any other graphs, the weights $\{w_{ij}\}$ associated to nodes pairs are still measuring a similarity but based on their spatial proximity. More precisely, for a graph with K nodes with positions $\{\boldsymbol{\mu}_k\}$, we will consider weights corresponding to the Euclidean distance, $w_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2$ if nodes indexed i and j are linked.

4.2.2 Linear algebra representations

To any graph $\mathcal{G}$ is associated a $K \times K$ matrix $\boldsymbol{A}$ fully encoding the graph information by embedding the relations between nodes. Such a matrix representation is called the adjacency matrix and is particularly useful in linear algebra contexts to have tractable representation of the graph. Elements of $\boldsymbol{A}$ are defined as

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

One can note that summing over the i^{th} row gives the degree of the node i, meaning that $\deg(i) = \sum_{j=1}^{K} a_{ij}$ and that $\sum_i \sum_j a_{ij} = 2|\mathcal{E}|$. In the particular case of simple graphs, $\boldsymbol{A}$ has the additional properties of being symmetric and has a zero-valued diagonal from the absence of loops. If weights are incorporated into the adjacency matrix then $\boldsymbol{W}$ fully characterises the graph $\mathcal{G}$ including node connections and the associated weights.

A second possible matrix representation is provided by the discrete graph Laplacian defined as the symmetric and semi-positive definite matrix

$$\boldsymbol{L} \coloneqq \boldsymbol{D} - \boldsymbol{A}, \tag{4.3}$$

where $\boldsymbol{D}$ is a $K \times K$ diagonal matrix with $d_{ii} = \deg(i)$ and $\boldsymbol{A}$ is the adjacency matrix defined above. This algebraic representation has led to numerous studies and is even at the origin of a subfield of research called "spectral graph theory" [see Chung, 1999; Brouwer & Haemers, 2012, for reviews]. The Laplacian matrix is hence a central operator whose eigenvalues are closely related with the macroscopic properties of the graph. For instance, the smallest eigenvalue of $\boldsymbol{L}$ is 0 and its multiplicity is associated to the number of connected subgraphs in $\mathcal{G}$ [Hagen & Kahng, 1992]. Based on extensions of this result, many efficient clustering algorithms emerged to partition a graph [see e.g. Alpert et al., 1999; Nascimento & De Carvalho, 2011; Saade et al., 2014] or to build projection of high-dimensional dataset in lower dimension [Belkin & Niyogi, 2001, 2003; Hein & Markus, 2007]. These last examples are based on the

established analogy between graphs and manifolds [Chung, 1999] that we will be exploiting in Sect. 4.3.1 to constrain the smoothness of the estimated principal graph in the proposed framework.

4.2.3 Some graph constructions

There are multiple ways to build a graph from a given set of spatial nodes $\boldsymbol{\mu} = \{\boldsymbol{\mu}_i\}_{i=1}^K$. We review below some remarkable graph topologies that are being used throughout this manuscript.

Complete graph. The complete graph is represented by a plain adjacency matrix, $\forall (i,j) \in \{1,\dots,K\}^2, a_{ij} = 1$, meaning that all possible pairs of nodes are linked together. Even though weights can be chosen to stress more local relations based on the proximity of nodes, this graph encodes a lot of redundant paths and do not emphasise a particular one in the data. Moreover, even if dense graphs have interesting properties (like robustness), this construction induces a plain adjacency matrix which, for applications including a large number of nodes (i.e. galaxies for instance), becomes quickly hardly tractable in both time and memory.

ϵ-neighbourhood. Two nodes i and j are linked together if $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 < \epsilon$ (see illustration in the left panel of Fig. 4.2). Despite its natural definition, where close nodes are linked together, one of the main drawbacks of this graph construction is that the represented pattern highly depends on the choice of ϵ that can be arbitrary and scale-dependant.

k-nearest neighbours. The k-neighbourhood (k-nn) of a node i is defined as the set of k nodes that are the closest in terms of the Euclidean norm. In this construction, two nodes i and j are hence linked together if i is in the k-neighbourhood of j or reciprocally. The choice of k is easier than in the ϵ-neighbourhood case but the k-nn has the disadvantage of creating long-range links for datapoints standing in almost-empty areas (as illustrated on the middle panel of Fig. 4.2).

Minimum Spanning Tree. Let us first define a subgraph $\mathcal{H} = (\mathcal{V}_\mathcal{H}, \mathcal{E}_\mathcal{H}, \mathcal{W}_\mathcal{H})$ of an arbitrary graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ as the graph such that $\mathcal{V}_\mathcal{H} \subset \mathcal{V}$ and $\mathcal{E}_\mathcal{H} \subset \mathcal{E}$. We call a spanning tree a particular type of subgraph in which the set of vertices is the same as the original one but with the minimum number of edges. That is, the graph $\mathcal{G}$ admits a spanning tree $\mathcal{T} = (\mathcal{V}, \mathcal{E}_\mathcal{T}, \mathcal{W}_\mathcal{T})$ with $|\mathcal{E}_\mathcal{T}| = |\mathcal{V}| - 1$ linking all nodes of $\mathcal{V}$ together. Building on that knowledge, a minimum spanning tree [MST, Borůvka, 1926] is a spanning tree with the minimum total weighted length $\sum_i \sum_j w_{ij}$. The MST is, by definition, a simple graph and is unique only if all weights are different. This definition can be rephrased in terms of an integer programming problem aiming

at finding the adjacency matrix $\boldsymbol{A}_{\text{MST}}$ such that

$$\boldsymbol{A}_{\text{MST}} = \underset{\boldsymbol{A}}{\operatorname{argmin}} \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ij} w_{ij} \tag{4.4}$$
$$\text{s.t.} \begin{cases} \boldsymbol{A} = \boldsymbol{A}^T, \\ a_{ij} \in \{0,1\}, \\ \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ij} = 2\left(K-1\right), \\ \forall k \in \{1,\dots,K\}, a_{kk} = 0, \\ \forall S \subseteq \mathcal{V}, \sum_{i\in S}\sum_{j\in S} a_{ij} \leq |S| - 1, \end{cases}$$

The first two constraints encode the undirected graph definition, the third one denotes a graph with $K-1$ edges, the fourth the absence of self loops in the structure and the last one requires that any subset of $|\mathcal{S}|$ vertices has at most $|\mathcal{S}|-1$ edges. By relaxing the integer constraint to $a_{ij} > 0$, the problem can be solved in quasi-linear time with the number of edges using Kruskal's algorithm [Kruskal, 1956].

As pointed out in its definition, the MST is a subgraph, meaning that it is built from the set of nodes, edges and weights of another graph. When considering a complete spatial graph embedded in $\mathbb{R}^D$ and weights $\{w_{ij}\}$ based on the Euclidean distance, the MST is called Euclidean and is linking all the input points together with the minimum total Euclidean distance to do so. Because in applications including a large number of graph nodes K, the use of a complete graph can make the computation of the MST costly, we rely on a graph built from the Delaunay Triangulation [Cavendish, 1974] from which the MST is theoretically a subgraph [Aurenhammer et al., 2013]. In particular, since the Kruskal algorithm has a time complexity of $O(|\mathcal{E}|\log|\mathcal{E}|)$, reducing the number of edges to explore from the complete graph with K^2 edges to the $O(K)$ of the Delaunay Triangulation allow to maintain an overall complexity of $O(K\log K)$ for the computation of the Euclidean MST.

The Euclidean MST (that we simply call MST in the rest of the manuscript) graph construction has the advantage of being scale-independant and parameter-free. These properties made it an especially well-suited tool to study the filamentary part of the cosmic web in the past decades [Barrow et al., 1985; Colberg, 2007; Alpaslan et al., 2014b; Bonnaire et al., 2020; Pereyra et al., 2020a]. The right panel of Fig. 4.2 allows to appreciate the preferred path obtained for the "S" shape and gives a good intuition that, by post-processing it with simple operations (like pruning), the main underlying structure can be obtained, even though locally spiky. The main focus of the next section is precisely to obtain a locally smooth representation of the graph structure by embedding it into the probabilistic framework of mixture models (see Sect. 3.2). The limited tree topology of the MST for the representation of cycles will be the topic of Sect. 4.4 in which we present how to extend it by combining several MSTs obtained from random sub-samplings.

4.3 Graph regularised mixture models

4.3.1 Full model and formalism

We assume in this section that we are given a dataset $\boldsymbol{X} \in \mathbb{R}^{N\times D}$, resulting from the noisy sampling of a continuous unknown one-dimensional manifold. As previously exposed, the

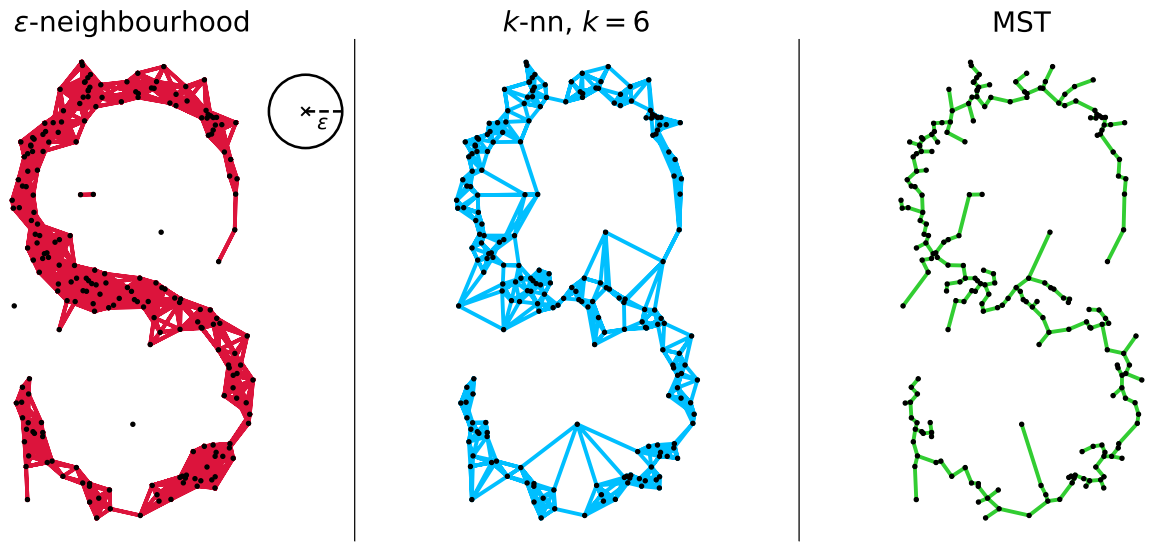

Fig. 4.2. Several graphs built from the set of datapoints in black showing a noisy "S" shape with $N = 220$ points among which 20 are outliers. From left to right: the ϵ-neighbourhood with ϵ visually shown, the k-nearest-neighbours ($k = 6$) and the minimum spanning tree.

direct application of a graph construction may not represent the topological and geometrical properties of the underlying pattern. To build a coherent-with-data graph representation, we seek to combine the mixture model formalism (Sect. 3.2) with a graph representation of the manifold, hence assuming a given topology, to extract a principal graph from the data (as illustrated in Fig. 4.3).

The data generation model

We first model the data using a mixture model with a set of K Gaussian clusters with their own centres $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$. Figure 4.3 illustrates the proposed model of the data distribution with spherical Gaussian clusters $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}_D$ hence assuming that the sampling noise around the 1D manifold is Gaussian and isotropic. The isotropy is characteristic of the tubular structures observed in many datasets like blood vessels or filaments and this spherical assumption will hold for the rest of the dissertation.

One of the drawbacks of the principal curves formulation presented in Sect. 4.1.2 is their sensitivity to outliers tending to bias the curve estimation. Real-world datasets, however, often come with outliers or noisy measurements that are not part of the underlying pattern that generated the data we aim to extract. In the mixture model formalism, we can tackle this problem by directly assuming an explicit distribution for the outliers. The choice of this distribution can be specific to the problem at hand and adapted depending on the knowledge one has on their generation. In the present work, we consider a uniform distribution of outliers over the data support volume. Hence, in addition to the K Gaussian clusters paving the distribution, we make use of a uniform background component to take into account observations that should not be part of the pattern of interest. The full mixture model can hence be written

$$p(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}, \boldsymbol{\theta}_k) + \alpha\rho(\boldsymbol{x}), \tag{4.5}$$

with $\boldsymbol{\Theta} = \{\pi_1, \dots, \pi_K, \alpha, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ the set of all parameters of the model with normalised and non-negative amplitudes $\sum_{k=1}^{K} \pi_k + \alpha = 1$ and $\rho(\boldsymbol{x}) = \left[\int_{\mathbb{R}^D} 1_{\mathcal{X}}(\boldsymbol{x})\right]^{-1}$. Consequently,

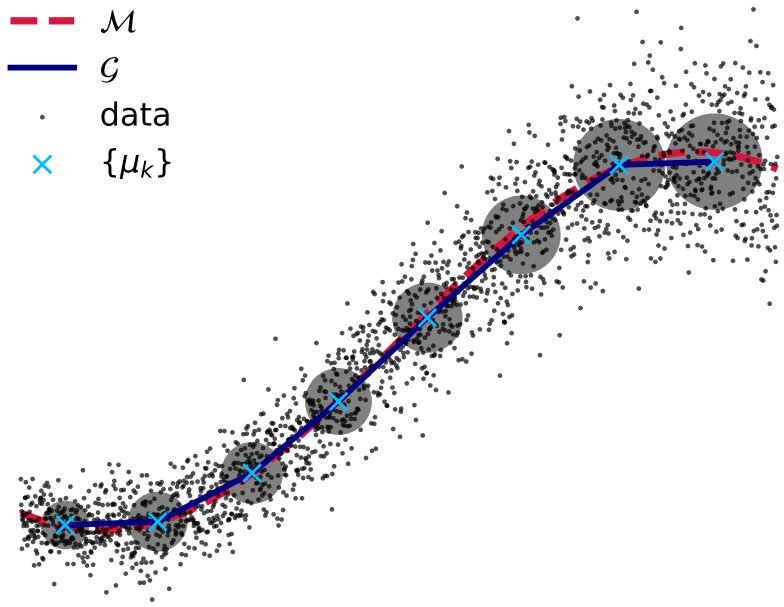

Fig. 4.3. Schematic view of the principal graph modelling proposed in this work. Datapoints are in black and represent a noisy heteroscedastic sampling of an underlying sinewave shown in dashed red. The model we propose is based on a coupling between a mixture model and a graph $\mathcal{G}$ linking Gaussian clusters together with an MST topology prescription. Blue crosses are averages of Gaussian $\{\boldsymbol{\mu}_k\}$ and grey shaded areas are the associated spherical covariance matrices $\{\boldsymbol{\Sigma}_k\}$. Edges of the MST linking Gaussian clusters are shown in dark blue.

the only information required to handle the outliers is the inverse of the data support volume, its amplitude being part of the parameters $\boldsymbol{\Theta}$, adjusted during the learning. In practice, $\rho(\boldsymbol{x})$ is estimated as the inverse volume of the convex hull of the input dataset. One of the key advantage of the proposed formulation is that outliers do not need to be removed in a pre-processing step as it is for instance the case in Chen et al. [2014] in which an arbitrary density threshold over a kernel-density estimate of the probability distribution is used. The proposed idea of built-in robustness is also explored in Gorban et al. [2016] and Albergante et al. [2020], who propose to discard from the update of a node position all datapoints beyond a robustness radius R_0. Still, the choice of R_0 is not trivial, scale-dependant and require a careful tuning to deal with uniform background noise [Albergante et al., 2020]. In our formulation, the parameters related to outliers, α and $\rho(\boldsymbol{x})$, are fully part of the model and do not require any heavy prior information nor fine-tuning to deal with uniformly distributed outliers.

Graphs as approximations of manifolds

As of now, we simply modelled the data generation process through a mixture model and wrote its likelihood. However, we additionally assume that the D-dimensional dataset $\boldsymbol{X}$ is lying on an one-dimensional Riemannian manifold $\mathcal{M}$. In simple words, manifolds extend the notion of curves, surfaces and volumes to any dimensions. When locally Euclidean and equipped with a metric, they are called Riemannian which allows for the definition of common geometric notions such as lengths, areas, volumes or curvatures. To avoid over-fitting and to get a smooth estimate of the manifold, we can use a quadratic regularisation term, as exposed in Sect. 3.1.2. Ideally, we would directly try to identify the continuous manifold $\mathcal{M}$ that generated the dataset and constrain its smoothness through the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$. In practice, $\mathcal{M}$ is not known and because of the finite nature of the considered datasets at hand, we need to

resort to an approximation. Instead, we model it as a graph structure $\mathcal{G}$ and the Laplacian of the graph can be seen as a discrete approximation of the Laplace-Beltrami operator defined on the manifold [Chung, 1999; Belkin & Niyogi, 2001]. Noting $\boldsymbol{\mu} \coloneqq (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)^{\mathrm{T}} \in \mathbb{R}^{K \times D}$, we therefore constrain the graph smoothness through

$$\begin{aligned} \|\boldsymbol{\mu}\|_{\mathcal{G}}^2 &= \sum_{i=1}^{K} \sum_{j=1}^{K} a_{ij} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 \\ &= 2\,\mathrm{Tr}\{\boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{\mu}\}, \end{aligned} \tag{4.6}$$

which converges to $\int_{\mathcal{M}} \|\nabla \boldsymbol{\mu}\|^2$, a measure of the manifold smoothness, in the infinite data limit [Hein et al., 2005; Singer, 2006]. This quantity can also be seen as the total length of the graph since $\|\boldsymbol{\mu}\|_{\mathcal{G}}^2 = \sum_{i,j=1}^{K} w_{ij}$. Minimising $\|\boldsymbol{\mu}\|_{\mathcal{G}}^2$ together with a data-driven term would hence intuitively lead to a shorter and smoother graph, similarly to what is done in Kegl et al. [2000] but without strictly fixing the length during the optimisation.

Regularisation terms

Early formulations of principal curves, based on the minimisation of the mean-squared projected distance, already suggested the need of regularisation [Hastie & Stuetzle, 1989]. In particular, Duchamp & Stuetzle [1996] show that principal curves in the plane are saddle points of the mean-squared objective function, making regularisation an indispensable way to constrain the form of the solution (see Sect. 3.1.2) when this kind of cost function is used[1]. In the present probabilistic formulation, the "curve" is introduced as a graph prior structure embedded through the regularisation term. It imposes a constraint on the behaviour of the average position of Gaussian clusters $\boldsymbol{\mu}$ through the graph smoothness defined by Eq. (4.6) to be added to the log-likelihood. This prior allows the inclusion of the geometric assumption has about the observed distribution of being generated by an underlying one-dimensional structure. In a Bayesian setup, this term can also be seen as a prior distribution over the parameter space, and more precisely as a Gaussian prior on the graph weights through the interactions between parameters of the model $\boldsymbol{\mu}_k$,

$$\log p(\boldsymbol{\mu}) = -\frac{\lambda_\mu}{2} \|\boldsymbol{\mu}\|_{\mathcal{G}}^2, \tag{4.7}$$

with λ_μ the precision parameter of the Gaussian prior distribution. We later refer to this parameter either as the precision of the prior on $\boldsymbol{\mu}$ or as a regularisation term on the log-likelihood, the two being equivalent. As we saw previously, this term corresponds to the length of the graph and will be added to the log-likelihood that we aim at maximising. It is hence equivalent to a soft minimisation of the total length of the graph structure. In the elastic map formulation of Gorban & Zinovyev [2005], a physics analogy is proposed in which Eq. (4.7) can also be seen as a stretching energy. In this analogy, graph nodes are considered as linked together by elastic bonds with elasticity coefficients λ. Such a formulation of the principal graph problem alleviates some of the drawbacks of the principal curve in which self-intersecting curves are not allowed [Hastie & Stuetzle, 1989; Kegl et al., 2000]. By restricting the graph to a chain topology with elements of the Laplacian matrix given by $l_{ij} = 2\delta_{i,i}^{\mathrm{K}} - \delta_{i,j+1}^{\mathrm{K}} - \delta_{i,j-1}^{\mathrm{K}}$, with $\delta_{i,j}^{\mathrm{K}}$ the Kronecker delta function, we can also derive the same type of curve constraints as in Yuille [1990] and Tibshirani [1992], making the graph formulation more general in terms of handled topologies.

[1] See Gerber & Whitaker [2013] for a regularisation-free formulation of the principal curve in which an alternative of the mean-squared projection distance cost is proposed.

The direct application of the EM procedure to maximise the regularised log-likelihood defined by the sum of the log-likelihood and Eq. (4.7) leads to an estimate of $\boldsymbol{\Theta}$. In this setup, the variances learnt by the procedure would depend only on the local distribution of datapoints. However, natural manifestations of heteroscedastic patterns, such as trees or blood vessels, exhibit a linear evolution of the sampling size without sudden variations. For instance, the width of a tree is continuously expanding from small branches to the trunk and the size of the vessel network is smoothly enlarging from capillaries to wide arteries. For these physically motivated reasons, the local size of the underlying continuous structure as measured by the variances of the model can be considered to evolve smoothly along the graph. To incorporate this idea in the formalism, we use again the non-Euclidean proximity measures on the graph structure for variances update. To this end, an additional prior distribution on variances is used, based on the local neighbourhood of a node. To obtain a closed-form expression of the new update equation, we use the conjugate prior for variances of a Gaussian likelihood.Formally, we rely on the inverse-Gamma distribution distribution with shape parameter $1 + \lambda_\sigma$ and scale parameter $\lambda_\sigma \sigma^2_{\mathcal{N}_k}$ defined such that the mode of the distribution is located at the mean variance of the neighbouring nodes, namely $\sigma^2_{\mathcal{N}_k} = 1/|\mathcal{N}_k| \sum_{i \in \mathcal{N}_k} \sigma_i^2$ with $\mathcal{N}_k = \{i \,|\, a_{ik} = 1\}$ and $|\mathcal{N}_k| = d_{kk}$ the degree of node k. Mathematically, we write

$$\log p(\sigma_k^2) = -\lambda_\sigma \left[\log \sigma_k^2 + \sigma^2_{\mathcal{N}_k}/\sigma_k^2\right] + \text{cst}, \tag{4.8}$$

the prior distribution for the σ_k^2 parameter and use the same $\lambda_\sigma \geq 0$ to constrain all Gaussian components.

Finally, a prior distribution is added for mixing coefficients to avoid singular solutions to happen when a node of the graph is paving an underdense region, like galaxies standing in voids for instance, making its amplitude going to 0. This can be achieved by assuming a Gaussian prior centred on uniform coefficients on the set of datapoints not represented by the background noise, namely $(1 - \alpha)/K$. Hence, we have

$$\log p(\pi_k) = -\frac{\lambda_\pi}{2} \left[\frac{1-\alpha}{K} - \pi_k\right]^2 + \text{cst}, \tag{4.9}$$

where $\lambda_\pi \geq 0$ controls the force of this prior.

The regularised mixture model

The full prior distribution on the parameter set $\boldsymbol{\Theta}$ can be written as the summation of all individual terms,

$$\log p(\boldsymbol{\Theta}) = \log p(\boldsymbol{\mu}) + \sum_{k=1}^{K} \log p(\sigma_k^2) + \sum_{k=1}^{K} \log p(\pi_k). \tag{4.10}$$

As we saw in Sect. 3.3, the optimal values of parameters can be estimated using the EM algorithm allowing the maximisation of log-likelihood in an alternating procedure. In the context of the proposed graph regularised mixture model (GRMM), we aim at maximising the log-posterior (that we equivalently call regularised log-likelihood) $\log p(\boldsymbol{\Theta} \,|\, \boldsymbol{x}) \propto \log p(\boldsymbol{x} \,|\, \boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta})$. For this purpose of maximum *a posteriori* estimation, EM can still be used with only minor changes to the introduced equations (3.12) and (3.13). Note that the E-step remains unchanged since the added term only depends on $\boldsymbol{\Theta}$ and that this first step is a maximisation of the lower bound over the latent variables only. In the M-step, we now seek to solve

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\text{argmax}}\ Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) + \log p(\boldsymbol{\Theta}), \tag{4.11}$$

where $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$ is defined through Eq. (3.12) as the expectation of the completed log-likelihood over the latent variable distribution $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta})$. We recognise here the general form for optimisation problems teased in Eq. (3.3) of Sect. 3.1.2 with the quadratic association cost hidden in the $\log \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\theta}_k)$ term of the log-likelihood and where $\log p(\boldsymbol{\Theta})$ constrains the form of the solution with strengths $\{\lambda_\mu, \lambda_\sigma, \lambda_\pi\}$.

For the log-likelihood as defined by the particular mixture model of Eq. (4.5) and the prior defined by Eq. (4.10), responsibilities of Gaussian and uniform background components, respectively noted p_{ik} and p_i^{bkg}, can be computed during the E-step as $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^{(t)})$ using Bayes' theorem and the current parameter values $\boldsymbol{\Theta}^{(t)}$

$$\begin{cases} p_i^{\text{bkg}} = \dfrac{\alpha\rho(\boldsymbol{x}_i)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j) + \alpha\rho(\boldsymbol{x}_i)}, \\ p_{ik} = \dfrac{\pi_k \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j) + \alpha\rho(\boldsymbol{x}_i)}. \end{cases} \tag{4.12}$$

Update equations for each parameter are then derived during the M-step of Eq. (4.11) as

$$\begin{cases} \alpha^{(t+1)} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} p_i^{\text{bkg}}, \\ \pi_k^{(t+1)} = \dfrac{1/N \sum_{i=1}^{N} p_{ik} + \lambda_\pi \left(1 - \alpha^{(t+1)}\right)/K}{1 + \lambda_\pi}, \\ \boldsymbol{\mu}_k^{(t+1)} = \dfrac{\sum_{i=1}^{N} \boldsymbol{x}_i \, p_{ik}/\sigma_k^2 + 2\lambda_\mu \sum_{j=1}^{K} a_{kj} \boldsymbol{\mu}_j^{(t+1)}}{\sum_{i=1}^{N} p_{ik}/\sigma_k^2 + 2\lambda_\mu \sum_{j=1}^{K} a_{kj}}, \\ \sigma_k^{(t+1)} = \left[\dfrac{\sum_{i=1}^{N} p_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2 + 4\lambda_\sigma \sigma_{\mathcal{N}_k}^2}{D \sum_{i=1}^{N} p_{ik} + 4\lambda_\sigma} \right]^{1/2}. \end{cases} \tag{4.13}$$

When not specified, parameters not indexed by time correspond to time t. This is especially important in the E-step in which all parameters are expressed at time t and in the computation of $\boldsymbol{\mu}_k^{(t+1)}$. It is indeed interesting to note the contribution of all $\boldsymbol{\mu}_j^{(t+1)}$ in the update of $\boldsymbol{\mu}_k^{(t+1)}$ in Eq. (4.13). It is hence more convenient to write this update equation matricially for $\boldsymbol{\mu} \in \mathbb{R}^{K \times D}$.

$$\boldsymbol{\mu}^{(t+1)} = \left[\boldsymbol{\Gamma}\boldsymbol{S}^{-1} + 2\lambda_\mu \boldsymbol{L}\right]^{-1} \boldsymbol{S}^{-1} \boldsymbol{R}^{\mathrm{T}} \boldsymbol{X}, \tag{4.14}$$

where $\boldsymbol{S}$ is a diagonal $K \times K$ matrix with $s_{kk} = \sigma_k^2$, $\boldsymbol{\Gamma}$ a diagonal $K \times K$ matrix of average data-points explained by the k^{th} component, i.e. $\gamma_{kk} = \sum_{i=1}^{N} p_{ik}$ and $\boldsymbol{R} \in \mathbb{R}^{N \times K}$ the responsibility matrix for Gaussian components such that $r_{ik} = p_{ik}$.

4.3.2 Algorithm and illustrative results

Algorithm (1) sums up all the steps for the learning of the proposed principal graph from a given set of measurements $\boldsymbol{X}$ and a graph construction prior $\mathcal{G}$. It extends the GMM by embedding a topological prior through $\mathcal{G}$ constraining the position of centres, propose an handling of outliers and the learning of the local width of structures. In this model, the topology of

Algorithm 1 Graph regularised mixture model (GRMM)

Input: Data: $\boldsymbol{X} \in \mathbb{R}^{N \times D}$, hyper-parameters: $\mathcal{Y} = \{\lambda_\mu, \lambda_\sigma, \lambda_\pi\}$, initialisation: $\boldsymbol{\Theta}^{(0)}$, $\mathcal{G}^{(0)}$.
Output: $\boldsymbol{\Theta}^{(t)}$ and $\mathcal{G}^{(t)}$.
while $\mathcal{R}_t \geq \epsilon$ **do**
Compute the adjacency matrix $\boldsymbol{A}$ of $\mathcal{G}^{(t)}$
E-step: Compute responsibilities p_{ik} and p_i^{bkg} using equations (4.12)
M-step: Compute new parameters $\boldsymbol{\Theta}_t$ based on responsibilities using equations (4.13) and (4.14)
Compute the increment in log-posterior $\mathcal{R}_t = \log\, p(\boldsymbol{\Theta}^{(t)} \,|\, \boldsymbol{X}) - \log\, p(\boldsymbol{\Theta}^{(t-1)} \,|\, \boldsymbol{X})$
Optional: Update graph topology by recomputing $\mathcal{G}^{(t)}$ on the new set $\{\boldsymbol{\mu}_k\}_{k=1}^K$
end while

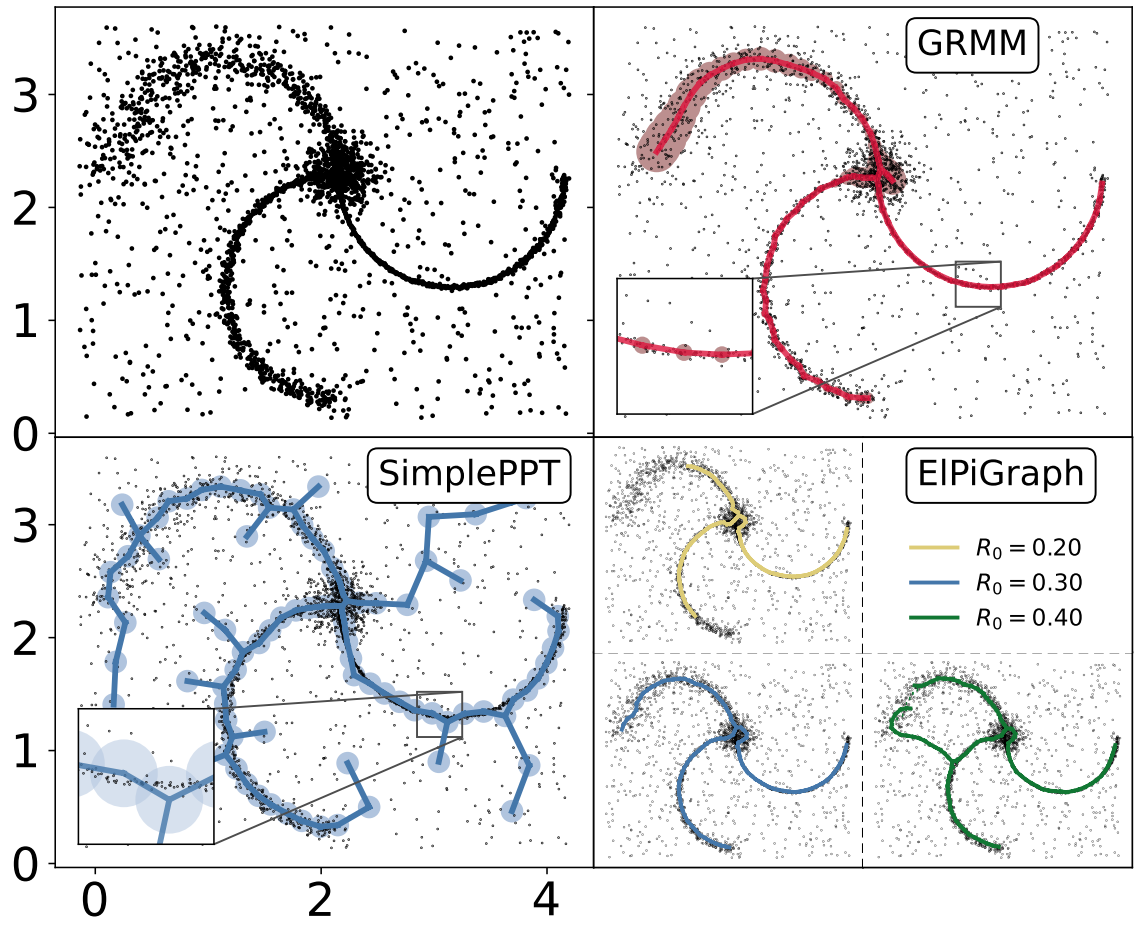

Fig. 4.4. Illustrative comparison of three procedures to learn a principal graph on an artificial dataset made of $N = 2666$ datapoints with three branches of linearly evolving standard deviations converging to a spherical Gaussian cluster shown in the top left panel. Top right is the principal graph learned by the proposed GRMM algorithm (1) with $K = 100$ initialized randomly over $\boldsymbol{X}$, $\mathcal{Y} = \{5/\sigma_0^2, 10, 1\}$ and $\sigma_0^2 = 0.01$. Bottom left is the one from the SimplePPT algorithm [Mao et al., 2015, 2017] with $\sigma = \sigma_0$ and identical initialisation nodes. In both cases, the shaded areas correspond to the 1-σ_k circles centred on $\boldsymbol{\mu}_k$. The bottom right panel is the result from the ElPiGraph procedure [Albergante et al., 2020] with 100 nodes initialized with 70 taken randomly over $\boldsymbol{X}$ and for different values of the trimming radius R_0 with the same elasticity parameters $\lambda = 0.01$, the length constraint and $\mu = 0.01$ the bending constraint.

the graph can be either fixed or updated during the optimization. This is particularly important since the prior approximation on the noisy pattern with outliers may not be suitable to describe the final smooth underlying structure. For instance, a bifurcation in the prior graph may not be relevant in the final one. Consequently, it may be useful to update the topological prior based on the current estimate of $\{\boldsymbol{\mu}\}_{k=1}^{K}$. When updating the topology at each iteration, $\boldsymbol{A}$ cannot be understood as a prior anymore but has to be included as a parameter in the model making the chosen graph construction itself acting like a prior. In the case of fitting an MST topology, we can embed the integer programming optimization problem definition of the MST from Eq. (4.4) in the regularized log-likelihood hence leading to estimate both $\boldsymbol{\Theta}$ and $\boldsymbol{A}$ in the M-step which simply involves computing the new graph $\mathcal{G}_t$ at iteration t given a graph construction process and the positions $\{\boldsymbol{\mu}\}_{k=1}^{K}$. When refining the graph, update equation for $\boldsymbol{\mu}_k$ of the SimplePPT algorithm proposed in Mao et al. [2015] can be derived from Eq. (4.14) by considering fixed variances, no outlier distribution and an MST prior.

Figure 4.4 provides an illustrative comparison of several algorithms to identify a principal graph in point-cloud datasets. The top right panel of Fig. 4.4 shows the graph learned by the GRMM Algorithm (1) for a dataset made of three branches with linearly evolving standard deviations $\sigma \in [0.015, 0.15]$ and 25% background noise uniformly added in the square. The bottom left panel of Fig. 4.4 illustrates the principal graph learned by the SimplePPT algorithm on the same dataset and using the same parameters and initialisation as our construction. Many branches are falling in the background noise and, because of its fixed-variance scheme, it fails at capturing the ridges of both the large and small variance branches. Choosing a large value for σ_0^2 biases the graph in the small branch, while a low one irremediably creates a lot of spurious branches in the large-variance part of the pattern, even in the absence of outliers. With its contracting iterative scheme, the Subspace Constrained Mean Shift [SCMS, Ozertem & Erdogmus, 2011; Genovese et al., 2014; Chen et al., 2014] algorithm would suffer less from the latter exposed problem. Yet, it does not provide a properly-speaking curve, but a set of independant projected points that are to be linked *a posteriori* and the SCMS is also usually used together with a pre-processing step to filter datapoints standing in low-dense areas such as proposed in Chen et al. [2015] hence involving an additional non-trivial parameter impacting the quality of the result. The sensitivity of the graph to outliers is handled by the ElPiGraph framework [Albergante et al., 2020] through a trimming radius R_0 discarding distant datapoints from the update of a node position. In the bottom right panel of Fig. 4.4 can be found three realisations of this algorithm with different values of R_0, showing the robustness of ElPiGraph to uniform background noise with a fine-tuned value. The choice of R_0 however remains scale-dependent and can alter the recovered structures in case of heteroscedastic patterns. It is also not obvious to tune in real applications and comes in addition to other hyper-parameters already complex to choose. Moreover, the resulting graph do not embed a description of the local width as opposed to the GRMM which fits all parts of the pattern in its center. Note nonetheless that the ElPiGraph algorithm handles better the dense area at the center of the dataset when $R_0 = 0.3$ with a single bifurcation thanks to an added terms to the cost function minimising the overall graph complexity. The algorithm we develop here [Bonnaire et al., 2021b], by unifying all these aspects of robustness and local size description in a single formulation, is able to provide a principal graph that visually shows a smoothly evolving adaptive scale along the graph structure. Even though the graph is initialized with only $K = 100$ nodes taken randomly over $\boldsymbol{X}$ with variances $\forall k, \sigma_k^{(0)} = 0.08$, there are no final nodes standing outside of the pattern and the resulting estimate of the background level is 25%. Note also that, on patterns showing no outliers and with branches of the same sizes, all algorithms provide similar results with a smooth graph passing in the middle of the structure.

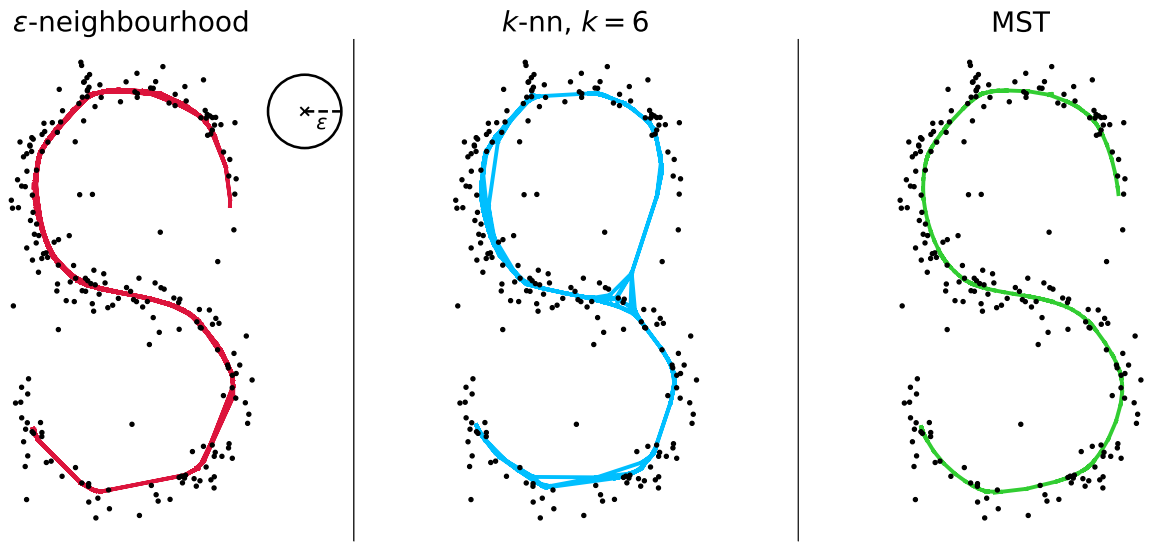

Fig. 4.5. The regularised versions of the graph priors shown in Fig. 4.2 obtained from Algorithm (1) with $K = N = 220$, $\mathcal{Y} = \{0.5/\sigma_0^2, 0, 0\}$ and $\sigma_0^2 = 0.1$. From left to right: the ϵ-neighbourhood with ϵ visually shown, the k-nearest-neighbours ($k = 6$) and the minimum spanning tree.

4.4 About graph priors

4.4.1 Basic graph constructions

The choice of the prior graph construction depends on the guess of the underlying manifold topology on which the data presumably live. Manifolds learning techniques for dimensionality reduction based on graphs usually rely either on a complete graph with weights obtained from a heat-kernel $w_{ij} = \exp\{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2/s\}$ with the choice of s being data-dependant Belkin & Niyogi [2003], either on a k-nearest neighbours graph. Such constructions are convenient for the precise purpose of dimensionality reduction in which the mapping between the high and low dimensional spaces is supposed to preserve the local neighbourhood of the data, assuming the manifold to be locally Euclidean [Belkin & Niyogi, 2003]. In our context of ridge estimation, we find that such local constructions, in addition to depend on a free parameter, are also creating a lot of redundant and spurious edges hence degrading the pattern extracted from Algorithm (1), as shown in the middle panel of Fig. 4.5 where the "S" shape of the pattern is not correctly extracted. Note however that, with a proper post-processing of the graph prior, pre-processing of the input dataset, or fine-tuning of the parameter for the graph prescription (ϵ or k), we could end up with a clean estimate of the ridge. The MST prior, on the other hand, is fitting well the expected underlying shape and has the advantage of being parameter-free, scale-free and can handle general patterns by assuming a tree-like topology.

4.4.2 The average graph prior

Although the minimum spanning tree exhibits some nice features discussed above, it has a limited representative power for general datasets due to this latter-mentioned tree topology that cannot represent cycles, as shown in the top left panel of Fig. 4.6. By using a non-cycling topology, the regularised graph becomes even more inaccurate if the data embeds holes because of the optimisation tending to shorten graph extremities hence emphasising the absence

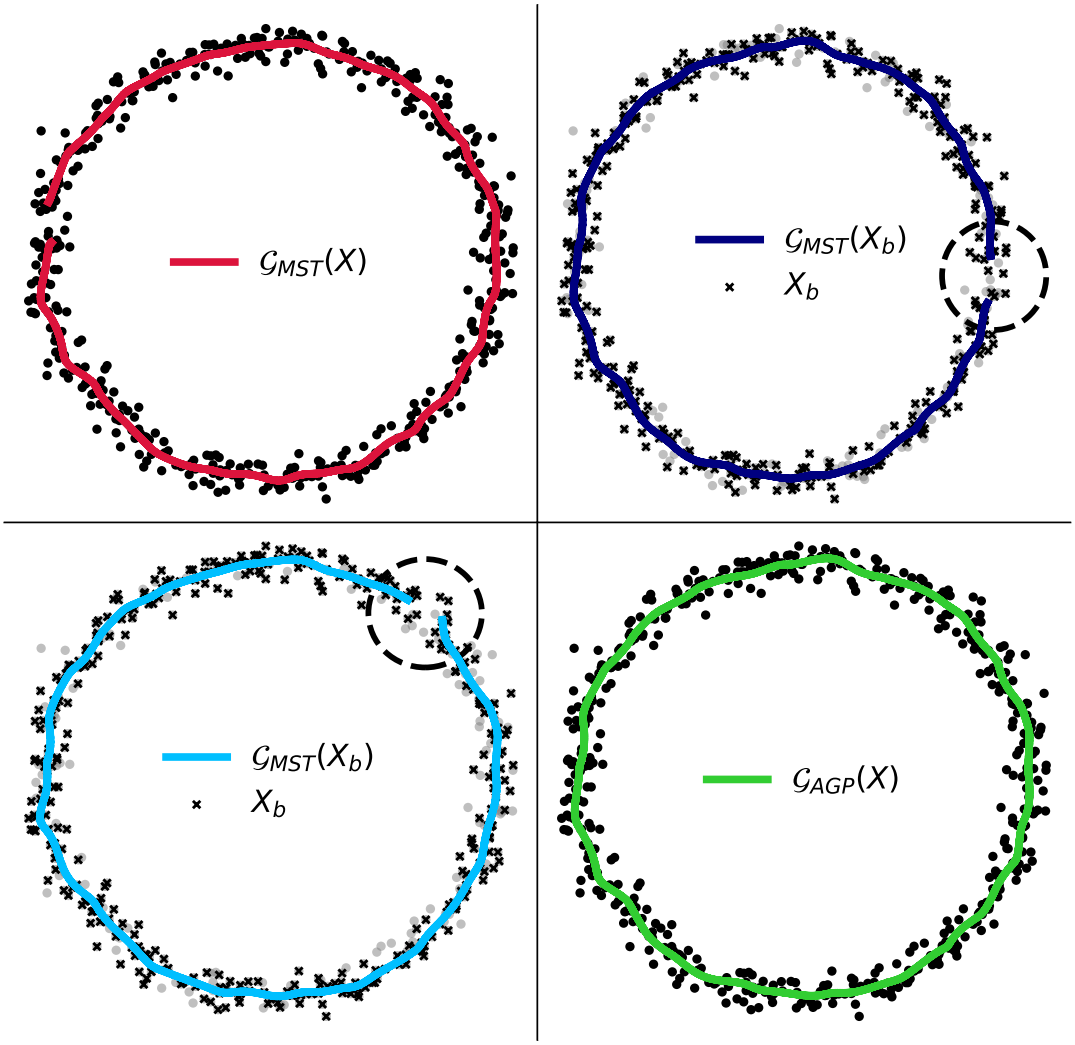

Fig. 4.6. Illustration of the interest of combining MSTs obtained from random sub-samplings of the data. *(top left)* The regularised MST computed over the all the input datapoints. *(top right)* The regularised MST computed over a random sub-sampling of the input with $N/N_b = 0.75$ (black crosses). Grey datapoints are those not used. *(bottom left)* Same as top right for another sub-sampling realisation. *(bottom right)* The graph obtained using the average graph prior. The two circles in the top right and bottom left panels illustrate that the hole not captured by the MST varies with the sub-samplings. This instability is exploited in the average graph prior to obtain a cycling topology.

of cycles. Several graph constructions allow such a feature, as for instance the k-nearest neighbours graph. This solution, in addition to create long-range edges for isolated nodes that would turn in branches reaching outliers or spurious cycles, is not based on an optimisation of the total length of the graph, which is convenient for convergence properties of the algorithm when updating the prior (see Sect. 4.5).

Because the MST results from a global minimization of the total Euclidean distance as can be seen in Eq. (4.4), the obtained preferred path is very sensitive to random removals of datapoints (see top right and bottom left panels of Fig. 4.6). We exploit this particular aspect by merging the idea of a graph with the minimum total length and the handling of one-dimensional holes in the dataset to propose an empirical construction based on the computation of MSTs obtained from a set of B realisations of random sub-samplings of the set

of nodes. Formally, the average adjacency matrix is $\bar{\boldsymbol{A}} = \sum_{b=1}^{B} \boldsymbol{A}_b / B$, where $\boldsymbol{A}_b$ is the adjacency matrix of the MST computed the b^{th} random sub-sampling of $\boldsymbol{\mu}$ using a fraction K_b/K of points. This approach, proposed in the context of pattern extraction to provide an uncertainty measurement [Chen et al., 2014; Bonnaire et al., 2020; Albergante et al., 2020], is exploited in the present case to combine all the realizations to obtain a single graph construction. This operation is also done *a priori* on non-regularized MSTs hence reducing the computational cost by avoiding the need of running Algorithm (1) multiple times.

The average adjacency matrix $\bar{\boldsymbol{A}}$ represents the frequency of appearance of an edge during B realisations of MST and hence the probability, for each pair of centres, to be linked. To illustrate the method, we build a discrete Voronoi dataset consisting of intersecting straight lines surrounding low-density areas visible on the top panel of Fig. 4.7. Such a dataset aims at visually reproducing the cosmic web with elongated intersecting 1D structures and was used to assess the quality of the detected structures in Aragón-Calvo et al. [2010] or Chen et al. [2015] but also to train a U-net architecture in Aragon-Calvo [2019]. The bottom panel shows the distribution of edge probabilities for several ratios K_b/K. For a large range of values of this ratio, we clearly distinguish two populations of edges: those with high probabilities and those with almost-zero probabilities. In the high probability population, we retrieve, in the three cases, all $K-1$ edges of the original MST computed over $\boldsymbol{\mu}$ plus **27** new ones which correspond to the exact number of closed cycles in the artificial dataset. These additional edges are shown in bold blue in the top panel of Fig. 4.7. When the ratio K_b/K becomes smaller, the high probability mode tends to be centred at lower and lower probability until the pattern is so much altered by the sub-sampling that it is not retrieved and the distribution becomes centred at low probability. The proposed method hence allows to take advantage of the inherent instability of the MST to recover additional edges inducing cycles independently of their scale. It solely and indirectly depends on the edge length required to close the cycle without imposing neither a formal definition for the cycle nor a hard threshold in edge length.

Even though observed in the example of Fig. 4.7, the retrieval of all the $K-1$ edges of the MST is not theoretically guaranteed. To ensure it, the resulting set of edges is obtained by the union of all MST edges and those in the high probability mode leading to the non-singular symetric adjacency matrix

$$(\boldsymbol{A})_{ij} = \max\left((\boldsymbol{A}_{\text{MST}})_{ij}, (\bar{\boldsymbol{A}}_{>m})_{ij} \right), \tag{4.15}$$

with $\bar{\boldsymbol{A}}_{>m}$ the thresholded average adjacency matrix $\bar{\boldsymbol{A}}$ at level m such that it isolates the high probability mode only. By doing so, the proposed graph construction is an extension of the MST containing all its edges with additional ones sharing similar probabilities of appearance during all the random computations.

For the sake of illustrative comparison, we add, as the turquoise blue line of the top panel of Fig. 4.7, the result of the cycling topology as defined by the framework of persistent homology Edelsbrunner et al. [2002]. In particular, we use the 1D-homologically persistent skeleton (HoPeS) of Kurlin [2015] built as the MST completed with edges that are creating persistent 1D homologies in the dataset. After the extraction of those critical edges, it is not trivial to obtain a persistence threshold that captures the desired cycles. Here, we rely on the bootstrap procedure that, together with the stability of the persistence diagram allow a statistically well-defined computation of a threshold [Chazal et al., 2018]. We see that some additional edges are creating high-scale cycles due to the noisy datapoints leading to inaccuracies in the representation of the underlying pattern. Even though these undesired cycles could be handled using the sub-level sets analysis of the distance-to-measure function of Chazal et al. [2011, 2018], it does not allow an easy extraction of a graph on the inputted datapoints as required by our

application. A pre-processing of the data by removing points in low-density areas or a careful fine-tuning of the persistence threshold are also possible but this is at odds with the proposed formalism supposed to handle such datapoints. One hint that these spurious branches come from the noisy measurements is that, when varying the randomness of the sampling, we end up with a different pattern (plotted as the orange line on the top panel of Fig. 4.7) while the one traced by the proposed average graph remains the same. It is also worth emphasising that, in both realisations of the 1D persistent skeleton, the two small-scale cycles at the center of the dataset are not captured because of the too high persistence threshold imposing a lower-bound on the scale of the detected cycles while they are observed in our average graph prior.

4.5 Convergence and time complexity

4.5.1 Convergence analysis

Convergence towards a local maximum and monotonic increase of the log-posterior through iterations are guaranteed by the EM procedure when fixing $\forall t, \mathcal{G}_t = \mathcal{G}_0$ [see McLachlan & Krishnan, 1997, for a detailed analysis of EM convergence properties]. When updating the topology at each iteration and using a prior based on the MST, convergence as well as monotonic increase of the regularised likelihood are still guaranteed [Mao et al., 2017], but it does not remain true for any general graph construction if not based on optimisation procedures such as defined by Eq. (4.4). The left panel of Fig. 4.8 shows the convergence for the learning of the principal graph from Fig. 4.4. We see that the regularised log-likelihood given by $\log p(\boldsymbol{X} \mid \boldsymbol{\Theta}) + \log p(\boldsymbol{\Theta})$[2] is getting maximised during the procedure, as theoretically predicted by EM. Since this quantity can be computationally costly to evaluate, we can also rely on the increment in parameter values as a stopping criterion. As an example, the evolution of $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(t-1)}\|_2$ is shown in dashed red on the left panel of Fig. 4.8, indicating that nodes of the graph are not significantly moving after roughly 50 iterations, which is also where the regularised log-likelihood gets stable.

4.5.2 Time complexity

The E-step equation (4.12) requires the computation of the responsibilities taking $O(NDK)$ operations to complete while the M-step equations (4.13) respectively require $O(NK)$, $O(NK)$, $O(K^3)$ and $O(NKD)$ operations. The added term on the log-likelihood hence leads to an increase of the complexity upon the usual EM update of center means. Considering T iterations for the algorithm to converge and a complexity of $C_{\mathcal{G}}$ for the graph structure computation, Algorithm (1) needs $O\left(T\left[K^3 + NKD + C_{\mathcal{G}}\right]\right)$ operations. $C_{\mathcal{G}}$ acts on K and can take very different forms depending on the used topological approximation. For the minimum spanning tree, it takes $O(K \log K)$ operations to build the Delaunay Tessellation and $O(K^{1+D/2} \log K)$ to find the MST with Kruskal's algorithm. Naturally, the proposed average graph being obtained from B random sub-samplings of the MST, it requires $O\left(B\, C_{\text{MST}}(K_b) + C_{\text{MST}}(K)\right)$ operations, with $C_{\text{MST}}(K)$ the complexity of getting an MST over K vertices.

[2] The first term is computed through Eq. (4.5) and the second one is the prior from Eq. (4.10).

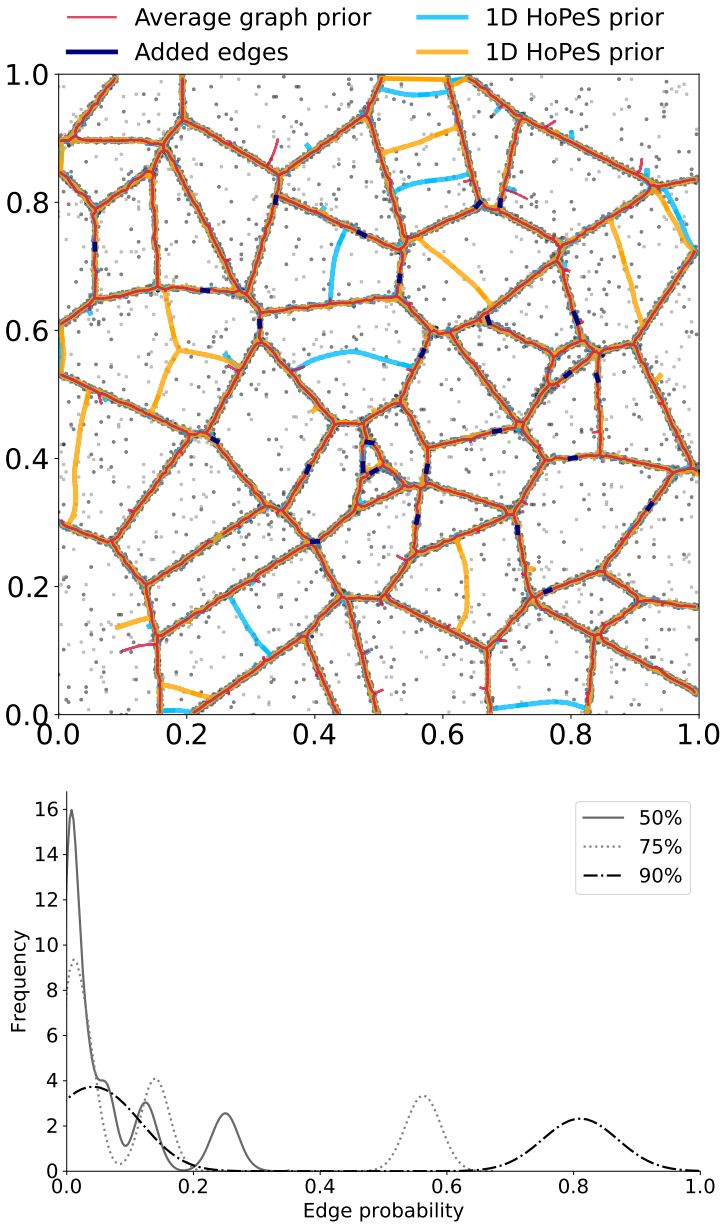

Fig. 4.7. Illustration and comparison of the average graph prior. *(top)* Black points and grey crosses are two sampling realisations of an artificial 2D dataset obtained from a Voronoi pattern. Red edges are those from the MST and bold dark blue ones are those added by the high probability mode with $B = 500$, $K_b/K = 0.75$ and thresholded at $m = 0.35$. The turquoise blue and orange lines are the set of edges from the regularised graph obtained with a prior given by the 1D-homologically persistent skeleton [Kurlin, 2015] with the two realisations of the dataset. When not explicitly visible, it means that the three lines overlap perfectly. *(bottom)* Probability distributions of edge probabilities $\left(\bar{\boldsymbol{A}}\right)_{ij}$ for several ratios K_b/K.

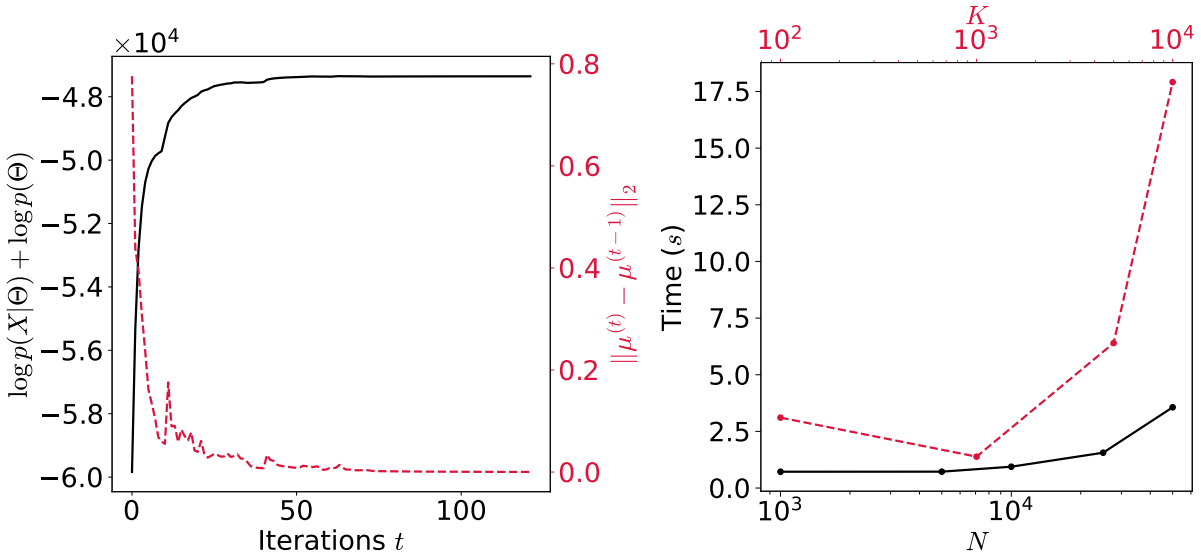

Fig. 4.8. *(left)* Convergence of the regularised log-likelihood (plain black) and of graph node positions (dashed red) with the iterations to obtain the principal graph from Fig. 4.4. *(right)* Runtimes from the Python implementation of Algorithm (1) for several values of N (black solid line) with fixed $K = 1000$ nodes and for several values of K (red dashed line) at fixed $N = 10000$. Runs have been carried out on 2D Voronoi-like mock data already used in Sect. 4.4 and shown on Fig. 4.7.

4.5.3 Runtimes

A Python implementation[3] of the proposed Algorithm (1) provides the principal graph from Fig. 4.7 in less than three minutes on a modern laptop with our single-core implementation. Figure 4.8 quantifies the evolution of the runtime for different number of input points, with fixed $K = 1000$ nodes in black and the evolution with K for fixed input $N = 10000$ in red. Using the MST prior computed using the Kruskal algorithm applied on the Delaunay Tessellation yields a fast algorithm, able to deal with large-scale datasets, as it can be the case in cosmological applications. The implementation additionally exploits adapted data structures to encode the sparse matrices encountered in the algorithm, like the adjacency or Laplacian matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ making it also efficient from a memory point of view. Besides, it makes use of binary search tree structures [Bentley, 1975] to optimise the computation of responsibilities requiring the evaluation of squared distances $\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$ under our assumption of spherical covariance matrices in the Gaussian mixture model.

4.6 Hyper-parameters and initialisation

4.6.1 The impact of parameters

Hyper-parameters of the full model are K and $\mathcal{Y} = (\lambda_\mu, \lambda_\sigma, \lambda_\pi)$. K denotes the number of Gaussian components used in the mixture model while $\lambda_\mu, \lambda_\sigma, \lambda_\pi$ are all related with shapes of prior distributions on the parameter indicated as subscripts.

[3] Available at *https://git.ias.u-psud.fr/tbonnair/t-rex*.

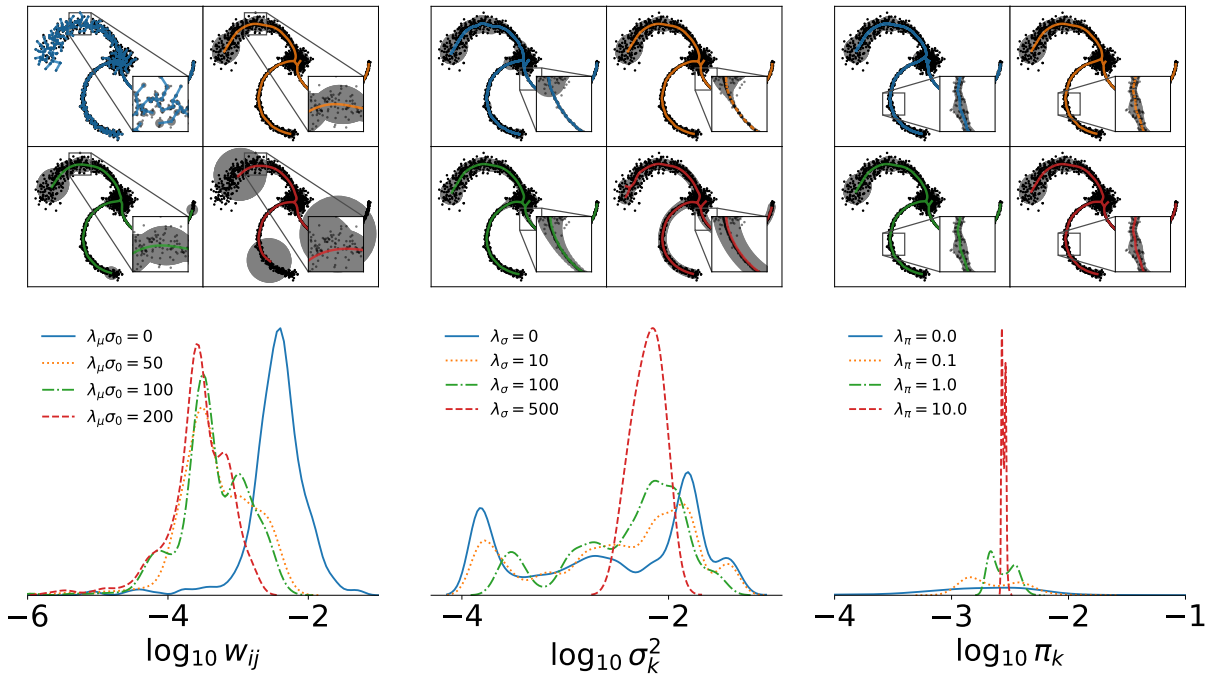

Fig. 4.9. Illustration of the impact of hyper-parameters $\mathcal{Y} = \{\lambda_\mu, \lambda_\sigma, \lambda_\pi\}$ in Algorithm (1) for an artificial dataset made of three branches with linearly evolving standard deviations converging to a spherical Gaussian cluster. In all cases, $K = 350$ nodes are used with the same random initialisation over $\boldsymbol{X}$. Black dots are datapoints, coloured lines are the set of edges learned by Algorithm (1) and gray shaded areas correspond to the 1-σ_k circles centred on $\boldsymbol{\mu}_k$. *(top row)* From left to right, quadrants corresponds to several values of λ_μ, λ_σ and λ_π. *(bottom row)* Probability distributions of the impacted parameters. From left to right: edge weights $w_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2$, variances of components σ_k^2, and mixing coefficients π_k.

Effect of λ_μ. λ_μ controls the force of smoothness constraint and corresponds to the precision of the prior Gaussian distribution put on edge weights. From Eq. (4.7), and emphasised by the upper left quadrant of Fig. 4.9 showing regularized graphs obtained with increasing values of λ_μ from 0 to $200/\sigma_0^2$, we see that the higher λ_μ, the more $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are pulling themselves along the graph structure leading to constrain the graph length $\sum_{ij} w_{ij}$ by shortening its extremities. The bottom left panel of Fig. 4.9 shows the final distribution of edge weights, in other words the distances between linked nodes in the graph, w_{ij}. We clearly observe the mode of the distribution being translated to lower values when λ_μ increases, indicating lower and lower distances between connected nodes.

Effect of K. At a fixed value of λ_μ, a large increase of the number of nodes K can induce over-fitting. Sticking with a similar extracted pattern would hence require the increase of λ_μ. This is for instance illustrated when comparing Fig. 4.4 obtained with $K = 100$ and Fig. 4.9 with $K = 350$ and required a larger value of λ_μ to provide a similar pattern. One way to reduce the interlink between these two parameters and the dependency of the pattern on K is the growing grammar proposed in Gorban & Zinovyev [2009] and followed-up by Albergante et al. [2020] in which the graph is initialized with a small value of K to then add some nodes iteratively using predefined rules. It also comes with a way to penalise "complex" graphs by adding a term to the log-likelihood and has the advantage to diminish the computational cost at early iterations.

Effect of λ_σ. λ_σ acts on the shape of the inverse-Gamma prior distribution of Gaussian variances σ_k^2. Equation (4.13) teaches us that, as for other hyper-parameters, when $\lambda_\sigma \to 0$, the prior is cancelled and variances are updated through the usual EM equation. When λ_σ increases, the shape of the inverse-Gamma law is more and more constrained leading σ_k^2 to be updated mainly through $\sigma_{\mathcal{N}_k}^2$ and, $\forall k, \sigma_k^2 \simeq \sigma_0^2$ with σ_0^2 the value used to initialise the algorithm. Eventually, a high enough value for this parameter thus leads to a fixed scale version of the algorithm, similar as those presented in Mao et al. [2017] and Bonnaire et al. [2020], as shown in the middle bottom panel of Fig. 4.9 where the distribution of $\log_{10}\sigma_k^2$ is more and more centred around $\log_{10}\sigma_0^2$. As illustrated in the upper middle quadrant of Fig. 4.9, when λ_σ increases, the structure shows a smoothly evolving variance along the graph structure with multiple scales described at the same time. This is also highlighted by the lower panel where the variance distributions obtained from regularized graphs with small values of λ_σ are spreading over several order of magnitudes and show three distinct modes corresponding to the characteristic scales of the three branches in the dataset. These estimates tend to be more and more biased toward σ_0 when λ_σ increases.

Effect of λ_π. Finally, λ_π is acting on the amplitude π_k, corresponding to the proportion of points being represented by the Gaussian component k and is controlling the force of the prior of uniformly distributed Gaussian mixtures. This parameter has a very low impact on the overall result but can avoid singular solutions to happen in very low-density regions paved by graph nodes. The upper right panel of Fig. 4.9 shows the low impact of the parameter despite a large range of tested values, from 0 to 10. Although the obtained distributions of mixing weights π_k are clearly different, as seen in the lower right panel tending to be centred at $(1-\alpha)/K$ when λ_π increases, the obtained graph structure remain unchanged, so as the local extent measured by the variances.

Even though the impact of these hyper-parameters is interpretable and that there is a wide range of values providing similar results, the "quality" of the obtained graph still depends on the settings, and, to the best of our knowledge, there is no well-defined method to choose regularisation parameters independently from user tests or external information. In that regard, the theoretical work of Gerber & Whitaker [2013] in which is proposed a modification of the L_2-form of the objective function of the principal curve formulation. By minimising this new cost, from which admissible principal curves are minima, they are freed from any regularisation parameter and can hence be of interest to alleviate the complexity of the model selection process.

4.6.2 Initialisation

Initialisation rules

There are four sets of parameters to initialise, namely $\{\boldsymbol{\mu}_k\}_{k=1}^K$, $\{\sigma_k\}_{k=1}^K$, $\{\pi_k\}_{k=1}^K$ and α. A simple and direct strategy to initialise positions of Gaussian components is to choose $\boldsymbol{\mu}^{(0)} = \boldsymbol{X}$. By doing so, we ensure that the observed point cloud distribution is well paved by centres. Note however that, from Sect. 4.5, when $K \simeq N$, the complexity scales with N^3. For large datasets, it may be interesting to first reduce the complexity by initialising the model with $K \ll N$ using for instance sub-samplings, noise reduction techniques or simple clustering methods like the K-Means or fiducial GMM algorithms as the case $K = N$ will generally produce more clusters than needed to pave the dataset. When no prior knowledge on the local size of structures, variances can be initialized as $\forall k \in \{1, \dots, K\}, \sigma_k^{(0)} = \sigma_0^{(0)}$. In this

context, $\sigma_0^{(0)}$ can be chosen as a prior guess on the average size of structures or through rules borrowed from density estimation methods [Heidenreich et al., 2013]. In practice, we can choose $\sigma_0^{(0)}$ using a modified version of the Silverman's rule [Silverman, 1986] as

$$\sigma_0^{(0)} = A_0 \left[N(D+2)\right]^{\frac{-1}{D+4}} \sigma_{\min}, \tag{4.16}$$

where A_0 is a constant, N is the number of datapoints, D is the dimension of the data and $\sigma_{\min}$ is the minimum standard deviation over all directions. Taking $A_0 = 1$ leads to the Silverman's rule and is the optimal estimate for an underlying Gaussian distribution. As argued by Chen et al. [2015], when the data are not Gaussian anymore, A_0 should be optimised as a free parameter.

The α parameter is the evaluation of the level of outliers in the dataset that should not be paved by the Gaussian components. Its value depends on the application and data at hand. In our experiments, we fix $\alpha^{(0)} = 0.10$ and then let it adjust itself during the learning. As an example, results from Fig. 4.9 were obtained by starting with this guess and then quickly converges towards a value numerically indistinguishable from 0. Finally, mixing coefficients are initially assumed to be uniformly distributed and $\forall k \in \{1, \ldots, K\}, \pi_k^{(0)} = \left(1 - \alpha^{(0)}\right)/K$.

Impact of the initialisation on the learning

Because the proposed algorithm relies on the EM procedure to maximise the log-posterior, it shares the same drawbacks, namely the convergence towards a local maximum, as exposed in Sect. 3.3. This makes, in principle, the full optimisation scheme dependant on the initialisation of $\boldsymbol{\Theta}$. One way to obtain a graph independant from the initialisation is to use the prescriptions from statistical physics established in Sect. 3.4. As already stated in Chapter 3, such simulated annealing approaches were introduced to track the global maximum of the log-likelihood [Kirkpatrick et al., 1983; Ueda & Nakano, 1998]. This solution however comes with a sizeable increase of the computational cost and, in all our experiments, did not significantly improve the quality of the extracted pattern.

4.7 Illustrative application: Road network

In Chapter 5, we shall see the application of the proposed method for the detection of the filamentary pattern of the cosmic web based on halos or galaxies. In this section, the context of road network detection from noisy GPS measurement is used as a showcase. The detection of roads is usually carried on images acquired by satellites, a task for which many algorithms have been proposed [see for instance Merlet & Zerubia, 1996; Stoica et al., 2004; Wang et al., 2016]. However, with the expansion of GPS technologies and their integration in smartphones, the update of maps obtained by satellite imaging techniques became possible at small cost. One of the objective of map reconstruction methods is to produce street maps from a set of sampled positions collected by travelling vehicles (like taxis or volunteers). Reaching this goal requires sophisticated algorithms to be deployed to handle the complex patterns observed, the high level of noise and outliers present in these datasets but also the uneven sampling of several trajectories. Main roads are indeed traced by hundreds of trajectories, while some low-level ones are traced only by a few, as illustrated in the top left panel of Fig. 4.10.

When applying our Algorithm (1) with the average graph prior on the 67 193 datapoints Berlin dataset, we obtain the graph from the top right panel of Fig. 4.10. Note that no pre-processing of the dataset nor post-processing of the resulting graph structure have been carried out. We see that the principal graph could use some topological refinements to make

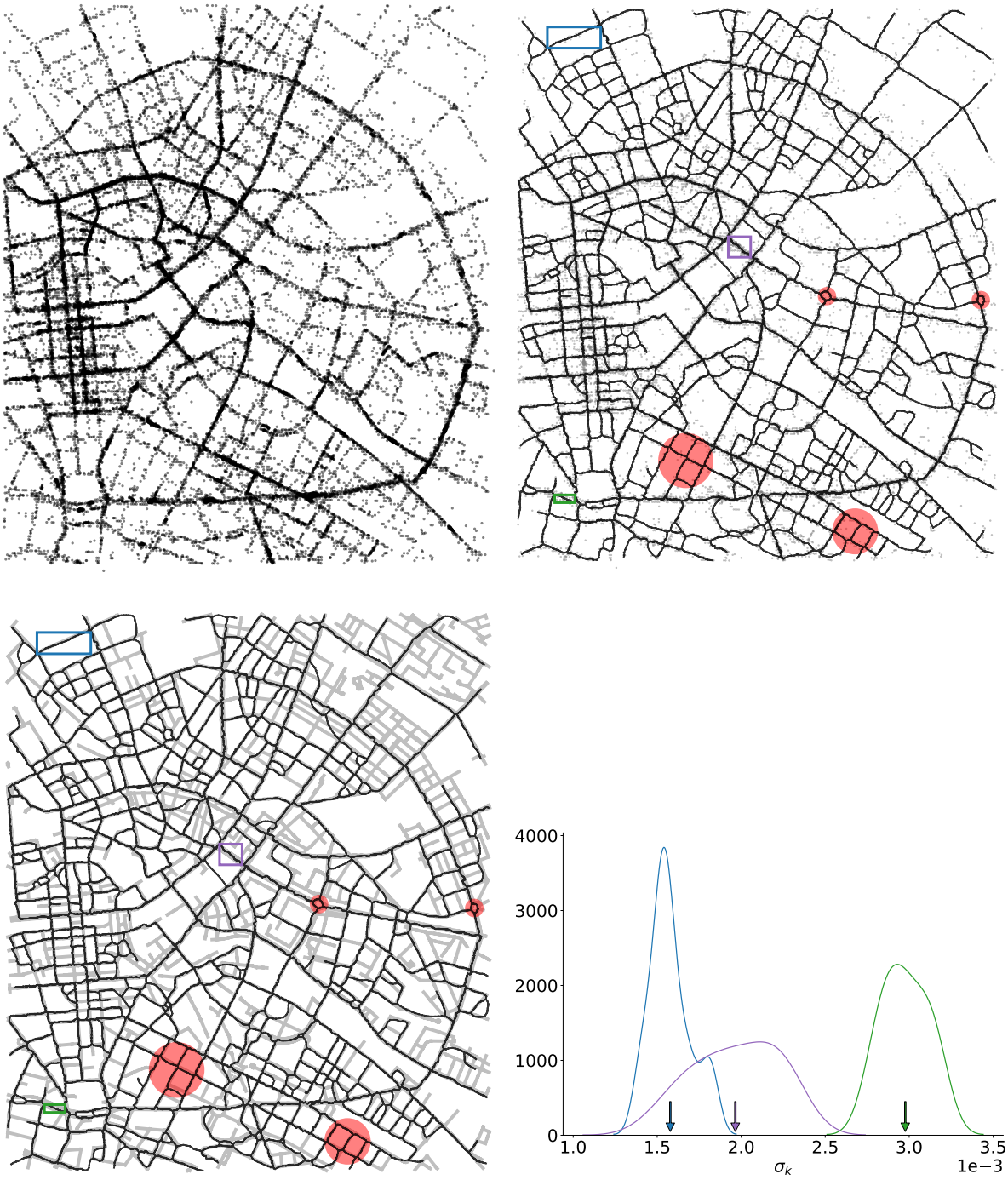

Fig. 4.10. Application of the principal graph learning to the extraction of road network from vehicle positions. *(top left)* The 67 193 input datapoints of the Berlin dataset. *(top right)* Black lines are edges of the regularized graph of Algorithm (1) overplotted on the raw datapoints. The graph was obtained using the average graph topology with $K = 7300$ and $\sigma_0 = 0.003$, $\mathcal{Y} = (10/\sigma_0^2, 5, 1)$. Red circles highlight features that could not be caught by a non-cycling topology. *(bottom left)* Same as for the top panel, but overplotted on the ground truth map extracted from OpenStreetMap. *(bottom right)* Probability distributions of variances for nodes in the corresponding circled coloured regions of the top panel.

it look smoother, removing waving branches and closing some of the remaining dangling branches. For more details about particular processing of inputs and outputs in the context of graph structure learning for road networks, see [Huang et al., 2018]. The obtained graph passes through the entire dataset, correctly paving regions with both high and low densities. We can also observe that the proposed algorithm recovers a large variety of cycle sizes, from very small ones, as for instance in the red shaded region with intersections or roundabouts, but also larger ones, as in the top-left corner. When comparing it with the ground truth map obtained from OpenStreetMap[4] in the bottom left panel, we clearly distinguish that some roads are not well-traced with several datapoints completely out with respect to the road map due to the poor quality of sampled data, as it is the case in the top right corner. However, the algorithm succeeds in proposing a robust version that do not take into account all such datapoints.

The proposed algorithm offers the possibility to estimate the local variance around the inferred principal graph. In the present dataset for instance, it is possible to investigate further small portions of roads where the estimated $\sigma_k^{(t)}$ is high or has a broad distribution to spot roads that are either wider or noisier than others. The bottom right panel of Fig. 4.10 reports the distribution of variances for nodes standing in rectangular regions of the top panel. When using variances of Gaussian clusters as a proxy of the road width, we conclude that main roads with multiple lanes like those in the green or purple rectangles have larger sampling standard deviation, by up to a factor of two, than the low-level road in the blue rectangle. These estimates of road widths however, depend on the quality of the sampling and some roads can appear wider than they actually are because of spurious points artificially increasing the estimate locally or at the extremities of some branches of the graph. This can be seen when inspecting the variance of the $\sigma_k^{(t)}$ distributions where the purple one is much larger than the others because of the local noise in the sampling. Finally, the proposed cycling graph captures correctly some topological features of the road network, such as the most prominent roundabouts or road intersections highlighted in shaded reddish areas on the two top panels of Fig. 4.10. The representation of such features would not be permitted by tree-based topologies.

4.8 Summary and prospects

In the present chapter, we merged concepts from mixture models and graph theory to build a new framework for principal curve estimation that alleviates some drawbacks of the existing definitions.

In summary, the proposed formulation:

1. Provides a natural connection of the points through a graph structure, as opposed to several ridge finders, that only results in a set of projected points stands linked *a posteriori* such as in the SCMS algorithm.

2. Embeds a built-in robustness to outliers of the pattern. Most of the algorithms require a pre-processing step [Stanford & Raftery, 2000; Ozertem & Erdogmus, 2011] to remove outliers exhibited by real-world datasets. Here, this difficulty is circumvented by directly handling outliers in the mixture model through an added uniform component.

[4] *https://www.openstreetmap.org/*

3. Is able to learn the local size of the one-dimensional structure through the variance of the Gaussian clusters making it a heteroscedastic version of the SimplePPT algorithm proposed in Mao et al. [2015, 2017]. In particular, we showed that the same update equations for nodes can be derived as a particular case of the proposed formalism.

4. Can work with any prior topology given by a graph construction. We additionally propose a graph that includes cycles of various scales and compare it with the mathematically well-defined framework of persistent homology [Edelsbrunner et al., 2002; Kurlin, 2015].

5. Relies on simple, fast and well-established procedures such as the Kruskal and EM algorithms making it suitable for handling a large number of datapoints.

By combining mixture models and graph theory, we have been able to take advantage of the probabilistic formulation to get a description of the local size of the sampling around the approximated one-dimensional manifold. One of the interesting perspectives of this work will be to extend the formalism for the detection of d-dimensional structures in $R^{>d}$ datasets with for instance applications to the detection of walls in the cosmic web that are 2D structures embedded in a 3D space. Leads of extensions in that direction may be found in the use of more general graph topologies, like those based on k-nearest-neighbours, and by relaxing the spherical assumption for Gaussian clusters used in the present version of the algorithm. These extensions may also trigger the need to adapt the cost function to constrain the full surface/volume of the data rather than solely the nodes of the graph.

The overall optimisation scheme of the proposed algorithm may be another avenue of future work. We have shown that it exhibits a stable behaviour for a large range of model's hyper-parameters and initialisation giving freedom to the user to retrieve the underlying patterns without requiring a cautious fine-tuning. The current optimisation, based on the tuning of free regularisation parameters, makes the extraction of a satisfactory pattern dependent on multiple trials and heuristics. Recent literature [Fischetti & Stringher, 2019] propose ways of alleviating these limitations by resorting to simulated annealing (also discussed in 4.3.1).

Finally, the proposed algorithm was devised for the specific task of detecting the filamentary pattern of the cosmic web. Even though these data show a high level of complexity with different local sampling densities, loops of various sizes, features of different scales, the proposed algorithm requires no pre-processing to remove spurious measurements and provides a global description of the pattern including properties such as local length or width. In Chapter 5, we will see how this algorithm was compared to other web-finder methods and used to build a graph representation of the cosmic web through the distribution of galaxies in simulations.

Part III

Analysis of the Cosmic Web pattern

Chapter 5.

The principal graph of the Cosmic Web

"A quoi bon faire des maths appliquées si c'est pour ne pas les appliquer ?"
V. Bonjean

In this chapter, we present how the algorithm developed in Chapter 4 can be used for the identification of the principal graph of the cosmic web. It partly exposes results from Bonnaire et al. [2020] and some from Gouin et al. [2021]. We show how the graph can be post-processed to identify the filamentary pattern as a whole together with an estimate of the positional uncertainty of the spine, but also to extract individual filaments and their properties for carrying statistical analyses. We apply and compare the filaments obtained from the galaxies of several large-scale hydrodynamical simulations, namely EAGLE, IllustrisTNG and Magneticum and then focus on the role of filaments in shaping galaxy clusters in simulations.

5.1 Context and motivations

We have stressed in Sect. 2.3 the primordial role of the cosmic web filamentary structure and Sect. 2.4 focused on the inherent challenges of carrying out an accurate detection of cosmic web elements. The sparse sampling of galaxies in observation is also an additional difficulty

that comes with its own systematic effects and limitations (like the completeness and spatial coverage of the survey). It is however essential to be able to extract cosmic web environments in both data and simulations to get a better understanding of how the large-scale structures formed and evolved through time. Because they are easier to detect, the nodes were the first to draw researchers' attention which led to a rather clear picture of their average properties like galaxy content, density profiles and gas content based on several catalogues [e.g. Abell et al., 1989; Rykoff et al., 2014] but also on their information content in terms of cosmological parameters [Yoshida et al., 2000; Peter et al., 2013]. Filaments on the other hand received an attention only more lately, permitted by the evolution of wide spectroscopic surveys [such as York et al., 2000; Colless et al., 2001] and by the elaboration of sophisticated tools to identify components other than nodes in the cosmic web. Filaments, acting like cosmic highways linking together large overdense nodes, play a key role in the dynamics of the Universe but also in shaping the evolution of galaxies that has been studied in both data and simulations. While the local density-mass relation between a tracer of the matter distribution and its environment is now well-established, making galaxies (or halos in simulations) more massive in dense environments like clusters [Dressler, 1980; Kauffmann et al., 2004; Baldry et al., 2006], the relationship between other properties of both tracers and large-scale structures is still under investigation. Many recent studies for instance point out that galaxies are less efficient in forming stars in the core of filaments than in their periphery [Alpaslan et al., 2016; Malavasi et al., 2017; Kraljic et al., 2018; Bonjean et al., 2020] suggesting a particular process occurring when a galaxy enters a filament. This result also matches a local density influence as filaments have been shown denser in their inner regions [Bonjean et al., 2020; Galárraga-Espinosa et al., 2020]. Beyond density, the local orientation of filaments has also an impact on the spin of galaxies reported by many works in both data and simulations [Hahn et al., 2007; Codis et al., 2012, 2015; Ganeshaiah Veena et al., 2018, 2019; Kraljic et al., 2020].

All these observations and conclusions allowing to push forward our understanding of the complex statistical interplay between matter tracers, like galaxies, and the large-scale structures predominantly made of nodes and filaments have only been possible thanks to the detection of such objects in the galaxy distribution [e.g. Abell et al., 1989; Rykoff et al., 2014; Tempel et al., 2014; Chen et al., 2016; Rost et al., 2020; Malavasi et al., 2020b; Duque et al., 2021]. It is however crucial, given the different results provided by many methods to identify filaments [Libeskind et al., 2017], to corroborate the correlations observed in various datasets and by several methodologies in order to avoid any bias induced by the definition of filaments and by the choice of the algorithm.

In this chapter, we first present T-ReX, a filament detection method based on the mathematical formulation exposed in Chapter 4. In the cosmological context of the detection of the filamentary pattern drawn from subhalos or galaxies, we begin by extracting the overall spine of the filamentary structure together with its spatial uncertainty. We assess the robustness of the method to sparse sampling, in addition to provide comparisons with other web-finders applied on same datasets. We then explain how individual filaments can be defined from the learnt principal graph to build catalogues gathering their positions with some geometrical properties like their lengths, curvatures, local and averaged widths. The proposed algorithm is also able to associate to each input datapoint (galaxy or halo) a probability to reside in a given filament. This association is successfully compared to the one provided by a physical rather than geometrical web finder (see Sect. 2.4.2) applied on the full dark matter distribution. T-ReX is able to retrieve 80% of the galaxies as belonging similarly to filaments, even though with a much

smaller number density of tracers as input. We then carry out a comparison of the properties of filaments identified individually from different samples of galaxies in several simulations that are EAGLE, IllustrisTNG and Magneticum. Finally, T-ReX is applied to the large-scale simulation of galaxies IllustrisTNG and is used to study the interplay between the nodes of the cosmic web and their surrounding filamentary pattern. In particular, we focus on the correlations between the connectivity of galaxy clusters and their physical and dynamical properties.

5.2 Filamentary pattern detection

In this first section, we focus on the detection of the filamentary pattern as a whole and explain how the algorithm presented in Chapter 4 can be used through multiple realisations based on random sub-samplings of the input dataset to obtain a picture of the filamentary structure of a dataset together with a spatial uncertainty.

5.2.1 T-ReX: Tree-based Ridge eXtractor

The minimum spanning tree (MST) is a graph with a tree topology that is linking all the input points together with the minimum total distance (see Sect. 4.2 for a more formal introduction). It has a long history in cosmology and has been the first method to exhibit the filamentary structure of the cosmic web in early galaxy distributions mapped in the sky by Barrow et al. [1985]. Among its most appealing features are its unicity, the absence of free-parameter, its scale-invariance but also its easiness to compute since it relies on fast and well-established methods like the Kruskal algorithm. All these features make the MST a well-suited tool for the study of the large-scale distribution of matter in the Universe as proposed by numerous papers on the topic [Pearson & Coles, 1995; Bhavsar & Splinter, 1996; Colberg, 2007; Park & Lee, 2009; Alpaslan et al., 2014b,a, 2016; Naidoo et al., 2020; Bonnaire et al., 2020; Pereyra et al., 2020b,a]. However, one of the main drawbacks of the MST formulation is that, by linking all datapoints, it results in a locally spiky geometry, as can be seen in the right panel of Fig. 5.1, that several works recently propose to alleviate. Pereyra et al. [2020b] for instance used a post-processing of the MST in which an *ad hoc* selection of tree branches are smoothed by means of a B-spline interpolation while Bonnaire et al. [2020, 2021b] and related Chapter 4 present a way of incorporating the MST in an optimisation scheme embedding the graph as a prior. In the cosmological context, the algorithm established in Chapter 4 is entitled T-ReX standing for "Tree-based Ridge eXtractor". Let us recall the main components of the model, the optimisation problem solved by the algorithm and how these can be interpreted in the context of cosmological datasets.

In the mixture model framework, the probability that a matter tracer stands at a position $\boldsymbol{x}_i$ is given by

$$p(\boldsymbol{x}_i \,|\, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\theta}_k) + \alpha\rho(\boldsymbol{x}_i), \tag{5.1}$$

where $\boldsymbol{\Theta}$ is the set of model's parameters, $\mathcal{N}(\boldsymbol{x}_i, \boldsymbol{\theta}_k)$ is the Gaussian probability density function centred on $\boldsymbol{x}_i - \boldsymbol{\mu}_k$ with variance σ_k^2, $\rho(\boldsymbol{x}_i)$ is the uniform distribution over the convex hull of the point-cloud distribution. The first term of this equation models the graph nodes standing on the filamentary pattern, paving the distribution while the second one handles

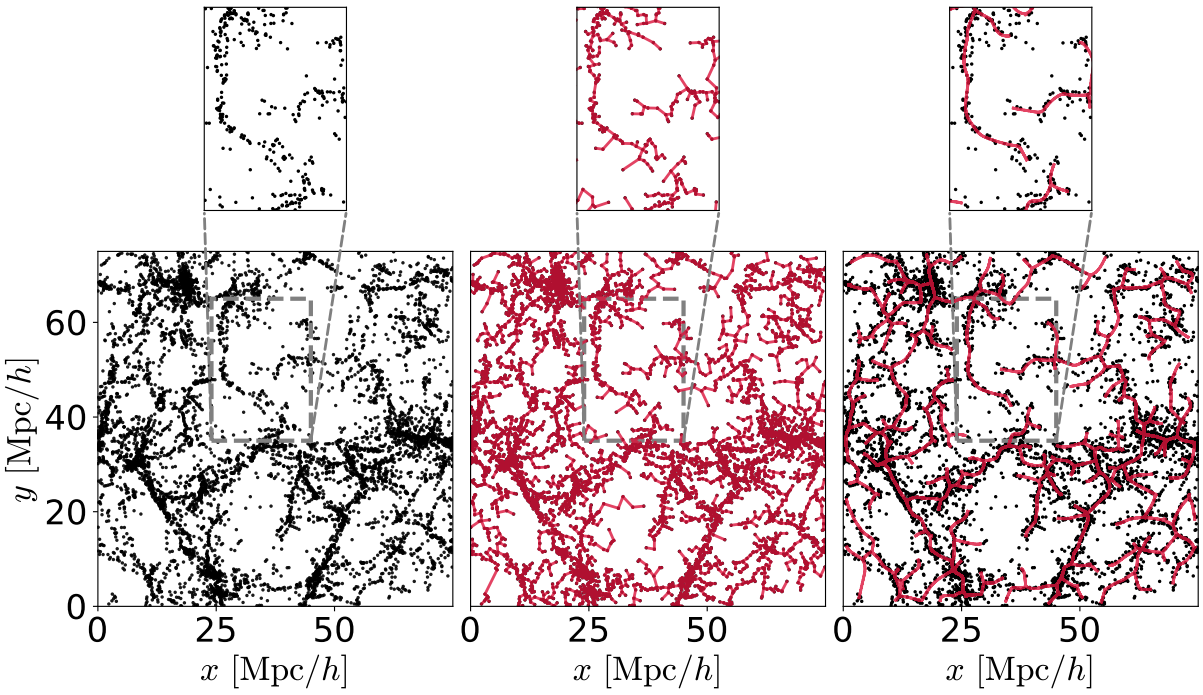

Fig. 5.1. Differences between the classical MST (middle panel) and the optimised one obtained from the T-ReX algorithm (right panel) with $\lambda_\mu = 5, \lambda_\sigma = 0, \lambda_\pi = 0$, $\sigma_0 = 1\,\mathrm{Mpc}/h$ and under a fixed-variance scheme. Both algorithms are applied on the 2D subhalos distribution of the Illustris simulation shown in the left panel.

datapoints that are not part of it, standing outside the pattern, in voids or walls embodied by a uniform distribution. We have seen in Sect. 4.3 that a double maximisation procedure can be used to estimate the smooth graph solving in particular

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\mathrm{argmax}} - \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}\frac{\|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2}{\sigma_k^2}}_{\text{data fidelity term}} - \underbrace{\lambda_\mu \sum_{i=1}^{K}\sum_{j=1}^{K} a_{ij}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2}_{\text{graph smoothness term}} + \ldots\,^{1}, \tag{5.2}$$

where $\boldsymbol{A}$ encodes the graph adjacency and λ_μ is the strength of the prior. More precisely, λ_μ regulates the trade-off between the data fidelity term representing how much the graph should pass close to the galaxies and how much it should be short and smooth. The T-ReX algorithm 1 exposed in Chapter 4 can be used to obtain the *best possible*[2] graph given an initialisation and a set of hyper-parameters λ_μ, λ_σ and λ_π. The application of T-ReX is shown in right panel of Fig. 5.1 based on the input datapoints shown in the left one. The obtained graph has a smooth adaptive behaviour, especially visible in the zoom regions when compared to the crude MST shown in the middle panel that is reaching all datapoints.

To summarise, the T-ReX algorithm keeps the fundamental idea of the MST-based approach by representing the topology of the cosmic web as an interconnected network based on discrete representation of tracers but provides additional features not allowed by previous definitions that are: (i) a smooth version of the MST incorporated in the core of the formalism; (ii) a handling of outliers to represent datapoints outside the filamentary pattern; (iii) a description

[1] Additional terms include the priors on uniform weights with strength λ_π, smooth evolution of variances along the graph with strength λ_σ, uniform background noise and weights π_k.

[2] In the sense of the local maximum *a posteriori*.

of the local width of filaments through variances σ_k^2 learnt during the optimisation; and (iv) the probabilistic association of tracers to individual filaments (see Sect. 5.3.3 for a more detailed presentation of this feature).

To initialise graph nodes in the algorithm, we use in the entire section a pruned version of the initial MST using all the input datapoints. It consists in a simple denoising operation cutting all the nodes standing in branches of the tree at a level l. Strictly speaking, we iteratively remove all nodes of degree one in the graph structure, similarly to Barrow et al. [1985] or Colberg et al. [2005]. By doing so, we remove the most spurious parts of the structure corresponding to nodes that are more likely to be found in physically irrelevant regions for the underlying pattern (i.e. underdense regions). This approach is iterative, meaning that a pruning at level l removes iteratively nodes of degree 1 a total of l times. To give a representative image of this procedure, it acts like the iterative peeling of an onion, attributing to each node a depth in terms of layers to peel before we reach it and starting from extremities [Hébert-Dufresne et al., 2016]. Note that, even though we prune the tree to initialise the number K and positions $\boldsymbol{\mu}^{(0)}$ of graph nodes, they are still used in the full optimisation and contribute to the shaping of the final smooth version of the graph.

Previous MST-based methods usually perform, in addition to this pruning, a removal of all edges above a given physical length. In our case, this operation is not applied to avoid the introduction of a new parameter that is not easy to tune, but also because we argue that all connections, even 'long' ones, can provide information about the underlying structure. Of course, as a result, if two unconnected parts of a network are given as an input to the presented method, they will end up connected.

In summary, the T-ReX method estimates a smooth MST and relies on the set of parameters that are: l the pruning level to initialise the graph which implicitly defines K the number of nodes and their initial position, $\mathcal{Y} = (\lambda_\mu, \lambda_\sigma, \lambda_\pi)$ the prior strengths and $\sigma_k^{(0)}$, the standard deviations of Gaussian graph nodes[3].

5.2.2 Filamentary pattern extraction from Illustris subhalos

In the rest of the section, we make use of the Illustris simulation outputs[4] [Vogelsberger et al., 2014]. It consists in a set of large-scale hydrodynamical simulations with different resolutions in which an initial set of particles distributed over a 75 Mpc/h box is evolved forward in time from high redshift to $z = 0$. From the resulting distribution at $z = 0$, halos of dark matter are identified using a Friend-of-Friend algorithm [FoF, Davis et al., 1985]. To mimic a galaxy survey, we consider structures inside halos, the subhalos, identified with the Subfind algorithm [Springel et al., 2008] and provided within the Illustris package. In the left panel of Fig. 5.1 is shown a thin 5 Mpc/h slice of the subhalo distribution obtained from the Illustris-3 simulation.

As previously mentioned, the MST highlights one particular path linking datapoints together but does not provide any idea of its uncertainty or reliability. It is also restricted by its tree topology that has no loops and cannot represent holes but only connected components in the cosmic web. Both of these issues can be overcome by introducing a robust representation that takes into account the variations in the input distribution. To do so, we build B different samples $\{\boldsymbol{X}_b\}_{b=1}^B$ from the initial one $\boldsymbol{X}$ and compute the regularised MST for each of them in a similar fashion as in bootstrap approaches, also evoked in Sect. 4.4.2.

From the B realisations of regularised MSTs, one can construct a map $\boldsymbol{I}$ characterising the

[3] Which appears as a parameter in the fixed-variance scheme, but as an initialisation when the variance is learnt during the optimisation.

[4] *http://www.illustris-project.org/data/*

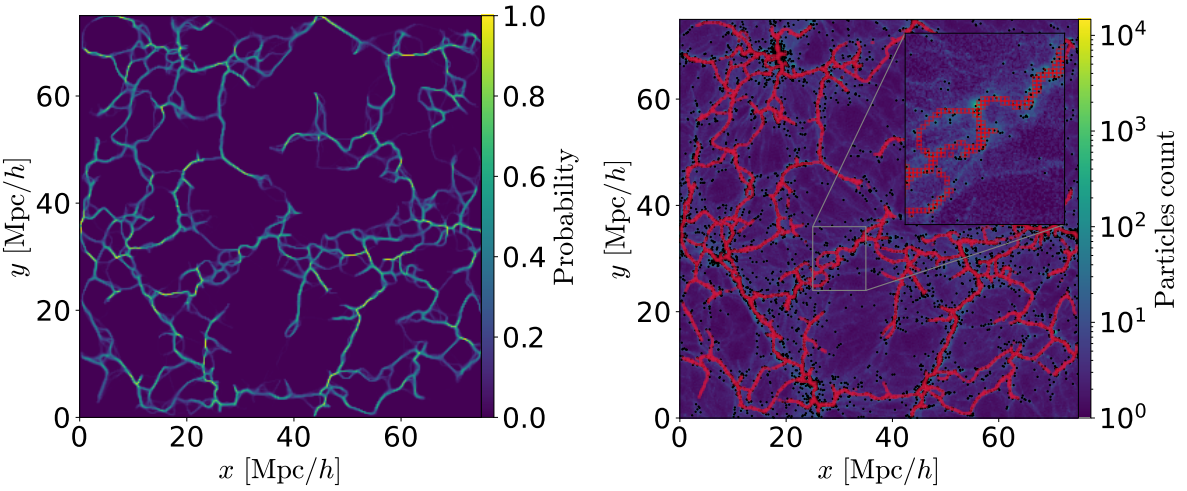

Fig. 5.2. *(left)* Probability map $\boldsymbol{I}$ obtained from subhalos displayed in the left panel of Fig. 5.1 with $B = 100$ and $N_b = 0.75$. The resolution of the probability map is $250\text{Kpc}/h$. *(right)* Superlevel set $\boldsymbol{\Gamma}_{0.25}(\boldsymbol{I})$ (red squares) overplotted on the DM distribution together with subhalos (black dots).

probability, in a frequentist meaning, of a position $\boldsymbol{x}_\text{g}$ on a grid to be crossed by a realisation of a tree:

$$i(x) = \frac{1}{B}\sum_{b=1}^{B} 1_{\boldsymbol{H}_b(\boldsymbol{x}_\text{g})=1}, \tag{5.3}$$

where 1_A is the indicator function and $\boldsymbol{H}_b$ is the binary histogram obtained from the graph nodes $\boldsymbol{\mu}_b$. The random nature of $\boldsymbol{I}$ thus comes from the random sub-sampling of $\boldsymbol{X}$ and not from Algorithm 1 that is a deterministic optimisation step.

The left panel of Fig. 5.2 shows a probability map obtained from a 5 Mpc$/h$ depth slice in which the intensity of each pixel corresponds to the frequency that an edge of the MST crossed it. This way, we quantify the reliability of the various paths in the input domain. In practice, to build $\boldsymbol{I}(\boldsymbol{x}_\text{g})$, we use both the graph nodes $\boldsymbol{\mu}_b$ and the set of edges linking vertices encoded that contains information on the paths used and consequently should be taken into account in the final distribution. Edges are thus sampled and counted in the computation of $\boldsymbol{H}_b$ for Eq. (5.3). In what follows, we may refer to a quantity called the superlevel set of those maps defined as $\boldsymbol{\Gamma}_p(\boldsymbol{I}) = \{\boldsymbol{x}_\text{g} \mid i(\boldsymbol{x}_\text{g}) \geq p\}$. Those sets are used to threshold the probability maps and keep only regions with a probability higher than p. We can see, in the right panel of Fig. 5.2, that the highly probable regions of the map are fitting what one would expect for the underlying distribution while the overlap of the superlevel set $\boldsymbol{\Gamma}_{0.25}(\boldsymbol{I})$ with the DM distribution allows us to see that high probability paths (above 0.25 in this case) are tracing the most prominent part of the network. It is worth noting that the agreement is particularly interesting given that the input of the algorithm are subhalos and not DM particles. The zoomed-in region clearly shows that small scales are also recovered where high probability paths follow the ridge in the DM distribution.

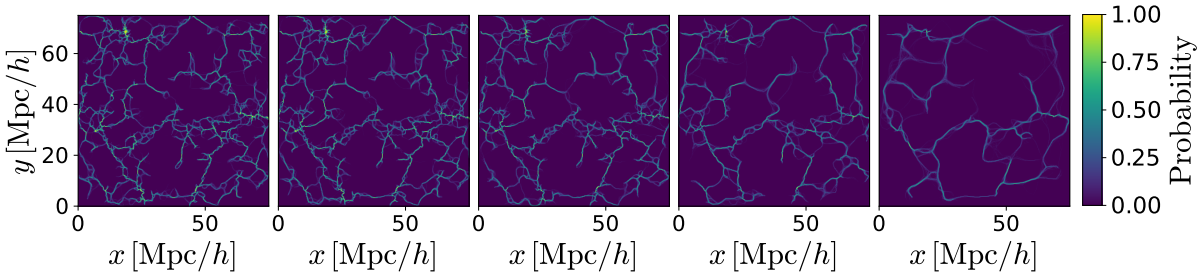

Fig. 5.3. Probability maps with increasing mass threshold M^{cut}. From left to right, $M^{\text{cut}} = \{0, 0.85, 1.35, 3.22, 11\} \times 10^{10} M_\odot/h$ corresponding, respectively, to 100%, 83%, 60%, 31%, and 10% of the total subhalos in the slice.

5.2.3 Performance evaluation

Robustness to sparse samplings

In order to assess the robustness of the method against the datapoint sampling density used for the detection of ridges, we reduce the number of subhalos in the initial dataset by keeping only those with a mass $M \geq M^{\text{cut}}$. In practice, we investigate how the original filamentary map is spatially close to the recovered ones when M^{cut} varies. Figure 5.3 shows probability maps obtained for increasing values of M^{cut} leading to sparser and sparser input, namely 100%, 83%, 60%, 31% and 10% of the initial subhalos in the slice respectively corresponding to $M^{\text{cut}} = \{0, 0.85, 1.35, 3.22, 11\} \times 10^{10} M_\odot/h$. Visually, probability maps show a nice stability, even when the sparsity is high: patterns are pretty much the same when we keep at least 60% of the most massive objects hence recovering the essential part of the structure.

In Fig. 5.4 is exhibited the spatial proximity between the different maps by representing, for each $\boldsymbol{I}_J$, where J denotes the fraction of galaxies we kept to compute the map, the cumulative distribution of $\{d^J_{\boldsymbol{x}}\}_{\boldsymbol{x} \in \boldsymbol{\Gamma}_{0.25}(\boldsymbol{I}_{100})}$ defined, for a position $\boldsymbol{x}$ in the set $\boldsymbol{\Gamma}_{0.25}(\boldsymbol{I}_{100})$, as

$$d^J_{\boldsymbol{x}} = \min_{\boldsymbol{x}' \in \boldsymbol{\Gamma}_{0.25}(\boldsymbol{I}_J)} \|\boldsymbol{x} - \boldsymbol{x}'\|_2. \tag{5.4}$$

Hence $d^J_{\boldsymbol{x}}$ corresponds to the closest distance from a position $\boldsymbol{x}$ in the original skeleton obtained by keeping all subhalos, namely $\boldsymbol{\Gamma}_{0.25}(\boldsymbol{I}_{100})$, to a given thresholded map $\boldsymbol{\Gamma}_{0.25}(\boldsymbol{I}_J)$. This way, the distribution of $d^J_{\boldsymbol{x}}$ measures how far the original pattern is from the one obtained with $J\%$ of the datapoints.

In more than 95% of the cases, the original pattern finds a closest point in the 83% and the 60% maps at less than 1.8 Mpc$/h$, showing that structures found in the three maps are spatially close and about the thickness of typical filaments [Cautun et al., 2014]. When M^{cut} increases, the filamentary pattern traces the most prominent parts of the structure with a loss of some small scales and hence highlights coarser and coarser structures. Even though the pattern is rough with only 31% of the datapoints used, we still observe a correlation with previous maps highlighting coherent structures with 90% of the original pattern being retrieved at less than 3 Mpc$/h$. As expected, an unrealistic scenario where we use only 10% of the datapoints associated with the most massive subhalos degrades the reconstruction of the filamentary pattern. Yet, the recovered structures show a coarse but coherent connectivity between regions. This illustrates the ability of T-ReX to recover the underlying structure with high stability with respect to missing information in the input distribution of datapoints.

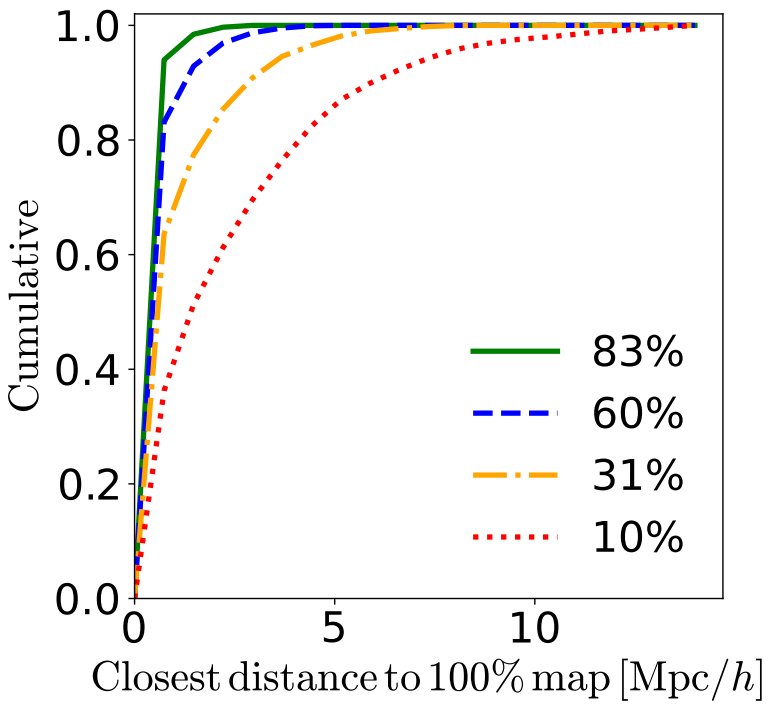

Fig. 5.4. Cumulative distribution of distances $\{d_{\boldsymbol{x}}^J\}$ between positions of the binary maps $\boldsymbol{\Gamma}_{0.25}(\boldsymbol{I}_{100})$ obtained with increasing mass threshold $M^{\rm cut}$ to the one with $J\%$ of the datapoints. $M^{\rm cut} = \{0.85, 1.35, 3.22, 11\} \times 10^{10} M_\odot/h$ leading respectively to 83%, 60%, 31% and 10% of the total number of subhalos in the slice.

Comparison with other algorithms

In this section, we apply T-ReX on the 3D distribution of halos obtained from a 200 Mpc/h Gadget-2 simulation[5] and compare our results with some other existing procedures. These have also been run on the same dataset in a review by Libeskind et al. [2017] to propose a comparison of the main existing procedures to classify elements of the cosmic web using either dark matter particles or halos as inputs. Although the review considers a dozen of different methods, we focus the comparison on three procedures, namely Nexus+, DisPerSE and Bisous, so that we have a broad set of different approaches using respectively scale-space representation, topological considerations, or stochastic approach to recover the filamentary pattern. Nexus+ [Cautun et al., 2013] is a classification algorithm inspired by image processing and based on filtering techniques leading to state-of-the-art environment classification able to identify clusters, filaments and walls. The main idea is to assume that the local morphology of the density field fully encodes the environmental information. Eigenvalues of the Hessian of the density field are thus used to compute an environmental signature in each voxel of the smoothed field. The key idea is to compute this signature for a set of smoothed fields with a log-Gaussian filter over a range of different scales to highlight structures of different sizes. Physically motivated criteria are then used to threshold signature values and attribute a classification to each volume element. Bisous [Stoica et al., 2007] is a publicly available[6] stochastic method based on halo positions that identifies the filamentary structure using a set of random parametric cylinders. Filaments are modelled as aligned and contiguous small cylinders of a given size in the galaxy distribution. The Bisous model generates two maps

[5] *http://data.aip.de/tracingthecosmicweb/doi:10.17876/data/2017_1*
[6] *https://www.ascl.net/1512.008*

allowing to extract filaments spine; one characterising the probability to find a filament at a given position called the visit map and an other one corresponding to the filament orientation field. This way, spines are defined as dense regions and are aligned with the axis of the different cylinders.

Note that not only these methods have very different mathematical definitions for what they all call clusters, filaments, and walls, but they also have been run with different input, using either DM particles or halos. We applied T-ReX to the full halo distribution of the 3D simulated box (281 465 halos in total) and built a $100 \times 100 \times 100$ grid map like other methods. For T-ReX, this means that the final probability map is computed over a 100^3 grid in which all visited voxels are considered part of the filamentary structure. As T-ReX is using 1D objects (segments of the smooth graph) sampled over the input space, it is preferable, for illustration and comparison, to give its filamentary pattern a 'thickness' by smoothing the obtained probability map. Whenever a voxel is classified as part of the filamentary structure, a smoothing is thus performed over its 26 direct neighbours. In what follows, we call this version T-ReX$_{\text{s}}$ while the original result is referred to as T-ReX$_{\text{us}}$.

For illustration, following Libeskind et al. [2017], we show in Fig. 5.5 the results of the classification provided by each method for a 2 Mpc/h depth slice from which FoF halos were extracted (top left panel of Fig. 5.5). We note that all methods have been run over the full 3D cube and this is a projected slice of the detection. It is also worth noting that T-ReX identifies the filamentary pattern as a whole and does not classify the environment into clusters, filaments and walls as Nexus+ and DisPerSE do. To perform the comparison, we must look at the full pattern provided by each method and compare it with our extracted skeleton. We observe that T-ReX provides a satisfactory connectivity of the halos through the slice. In its smoothed version, it leads to thicker filaments compared to the results of Nexus+ and Bisous but thinner ones than Disperse, and retrieves most of the structures (filaments, walls, and clusters) obtained by the Nexus+ algorithm.

Even though these methods have been developed with different approaches, it is interesting to see whether they agree or not in the detection of the filamentary pattern. To do so in a quantitative way, we could use the proximity measurement of Eq. (5.4) but as the resulting patterns are presented on a 2 Mpc/h grid, the distance between them would not be accurate. Hence, we introduce a similarity measurement as follows: considering the answers provided by two detection methods, H_1 and H_2, such that $H_\bullet(x) = 1$ if the cell centred at $\boldsymbol{x}_{\text{g}}$ is part of the filamentary structure and 0 otherwise, the similarity measurement is defined as:

$$\mathcal{S}(H_1, H_2) = \frac{|H_1 \cap H_2|}{|H_1|}, \tag{5.5}$$

where $|H_i|$ denotes the cardinal of H_i defined as $\sum_{\boldsymbol{x}} 1_{H_i(\boldsymbol{x})=1}$ and $|H_1 \cap H_2|$ is the cardinal of the intersection between H_1 and H_2 detections defined as $\sum_{\boldsymbol{x}} 1_{H_1(\boldsymbol{x})=1} 1_{H_2(\boldsymbol{x})=1}$. Hence, $\mathcal{S}(H_1, H_2)$ measures the proportion of H_1 detections that are contained in H_2 and is thus asymmetric. In other words, if we consider H_2 as a reference, $\mathcal{S}(H_1, H_2)$ represents the proportion of true detections provided by H_1. Of course, such a simple metric does not provide the full information on the similarity between the considered patterns. This measure must then be used in tandem with others, or with visual inspection, as we have done here.

Table 5.1 shows the similarity indices between all considered methods for the entire 3D cube. We observe that 85% of the detections provided by the unsmoothed version T-ReX$_{\text{us}}$ are contained in the Nexus+ skeleton and 81% of the Nexus+ detections are found by the smoothed version of T-ReX. This indicates that the smoothed version of T-ReX contains a large part of the Nexus+ skeleton but with a larger amount of the volume detected, explained by the smoothing

Table 5.1. Similarity index $\mathcal{S}(H_1, H_2)$ as defined in eq. (5.5) between the considered methods applied on the entire 3D cube. T-ReX$_{\rm us}$ refers to the unsmoothed version of the detection while T-ReX$_{\rm s}$ refers to the smoothed one over the 26 neighboring voxels.

H_1 \ H_2	T-ReX$_{\rm us}$	T-ReX$_{\rm s}$	Nexus+	DisPerSE	Bisous
T-ReX$_{\rm us}$	1	1	0.85	0.62	0.37
T-ReX$_{\rm s}$	0.48	1	0.62	0.62	0.24
Nexus+	0.53	0.81	1	0.62	0.30
DisPerSE	0.22	0.46	0.35	1	0.12
Bisous	0.66	0.87	0.86	0.62	1

leading to a thicker filamentary pattern. The same tendency is observed concerning Bisous for which the detections are mostly contained in other skeletons (last row of Table 5.1) but not reciprocally (last column). This is due to the sparse and unconnected detection provided by the Bisous method. The thick skeleton of DisPerSE also tends to contain a large fraction of other skeletons (fourth column of Table 5.1) but it fills so much volume that it is not contained in the latter (fourth line).

5.3 Identification of individual filaments

In the previous section, we have shown how the detection of the overall filamentary structure can be done with the graph-based algorithm and that the identification agrees with some other definitions. The gridded output representing the probability of a cell belonging to the pattern can be of interest for using it as a mask in large scale cross-correlation analyses. However, it is not trivial to define individual objects that are filaments from such an output. One way to achieve this is to use and post-process the principal graph learnt from the application of the T-ReX algorithm on the full dataset made of the N tracers at positions $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$.

5.3.1 A graph-based definition for filaments

In a graph, one can associate to each node i a degree $\deg(i)$ corresponding to the number of direct neighbours (see Sect. 4.2). Naturally, three types of nodes can be distinguished: i) Extremity nodes with $\deg(i) = 1$; ii) Junction nodes with $\deg(i) = 2$; and iii) Bifurcation nodes with $\deg(i) > 2$. One simple way to define filaments based on the MST is hence to use branches of the graph, a branch being the set of connected edges linking an extremity node to a bifurcation or a bifurcation to a bifurcation as illustrated in the left panel of Fig 5.6.

5.3.2 Characteristics of individual filaments

At this stage, we consider an individual filament as a set of M nodes which are actually Gaussian components and a subset of the K centres $\boldsymbol{\mu}$. Centres composing a filament are found at positions $\mathcal{F} = \{\boldsymbol{f}_m\}_{m=1...M}$ with $\mathcal{F} \in \mathbb{R}^{M\times D}$, a subset of the set of graph nodes $\boldsymbol{\mu} \in \mathbb{R}^{K\times D}$ where $\boldsymbol{f}_k \in \mathbb{R}^D$ and $M-1$ edges forming a branch of the resulting regularised graph provided by T-ReX. To each edge linking two nodes $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ is associated a weight corresponding to the

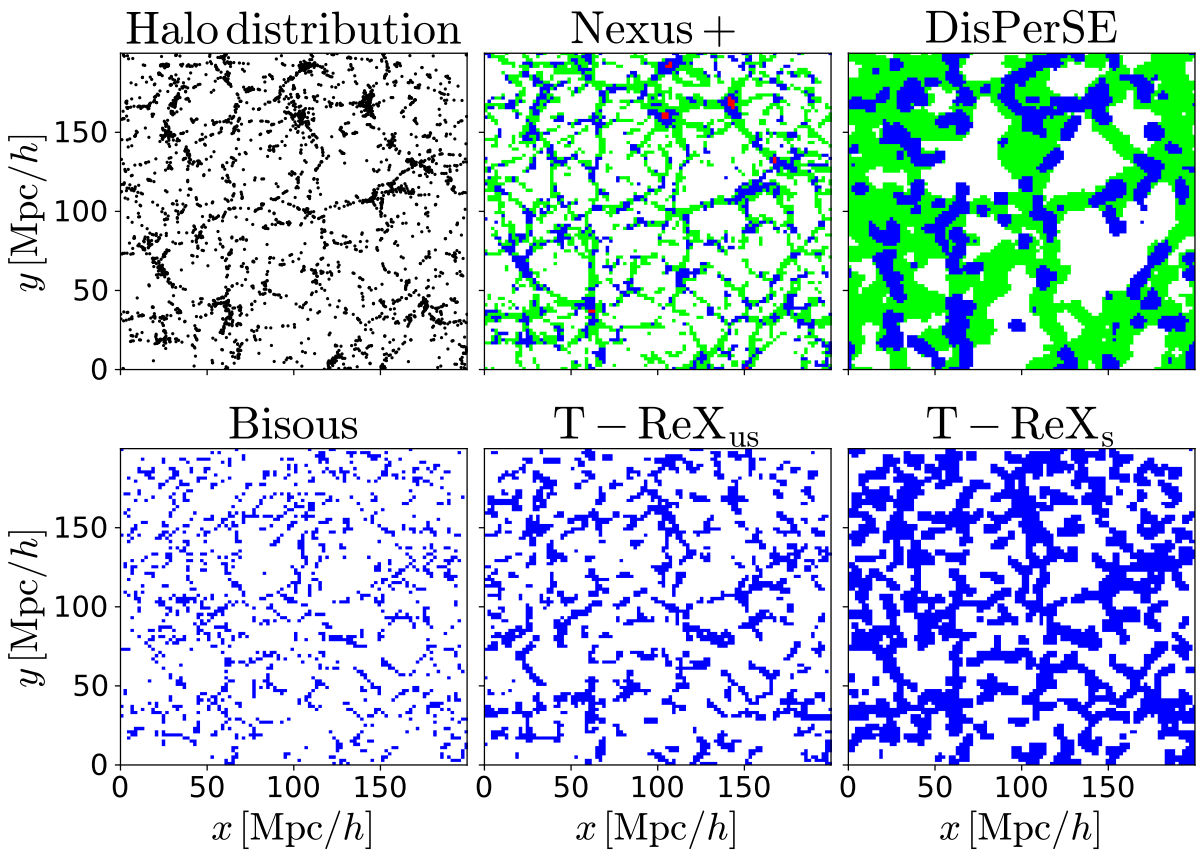

Fig. 5.5. Identification results provided by four detection methods on a randomly chosen 2 Mpc/h depth slice of the full 3D detection for each method. Green pixels are walls, blue are filaments, red are clusters and white are voids or unclassified regions.

Euclidean distance $w_{ij} = \|\boldsymbol{f}_i - \boldsymbol{f}_j\|_2$. Let us also define $\mathcal{R}$ the ridge of the filament corresponding to the one dimensional piecewise linear line connecting nodes. Filaments characteristics that can be extracted from the graph are listed and defined below.

Geodesic length. The geodesic length L of the filament is defined as the sum of all edge weights forming the ridge,

$$L = \sum_{i,j=1}^{M} w_{ij}. \tag{5.6}$$

Inspecting the left panel of Fig. 5.6, the length of a filament is for instance given by summing all edges of same colour.

Curvature. The simple proxy we use to describe the shape of the filament ridge is how much the geodesic length differ from the Euclidean length. The curvature is thus defined as

$$\gamma = 1 - \frac{\|\boldsymbol{f}_1 - \boldsymbol{f}_M\|_2}{L}, \tag{5.7}$$

with $\boldsymbol{f}_1$ and $\boldsymbol{f}_M$ the two extremities of the branch, i.e. a node of degree 1 or higher than 2, as illustrated in left panel of Fig. 5.6. Consequently, $\gamma \in [0; 1]$ and the closest it is to one, the more the filament deviates from a straight line.

Local width. The local width of a filament is defined by the set of variances of the Gaussian component composing the graph nodes paving its ridge, $\{\sigma_k^2\}_{k=1}^M$. A one-point summary of this distribution can be the average of standard deviations leading to an estimate of the average radius of the filament, noted

$$r = \sum_{m=1}^{M} 3\sigma_m / M, \tag{5.8}$$

such that locally, the radius is chosen to be represented by three times the standard deviation of the Gaussian cluster paving the ridge. Note that this definition of radius for filament takes into account the spatial variations in the extension of the filament through the local variances of Gaussian graph nodes and is not a fixed-length definition.

Radial distance. It is possible to define for each galaxy in the catalog a projected distance to the ridge $\mathcal{R}$ of the filament. This distance is what we call the *radial distance* and corresponds, for a datapoint at position $\boldsymbol{x}_i$, to

$$\Delta_i = d(\boldsymbol{x}_i, \mathcal{R}) = \min_{\boldsymbol{x}' \in \mathcal{R}} \|\boldsymbol{x}_i - \boldsymbol{x}'\|_2. \tag{5.9}$$

This measurement can be used to provide either the full distribution of datapoint distances around the filament or as a proxy to the radial extension of the filament using the mean of the distribution, that we note $\bar{\Delta}$, alternative to the previous definition of r.

Positional uncertainty. In Sect. 5.2.2, we used multiple realisations of the algorithm to obtain an estimate of the positional uncertainty of the ridge. We can nonetheless still obtain an idea of the uncertainty at the graph nodes level. Assuming B regularised graph realisations, we can associate to each node k of the full graph an uncertainty in the form of a confidence sphere $B(\boldsymbol{f}_k, h_k)$ centred at position $\boldsymbol{f}_k$ and with radius h_k, similarly to what was done in Chen et al. [2015]. Radii of the uncertainty spheres are provided by the mean projected distance of a given node k to the ridge for each bootstrap realisation that we can write

$$h_k^2 = \frac{1}{B} \sum_{b=1}^{B} d^2(\boldsymbol{f}_k, \mathcal{R}_b). \tag{5.10}$$

It is noteworthy to specify that, in practice, this step is realised before any filament extraction and is independant from it, operating on the full filamentary pattern. Thus, in Eq. (5.10), the k index refers to any node in the graph and R_b to the full 1d ridge of the filamentary pattern of the realisation b. Hence, the positional uncertainty of graph nodes can be defined as the union of all uncertainty spheres

$$U(p) = \bigcup_{k=1}^{K} B(\boldsymbol{f}_k, p h_k), \tag{5.11}$$

where p indicates the level of the uncertainty band, similarly to a p-sigma uncertainty.

5.3.3 Association of galaxies

T-ReX provides a probabilistic version of the regularised graph in which each node of the filament ridge is actually a Gaussian cluster with mean $\boldsymbol{\mu}_k$ and variance σ_k^2. It is also equipped with a robustness to outliers handled by an added uniform background distribution. In the EM procedure, we compute what is called the *responsibility* p_{ik} (see more details in Sect. 3.3

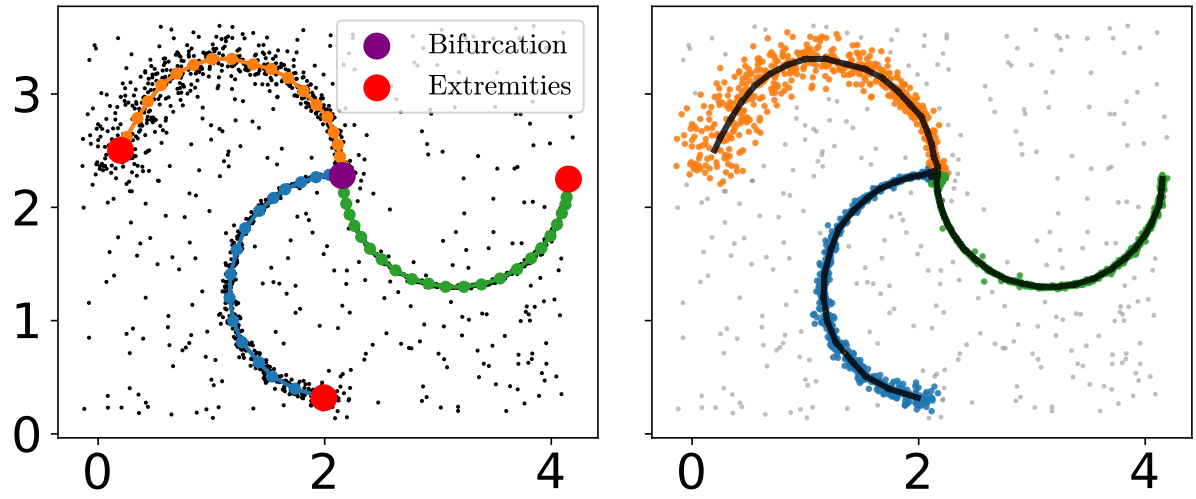

Fig. 5.6. Illustration of the definition of individual filaments on a toy dataset. *(left)* Black points are datapoints and coloured ones are those from the smooth graph structure of T-ReX. Filaments are defined as the branches linking bifurcations and extremities in the graph as discussed in Sect. 5.3.1. *(right)* Illustration of associated datapoints to each individual filament (coloured datapoints) or to the detected uniform background (grey datapoints) depending on the graph (black line) and associated variances (not represented) as exposed in Sect. 5.3.3.

and Sect. 4.3.1) characterising the probability that a given tracer $\boldsymbol{x}_i$ is drawn from a particular component of the mixture model, hence made of K Gaussian clusters and one uniform background component. Mathematically, we recall that we can write

$$p_i^{\text{bkg}} = p(z_i = K+1 \,|\, \boldsymbol{x}_i, \Theta) = \frac{\alpha\rho(\boldsymbol{x}_i)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_j) + \alpha\rho(\boldsymbol{x}_i)},$$
$$p_{ik} = p(z_i = k \,|\, \boldsymbol{x}_i, \Theta) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_i \,|\, \boldsymbol{\theta}_j) + \alpha\rho(\boldsymbol{x}_i)},$$

where $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \sigma_j^2)$ encodes the parameter of the Gaussian indexed j and z_i is the latent variable encoding the cluster attribution of datapoint $\boldsymbol{x}_i$. What hence interests us is to find the most probable value of this assignation for each datapoint. First, a datapoint can be attributed either to the overall filamentary pattern or to the background, namely if $\sum_{k=1}^{K} p_{ik} > p_i^{\text{bkg}}$ (resp. $<$), the tracer i has a higher probability of being generated by the Gaussian components (i.e. the filaments) than the background noise. We have seen that a filament can be defined as the set of graph nodes linking two extremities or bifurcations. A datapoint belonging to the filamentary structure can then be uniquely associated to the filament hosting the graph node maximising the probability p_{ik} over all nodes, namely $\hat{z}_i = \operatorname{argmax} p_{ik}$. For a given filament, we can thus derive a set of galaxies associated to graph nodes that are forming the ridge $\mathcal{R}$. Such points are shown as coloured datapoints in the right panel of Fig. 5.6 while those associated with the background are in grey.

Until now, the procedure was focused on associating a datapoint either to a filament or to the background. However, the graph can also be used to investigate a possible definition for nodes. These latter are defined in the cosmic web as dense regions linked together by filaments. In the graph, these may be represented as dense bifurcations, meeting points of several branches. To investigate further this hypothesis and to assess the overall environment classification scheme, we perform a comparison with a physical web finder, Nexus+ [Cautun et al.,

2013], applied on the EAGLE hydrodynamical simulation [Schaye et al., 2015]. The Nexus+ algorithm was run (Marius Cautun, private communication) on the dark matter density field and the classification was then propagated at the level of galaxies depending on the cell environment (node, filament, wall or void) they belong to. On the other side, T-ReX was run using as an input the set of galaxies with stellar mass $M_\star > 10^7 M_\odot$ which yields a total of $N_{\rm gal} = 142\,392$ galaxies in the $L = 100$ Gpc length box of EAGLE. We then extracted the filaments and bifurcations in the regularised graph computed with parameters $l = 25$, $\lambda_\mu = 100$, $\lambda_\sigma = 5$ and $\lambda_\pi = 1$. Note that the high value of λ_μ is required given the low stellar mass threshold used to trace the pattern with which we expect to find a large number of galaxies in walls and voids. We hence use a high pulling prior to avoid the detection of spurious filaments, even though it may prevent the algorithm from finding the most tenuous ones.

The bifurcations in the graph are defined as nodes with degree ≥ 3 and we additionally compute a local density for each of them defined as the number of galaxies residing in the $3\sigma_k$ region around their centre, with σ_k the learnt variance of the bifurcation divided by the enclosed volume of $4\pi\left(3\sigma_k\right)^3/3$. This results in a total of 285 bifurcations with various densities $n_{\rm bif}$. We can expect that only dense bifurcations trace nodes of the cosmic web. Defining the mean number density of galaxies $\bar{n}_{\rm g} = N_{\rm gal}/L^3 \simeq 0.143$, we explore three density thresholds factors of $\bar{n}_{\rm g}$, one keeping almost all bifurcations, even low dense ones, one intermediate and one very restrictive keeping only 10% of them, respectively given by $5\bar{n}_{\rm g}$, $30\bar{n}_{\rm g}$ and $100\bar{n}_{\rm g}$. The obtained classifications compared to the Nexus+ ones are shown in Table 5.2. We see that for a low value of the number density threshold (i.e. keeping almost all bifurcations), we retrieve most of the galaxies classified as belonging to nodes by Nexus+ as well (99.4%) but with a 26.5% contamination of galaxies standing in filaments according to the Nexus+ algorithm. These values tend to both decrease when the density threshold increases. These results illustrate the ability of T-ReX to identify nodes and galaxies residing in them as bifurcations in the graph structure. The filament classification in T-ReX is the result of galaxies satisfying the criterion $\sum_{k=1}^{K} p_{ik} > p_i^{\rm bkg}$, meaning that they have a higher probability of being part of the filamentary pattern than being in the uniform background, and that are not already classified as being in nodes, with $n_{\rm bif} > 30\bar{n}_{\rm g}$. We see that with this definition, we retrieve 70.2% of the galaxies classified as in filaments by Nexus+ and 34.3% of galaxies in walls. The missed galaxies in filaments can be explained by the high value of λ_μ hence leading to a graph where the tenuous filaments are missing because they are not well-traced by galaxies, but also by the density threshold for nodes galaxies in which 13% of them are classified in nodes. Exploring further the walls contamination shows that 74% (resp. 95%) of the galaxies identified by T-ReX in filaments are actually standing at less than 0.5 Mpc (resp. 2 Mpc) from a galaxy tagged as filament by Nexus+. It means that the "misclassification"[7] is only locally induced by the borders of walls close to filaments which are not precisely detected by T-ReX when using the galaxies.

We conclude that both methods agree well in classifying the galaxies in the different environments, even though T-ReX do not, by construction, distinguish between voids and walls and simply attribute the datapoint to the "background" component. Focusing on galaxies in filaments and nodes, Fig. 5.7 visually support the good agreement between the T-ReX and Nexus+ classifications in the 3D disposition of galaxies in the simulation (galaxies in filaments and nodes are respectively in green and red). The Nexus+ distribution of nodes is quite sparse, and this is the reason we did not follow the density threshold giving the smallest contamination to assign galaxies in filaments. It is also remarkable to recall that the T-ReX results are obtained using solely the galaxies to trace the pattern and do not resort to the dark matter

[7] Considering Nexus+ as a ground truth.

Table 5.2. Proportion of galaxies in the Nexus+ categories as identified in T-ReX as being in either nodes or filaments depending on the density of the bifurcation considered as nodes. Filaments classification are obtained with $n_{\rm bif} > 30\bar{n}_{\rm g}$.

Nexus+ \ T-ReX	Nodes $n_{\rm bif} > 5\bar{n}_{\rm g}$	Nodes $n_{\rm bif} > 30\bar{n}_{\rm g}$	Nodes $n_{\rm bif} > 100\bar{n}_{\rm g}$	Filaments
Nodes	0.994	0.963	0.962	0.037
Filaments	0.265	0.130	0.056	0.702
Walls	0.035	0.004	0.001	0.343
Voids	0.004	0.000	0.000	0.086

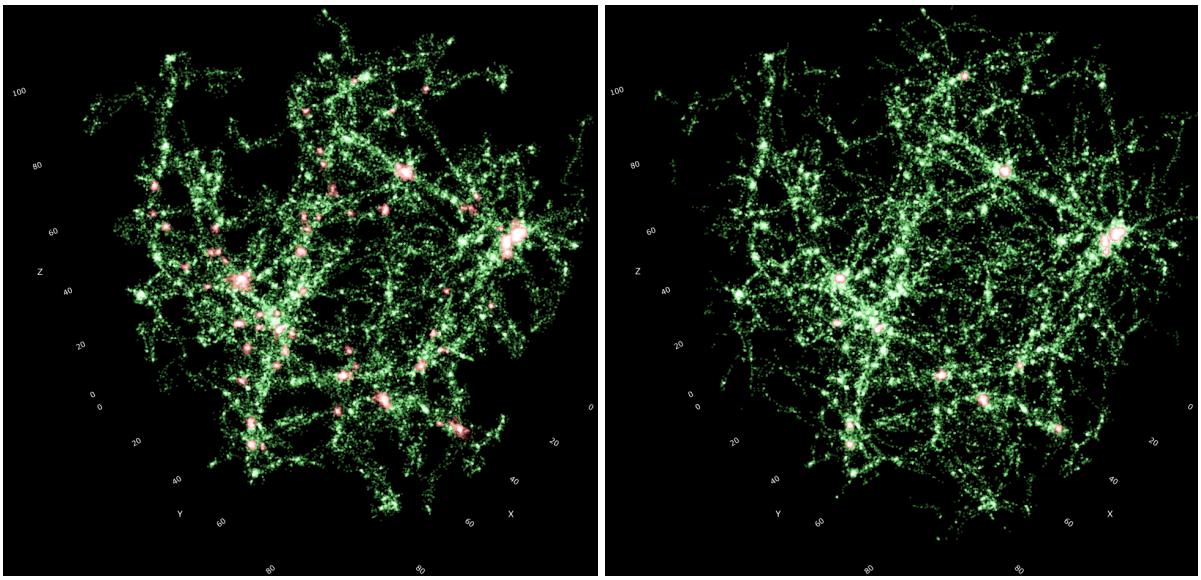

Fig. 5.7. Galaxy distribution of the EAGLE simulation coloured by their environments by T-ReX (left panel) and Nexus+ (right panel). Galaxies in green are those standing in filaments while red ones are found in nodes. Note also that we only show the galaxies classified in these environments, showing only around 65% of the galaxies in both cases.

particles which number density is nearly $24\,000$ times larger.

5.4 Filaments characteristics in simulations

In this section, we propose to study the statistical characteristics of filaments as defined in Sect. 5.3.2 obtained from the set of galaxies in three hydrodynamical simulations. In addition of different recipes for baryonic processes, these simulations have also different mass and volume resolutions consequently drawing different pictures of the cosmic web. The assessment of filaments characteristics is hence of importance to calibrate as well as possible the simulations with respect to observations and improve the quality of the physics governing the formation and evolution of galaxies.

5.4.1 Simulations and principal graphs

We use multiple datasets to compute several principal graphs, extract the corresponding filaments and their characteristics. These simulations are briefly described below, together with the parameters used for the running of the T-ReX algorithm to obtain the analysed filaments.

EAGLE. As already presented, EAGLE [Schaye et al., 2015] is a $(100\mathrm{Gpc})^3$ box hydrodynamical simulation. The cosmological parameters used to evolve the simulation are consistent with *Planck*15 results [Planck Collaboration XIII et al., 2016] with $\Omega_\Lambda = 0.693$, $\Omega_\mathrm{m} = 0.307$, $\Omega_\mathrm{b} = 0.04825$, $\sigma_8 = 0.8288$, $n_\mathrm{s} = 0.9611$ and $h = 0.6777$. We use the identified galaxies with stellar mass $M_* > 10^7 M_\odot$ yielding a total of 142 392 galaxies and consequently, a mean number density of galaxies of $\bar{n}_\mathrm{g} \simeq 0.142$. The catalogue of filaments is obtained by running T-ReX in initialised with a pruned MST with $l = 25$, and $\sigma_k^{(0)} = 1$ Mpc. Values of the hyper-parameters are $\lambda_\mu = 100$, $\lambda_\sigma = 5$ and $\lambda_\pi = 1$. We discard from the analysis all the small branches with $L < 2\sigma_k^{(0)}$ Mpc which leads to a total of 572 filaments in the catalogue.

IllustrisTNG. The suite of hydrodynamical cosmological simulations IllustrisTNG [Nelson et al., 2019] follows the evolution of dark matter, gas, stars, and black holes on a moving mesh from redshift $z = 127$ to $z = 0$. The input cosmology of the simulation is consistent with the *Planck*15 one as well with $\Omega_\Lambda = 0.6911$, $\Omega_\mathrm{m} = 0.3089$, $\Omega_\mathrm{b} = 0.0486$, $\sigma_8 = 0.8159$, $n_\mathrm{s} = 0.9667$ and $h = 0.6774$. We use here the largest simulation box IllustrisTNG300-1 with a size length of 302.6 Mpc. Galaxies are identified by means of the SubFind algorithm [Springel et al., 2001] and we restrict the analysis using galaxies with $M_* \geq 10^8 M_\odot$, to highlight a different setup and number density than the EAGLE case. In total, we get 603 630 galaxies leading to a mean number density $\bar{n}_\mathrm{g} \simeq 0.0218$ which is hence almost 5 times smaller than the EAGLE sample. The T-ReX algorithm is consequently run using $l = 15$, $\sigma^{(0)} = 1$ Mpc and with $\lambda_\mu = 30$, $\lambda_\sigma = 5$ and $\lambda_\pi = 1$. Note that these values are lower than in the EAGLE case due to the lower number density of galaxies and higher mass threshold which hence require less smoothing and denoising of the initial structure. Using the same threshold of $L > 2\sigma^{(0)}$, we end up with a total of 5 492 filaments identified.

Magneticum. As a last sample of galaxies, we use the $z = 0.066$ snapshot of the Magneticum simulation Box2/hr [Hirschmann et al., 2014] with a size of 500 Mpc. The simulation uses the cosmological parameters consistent with WMAP7 data [Komatsu et al., 2011] with $\Omega_\Lambda = 0.728$, $\Omega_\mathrm{m} = 0.272$, $\Omega_\mathrm{b} = 0.0456$, $\sigma_8 = 0.809$, $n_\mathrm{s} = 0.963$ and $h = 0.704$. Similarly to the case of IllustrisTNG, we restrict the analysis to the set of galaxies with $M_* \geq 10^8 M_\odot$ leading to a total of 1 363 468 galaxies and a mean number density $\bar{n}_\mathrm{g} = 0.011$, a value twice as small as that of IllustrisTNG. The used parameters are exactly the same as in the IllustrisTNG case leading to 11 675 filaments.

5.4.2 Comparison of filaments characteristics

From each obtained principal graph, we extract individual filaments together with their characteristics as defined in Sect. 5.3.2. The probability distribution functions of some of these properties are displayed in Fig. 5.8. The top left panel is focusing on the length L of filaments, which is showing an exponential tail also reported by many previous findings [such as Bond et al., 2010; Galárraga-Espinosa et al., 2020; Malavasi et al., 2020b; Rost et al., 2020]. At small lengths, all the samples provide similar distributions with a peak of the distribution around

5 Mpc, as also reported in Galárraga-Espinosa et al. [2020]. At larger lengths, the differences between the simulations can be attributed to their different volumes which is indeed larger for Magneticum than for TNG which in turn is also larger than the EAGLE simulation, hence allowing for longer filaments. The curvature distribution are closely similar with a large number of low-curvature filaments in all considered catalogues and only few highly curved ones, as shown in the top right panel of Fig. 5.8. In addition to the previous quantities for filaments, we also compute the mean galaxy density of each filament defined as $\rho_{\rm f} = n_{\rm f}/\left(L\pi 4r^2\right)$, where $n_{\rm f}$ is the number of galaxies associated individually to the filament. In the bottom left panel, we show the distribution of filament overdensities defined as $\rho_{\rm f}/\bar{n}_{\rm g}$ where $\bar{n}_{\rm g}$ is the average number density of galaxies in the overall input volume depending on the simulation box. The density of filaments spans a broad range in all cases, from tenuous ones with galaxy overdensities close to 0 and up to ~ 100 for dense bridges of matter also drawn in dark matter analyses of the cosmic web like Cautun et al. [2014]. In the bottom right panel, we can also see that the radii as defined by Eq. (5.8) are distributed over few Mpc around the filament spine in all the simulations with a peak slightly below 2.5 Mpc for TNG and EAGLE and around 2.8 Mpc for Magneticum. Even though depending on the definition used to define the width, these results are also consistent with previous measurements of filament extensions in simulations exhibiting peak of the radii distribution between 2 and 3 Mpc [Colberg et al., 2005; Bond et al., 2010; Cautun et al., 2014] using the dark matter. The differences between the three distributions are mainly due to the number density of objects that is higher for EAGLE than for TNG and Magneticum, this latter having the smallest. Note also that the radius of filaments has been shown highly sensitive to the time-evolution of structures with thinner filaments at low redshift as a result of the gravitational collapse and that Magneticum has a slightly higher redshift than the two other ones. These results on the spatial radial extent of filaments are also consistent with the observational findings of Bonjean et al. [2019] with a characteristic radius of 7.5 Mpc.

When investigating further the relations between individual properties of filaments, we find strong dependencies on the curvature and densities with the spatial extent of the filament. In the left panel of Fig. 5.9 is shown the evolution of the curvature γ with the length L while the right panel displays the evolution of the filaments overdensities $\rho_{\rm f}/\rho_{\rm b}$. All the filaments in the studied simulations show the same trends with shorter filaments appearing, on average, straighter and denser than their longer counterparts. This result is in line with those of Galárraga-Espinosa et al. [2020] who identify two populations of filaments with short ones mainly standing in overdense regions and longer ones connecting lower density regions in the Universe. We hence additionally show that small bridges connecting dense clusters are also respectively straighter than the long filaments. Note also that, even though the trends are similar for the different simulations, the absolute values of the quantities vary. This may be due to the different baryonic physics models that are impacting the distribution of matter around filament, as shown by Galárraga-Espinosa et al. [2020]. This work constitutes a very first step in studying the statistical properties of filaments drawn from different simulations. However, the end goal is to apply the identification tool on actual galaxy surveys. To do so, several observational effects must be taken into account that we leave for future investigations.

5.5 The impact of the cosmic web on cluster properties in simulations

In Sect. 5.3.3, we introduced a definition of nodes in the cosmic web as bifurcations in the graph structure based on the assumption that nodes can be found at the densest intersections

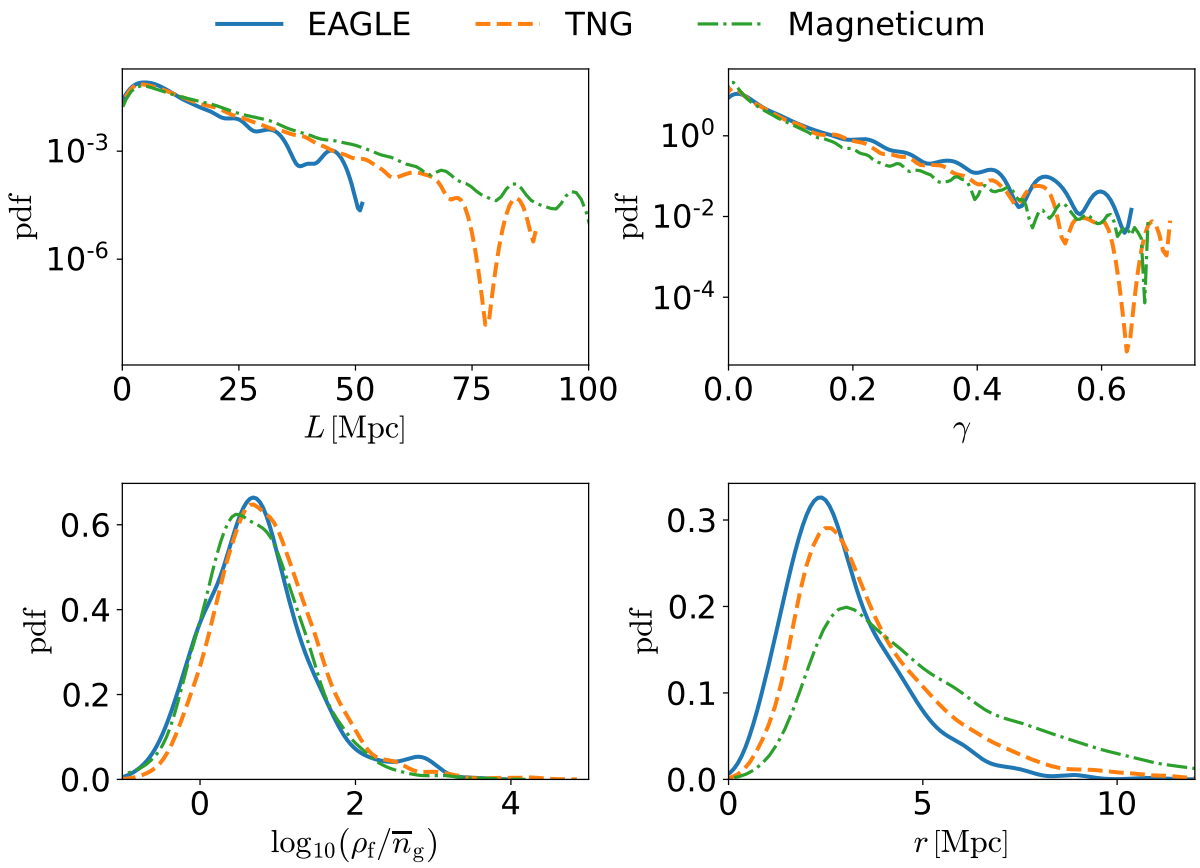

Fig. 5.8. Probability distribution functions of several quantities extracted from the three catalogues of filaments. Are shown the length L (top left), the curvature γ (top right), the overdensity ρ_f/ρ_b (bottom left) and the radii as defined by Eq. 5.8 (bottom right) of filaments.

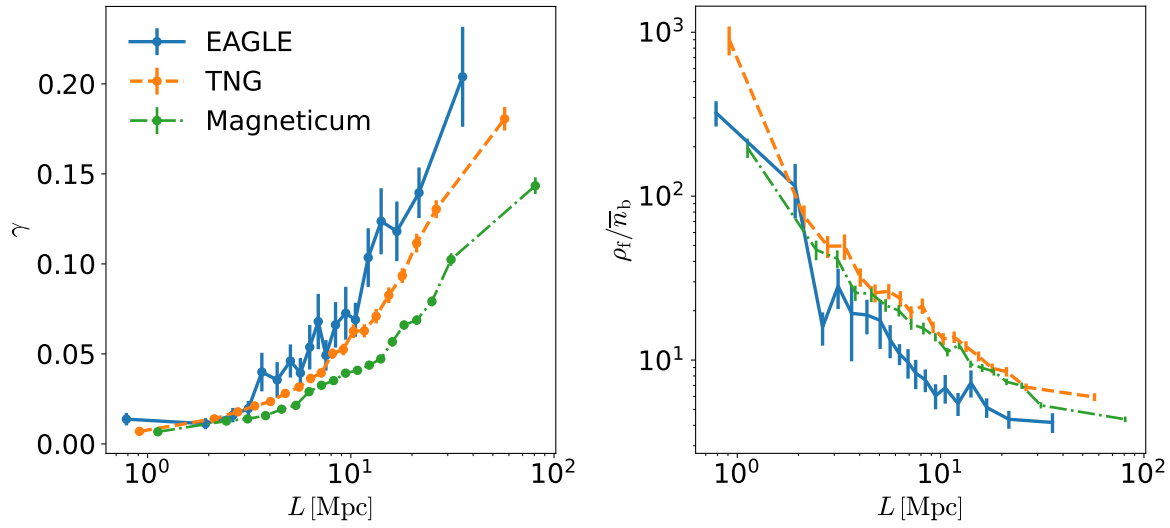

Fig. 5.9. Evolution of the average curvature (left panel) and density (right panel) of filaments as a function of their length L for the three catalogues. Error bars are the 68% bootstrap confidence interval on the mean in each length bin.

between filaments. The local topology and geometry of the density field near these massive halos can be probed by the *connectivity*, a quantity encoding the number of filaments a node is connected to. The connectivity of halos must decrease with the cosmic time due to the Universe accelerated expansion which disconnects cosmic nodes from the network of filaments [Pichon et al., 2010]. Therefore, the statistics of halo connectivity is expected to depend on the growth factor and hence constitute a topological constraint on dark energy [Codis et al., 2018]. On the the other hand, the filamentary pattern surrounding the nodes (quantified by the connectivity) can heavily impacts their characteristics such as their internal properties and density profiles [see e.g. Contigiani et al., 2021]. In this section, we summarise our investigation of the impact of the connectivity on the physical properties of galaxy clusters (morphology, dynamical state, and mass assembly history), detailed in Gouin et al. [2021].

5.5.1 Data, filamentary pattern and connectivity

For this analysis, we use the set of halos detected from the largest resolution of the IllustrisTNG simulations [Nelson et al., 2019] with a box size of 302.6 Mpc (see Sect. 5.4.1). Groups and halos are detected in the simulation by means of the FoF algorithm [Davis et al., 1985] with a linking length of 0.2 Mpc/h. To trace the filamentary pattern, we additionally make use of the set of subhalos identified with the Subfind algorithm to detect substructures inside host halos [Springel et al., 2001]. From the set of FoF sample, we select 2522 halos representing galaxy groups and clusters with masses $M_{200} \geq 1 \times 10^{13} M_{\odot}/h$ at $z = 0$. We define R_{200} as the radius enclosing a mass of M_{200} characterising a mean overdensity of 200 times the critical background density. Notice also that we discard 79 halos that are less distant than $3R_{200}$ from the edges of the simulation box to focus our analysis on the large-scale environments around halos.

The anisotropy in the large-scale environment of groups and clusters is quantified by the local number of filaments that are connected to them, the so-called connectivity κ. This proxy is a powerful tool to understand the geometry of the underlying density field, as discussed in Codis et al. [2018]. To detect the filaments in the simulation, we use the T-ReX algorithm applied on the set of subhalos with $M_* \geq 10^9 M_{\odot}$, following Galárraga-Espinosa et al. [2020, 2021]. From the output graph, we associate to each group and cluster the closest node of the graph, and the local connectivity κ is the number of intersecting filaments in a sphere of $1.5R_{200}$ radius around the graph node (a similar definition adopted in previous works like Darragh Ford et al. [2019]; Sarron et al. [2019]). In addition to the 79 previously removed halos, we remove 24 more that are low-mass groups considered as standing too far away from the graph in terms of projected distance (> 1 Mpc/h). As a result, our final sample of groups and clusters consists of 2419 halos, for which we estimate the connectivity. The cluster connectivity is illustrated in Fig. 5.10, and one can see that the T-ReX filamentary structure (computed from the galaxy distribution) traces well the DM distribution around halos. In addition, Fig. 5.11 shows the probability distribution function of the connectivity for four mass bins. The distribution of massive clusters ($M_{200} > 10^{14} M_{\odot}/h$) peaks around $\kappa \sim 3, 4$ meaning that they are mostly connected to around three or four filaments, and is spread over large κ values (up to $\kappa = 6$). In contrast, the connectivity statistics of the lowest mass groups ($10^{13} M_{\odot}/h < M_{200} < 2 \times 10^{13} M_{\odot}/h$) strongly peaks at $\kappa = 2$, suggesting that low-mass groups are preferentially located inside filaments, whereas massive clusters are more likely located at the nodes of cosmic web, connecting more than two filaments. This thus confirms that massive structures have a higher connectivity than low-mass ones, as also found by previous works such as Aragon-Calvo et al. [2010]; Codis et al. [2018]; Darragh Ford et al. [2019];

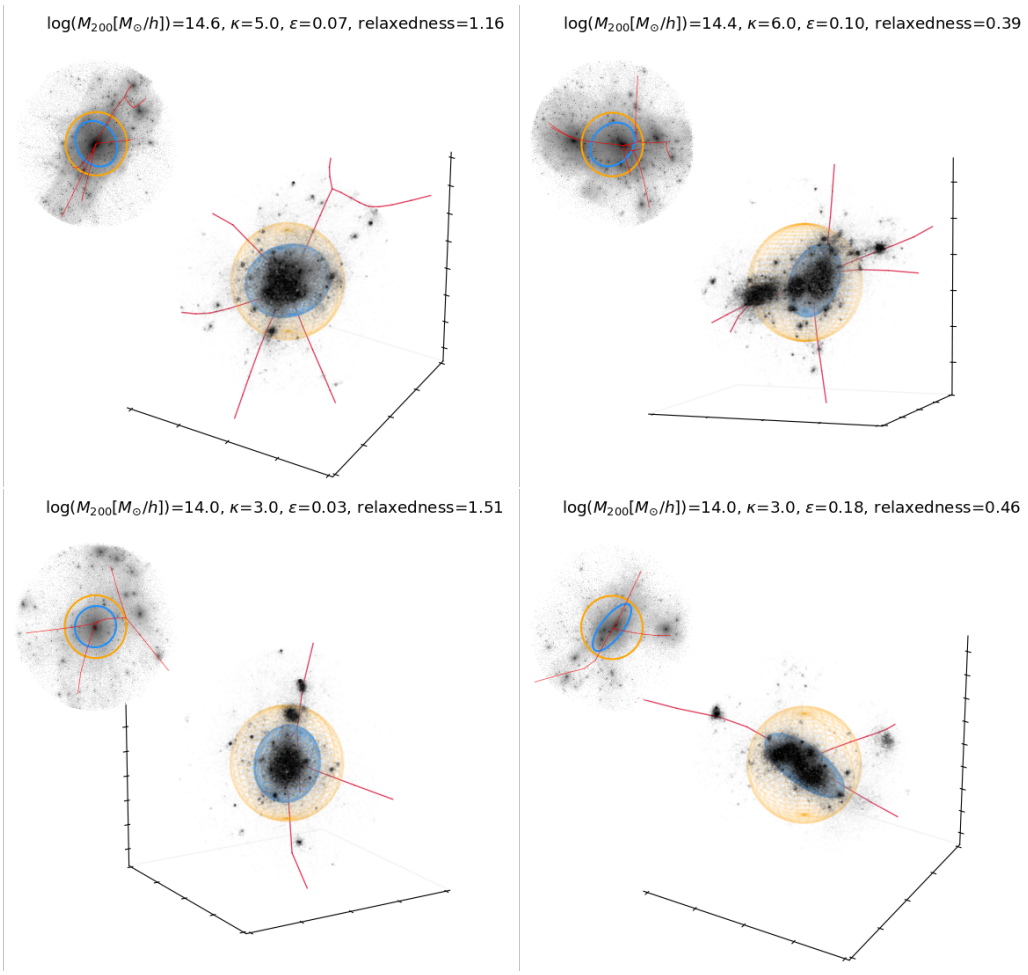

Fig. 5.10. Illustration of the ellipsoidal shapes and large-scale environments of four simulated galaxy clusters. The blue ellipsoids are computed by calculated the mass tensor of DM distribution, following [Jing & Suto, 2002]. The red lines represent cosmic filaments reconstructed by the T-ReX algorithm from the galaxy distribution. Yellow spheres have a radius of $1.5R_{200}$.

Sarron et al. [2019]; Malavasi et al. [2020b]. In practice, the connectivity of low-mass groups with $M_{200} \in [10^{13}; 10^{13.5}]\, M_\odot/h$ is low with nearly 80% of the objects standing in a filament or at its extremity with $\kappa < 2$ while the most massive ones with $M_{200} > 10^{14.2} M_\odot/h$ are more connected with $\kappa > 3$.

5.5.2 Impact of connectivity on the growth and shapes of clusters

It is already well known that the mass of halos are strongly correlated with their shape [Aragon-Calvo et al., 2010; Codis et al., 2018; Darragh Ford et al., 2019; Sarron et al., 2019; Malavasi et al., 2020b]. Beyond the driving mass effect, we investigate the influence of the large-scale cosmic web environment on the shape of the mass distribution in clusters. Their morphology, as defined by the ellipticity ϵ, measures the local anisotropy traced by the dark matter particles. Similarly to Jing & Suto [2002], we compute the ellipticity ϵ as the ratio $(\lambda_1 - \lambda_3) \,/\, (\lambda_1 + \lambda_2 + \lambda_3)$ where λ_i is the ith largest eigenvalue of the local mass tensor. Starting from a sphere, the procedure is made iterative, shrinking the axes, such that the ellipsoid encloses a total mass of M_{200}. An illustration of the resulting ellipsoids is shown in Fig. 5.10 for four clusters. The first row of Table 5.3 quantifies the correlation between κ and ϵ through the

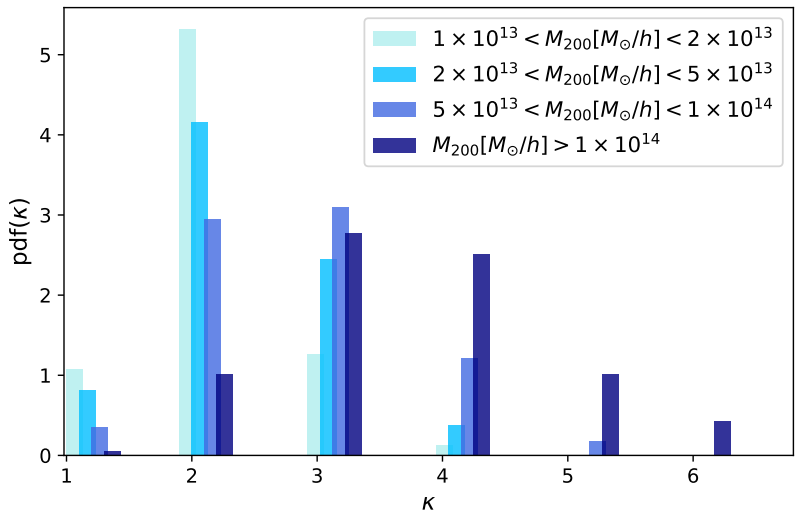

Fig. 5.11. Probability distribution function of groups and clusters connectivity, κ, for four mass bins.

Spearman rank correlation coefficients $\rho_{\rm sp}$ for two sets of clusters. The correlation coefficient increases when the mass increases and reaches $\rho_{\rm sp} \sim 0.24$ when $M_{200} \geq 10^{14} M_\odot/h$, hence stressing that the connectivity is also closely linked to the ellipticity, beyond the effect of the mass dependence. This is also shown in the left panel of Fig. 5.12 in which, in a given mass bin, groups with a high connectivity are more elliptical on average than low-connectivity groups and clusters.

The observed impact of filamentary structure on the cluster shape must be the result of different accretion phases and it is natural to think that the connectivity is also closely linked to the way clusters accrete matter. To confirm this assumption, we compare the connectivity to a proxy of the accretion of a halo at $z \sim 0$. We therefore introduce the instantaneous mass accretion rate defined as

$$\left(\frac{{\rm d}M_{200}}{{\rm d}t}\right)_{z\sim 0} = \frac{M_{200}(t+{\rm d}t) - M_{200}(t)}{{\rm d}t}. \tag{5.12}$$

To compute it, we use the ten last snapshots of the simulation and perform a linear regression to obtain the value of ${\rm d}M_{200}/{\rm d}t$. We show in the right panel of Fig. 5.12 the evolution of the the instantaneous mass accretion rate at $z = 0$ as a function of mass for three bins of connectivity (low connectivity with $\kappa < 3$, mildly connected with $\kappa = 3$ and highly connected with $\kappa > 3$). We see that, at fixed mass, highly connected clusters tend to grow faster than low-connectivity groups and clusters, also emphasised by the Spearman correlation coefficients from the second row of Table 5.3. Intuitively, one can assume that more mass is feeding a cluster when this latter is connected to a larger number of filaments.

These results hence highlight that the halo environment in the cosmic web is impacting its shape, independently of its mass: highly connected clusters accrete more matter and this fast accretion must disturb their mass distribution consequently increasing their ellipticity.

5.5.3 Impact of cluster dynamical states on the connectivity

Beyond the shape of a cluster as traced by its ellipticity, we can focus on the correlation between the connectivity of a node in the cosmic web and its dynamical state. We quantify

	$M_{200} > 1 \times 10^{13} M_\odot/h$		$M_{200} > 5 \times 10^{13} M_\odot/h$	
	ρ_{sp}	p-value	ρ_{sp}	p-value
ϵ and κ	0.11	2.15×10^{-7}	0.17	5×10^{-4}
κ and $dM/dt_{z\sim0}$	0.33	3.8×10^{-60}	0.35	4×10^{-13}
κ and χ_{DS}	-0.13	4.6×10^{-11}	-0.17	5×10^{-4}

Table 5.3. Spearman rank correlation coefficients ρ_{sp} and corresponding p-values between halo properties and the connectivity for all the groups and clusters in the sample ($M_{200} > 1 \times 10^{13} M_\odot/h$), and for the 408 most massive groups and clusters ($M_{200} > 5 \times 10^{13} M_\odot/h$).

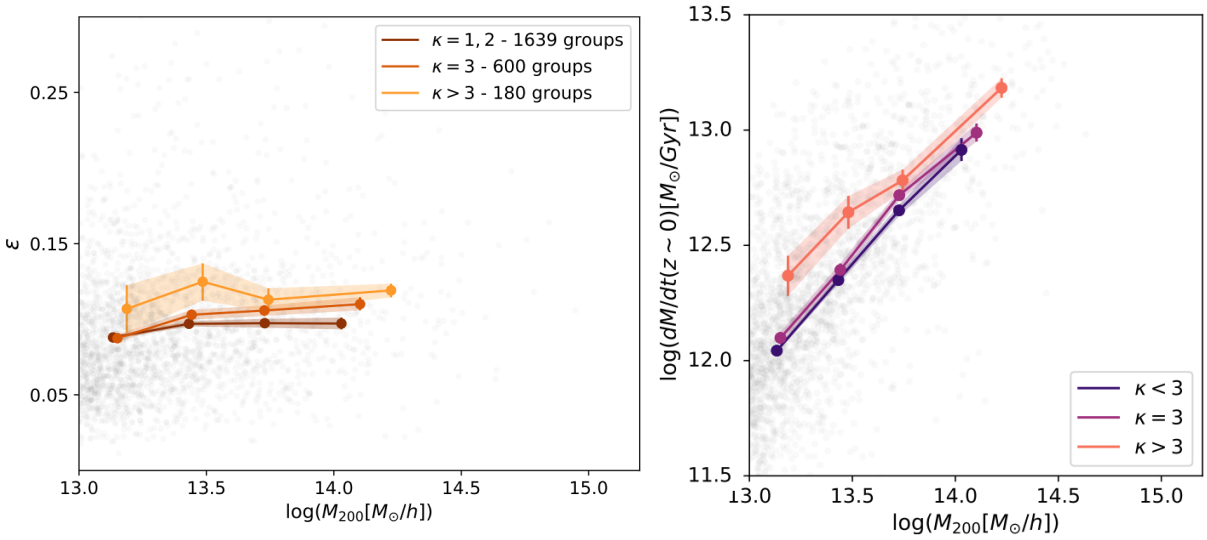

Fig. 5.12. *(left)* Ellipticity ϵ as a function of the group mass for three bins of connectivity: $\kappa = \{1, 2\}$, $\kappa = 3$ and $\kappa > 3$. *(right)* Mean instantaneous mass growth $\mathrm{d}M/\mathrm{d}t$ computed at $z \simeq 0$ as a function of the group mass for three same bins of connectivity. Error bars are the 68% errors on the mean, derived from bootstrap resampling.

the level of dynamical relaxation, called the *relaxedness*, defined by [Haggar et al., 2020] as

$$\chi_{DS} = \sqrt{\frac{3}{\left(\frac{\Delta_r}{0.07}\right)^2 + \left(\frac{f_{sub}}{0.1}\right)^2 + \left(\frac{\eta-1}{0.15}\right)^2}}, \tag{5.13}$$

where Δ_r is the centre of mass offset between the density peak and the barycentre of the mass distribution in the halo, divided by the virial radius, f_{sub} the subhalo mass fraction defined as $\sum M_{sub}/M_{tot}$ where $\sum M_{sub}$ is the sum of the subhalo masses except the most massive one, and η is the virial ratio characterising the level of virialisation of the halo defined as $\eta = 2T/|W|$, with T and W the kinetic and gravitational potential energy respectively. Intuitively, a cluster that has a low value of $f_{sub} < 0.1$ (few substructures), a low offset of the centre of mass $\Delta_r < 0.07$ and that is close to virial equilibrium with $|\eta - 1| < 0.15$ can be considered as relaxed, leading to $\chi_{DS} \geq 1$ [Kuchner et al., 2020]. Groups and clusters that do not fulfil this criterion are considered as unrelaxed ones.

It is admitted that the mass is also driving the relaxedness of clusters with more massive ones being less relaxed on average [see e.g. Power et al., 2012; Kuchner et al., 2020] since they formed later and are still in their formation phase. Beyond this mass dependence, we show in the left panel of Fig. 5.13 the evolution of the connectivity as a function of the mass for the

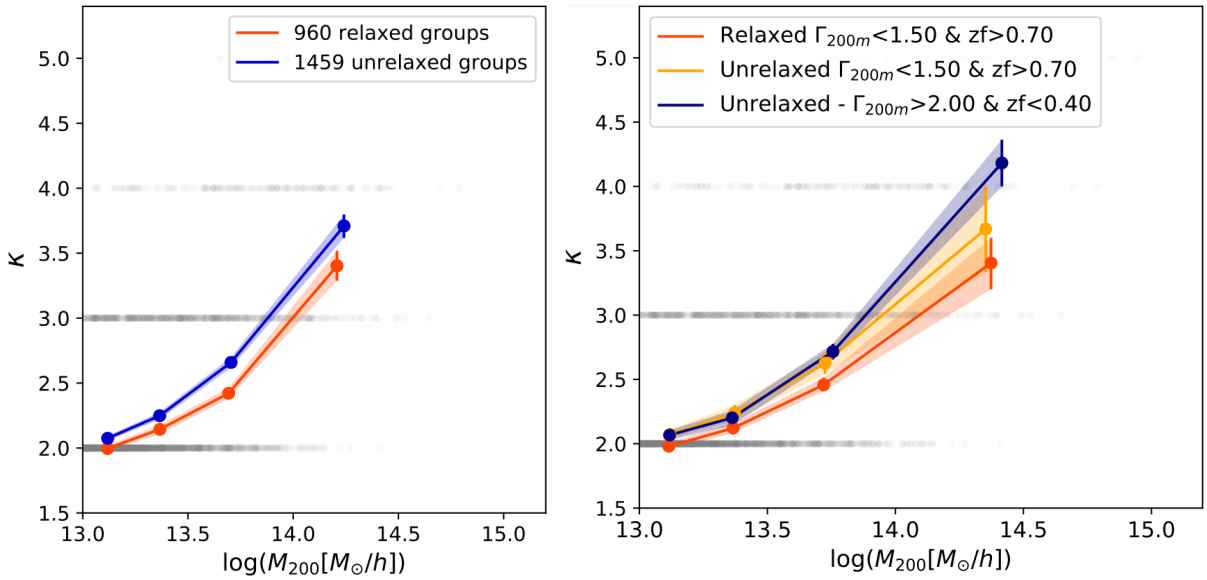

Fig. 5.13. *(left)* Evolution of the connectivity as a function of mass for relaxed and unrelaxed clusters. *(right)* Evolution of the connectivity as a function of mass for relaxed and unrelaxed clusters for three subsamples with different mass assembly histories. Error bars are the 68% errors on the mean, derived from bootstrap resampling.

two categories of relaxed and unrelaxed clusters. At a fixed mass bin, we see that unrelaxed clusters are more connected to the cosmic web than relaxed ones, a result which appears in the last row of Table 5.3 as a weakly-negative correlation between the two quantities whose absolute value grows with the mass threshold. This shows that independently of the mass, the connectivity is impacting the dynamical state of groups and clusters.

5.5.4 The influence of mass growth history

In addition to the higher accretion rate of high-connectivity clusters shown in Sect. 5.5.2, we exhibited in Sect. 5.5.3 that these latter are also less relaxed on average. All these results suggest that the connectivity is actually tracing different mass assembly histories (MAH) of clusters. Previous works already point out a relation between dynamical state and MAH [Power et al., 2012; Mostoghiu et al., 2019]. Therefore, we wish to investigate how a cluster large-scale environment in the cosmic web influences its MAH. In addition to the relaxedness, we focus here on two quantities linked with the MAH of clusters given by the formation redshift z_f at which the halo reached half of its M_{200} mass at $z = 0$ [Cole & Lacey, 1996] and the continuous mass accretion rate $\Gamma_{200} = \Delta \log M_{200} / \Delta \log a$, with a the scale factor. In the right panel of Fig. 5.13 we show the evolution of the connectivity for three subsamples of clusters with different MAH: (i) the early-formed, relaxed and slowly-accreting clusters (in orange); (ii) the early-formed, unrelaxed and fast-accreting clusters (in yellow); and (iii) the lately-formed, unrelaxed and fast-accreting clusters (in blue). We observe that older and more relaxed clusters are weakly-connected, while the unrelaxed ones are more connected to the cosmic web at fixed mass. Finally, the young and unrelaxed population of clusters in blue are significantly more connected to the cosmic web. These objects are strongly affected by the infalling matter and can be thought as the result of recent merger events disturbing the matter distribution and increasing the connectivity [Klypin et al., 2016; Darragh Ford et al., 2019; Vallés-Pérez et al., 2020].

All these observations advocate for a strong impact between the mass assembly history of clusters and the way they are embedded in the cosmic web. Old relaxed clusters are more spherical and slowly accreting matter with a small value of the connectivity while young and still in formation clusters are unrelaxed, more elliptical and connected in the cosmic web by numerous filaments feeding them abundantly.

5.6 Summary and perspectives

Given the primordial role of filaments in the cosmic web, we investigated in this chapter a possible definition based on the principal graph formulation established in Chapter 4 not only for the filaments but also for the nodes of the cosmic web. We particularly showcased two possible applications of the proposed method, entitled T-ReX, aiming either at detecting the filamentary structure as a whole, together with a positional uncertainty of the ridge, or at defining individual filaments as branches of the principal graph. We showed that this definition based on the probabilistic framework of T-ReX allowed the derivations of many interesting characteristics for the statistical analyses of filaments that we carried out from multiple samples of galaxies in simulations. Finally, we used the smooth graph learnt by the algorithm to perform a thorough analysis of the properties of galaxy clusters depending on their spatial embedding in the cosmic web as measured by the connectivity.

In this chapter, we covered several topics related to the filamentary structure of the cosmic web and, more precisely:

1. We showed that the T-ReX algorithm performs as good as other web finders for the detection of filament ridges but adding the information of the uncertainty on the ridge and using only a sparse distribution of the input traced by halos in simulations.

2. We exposed how the principal graph can be used to identify the cosmic web elements that contain the largest mass fraction, namely nodes and filaments, as dense bifurcations and branches in the graph structure respectively. By deriving from the obtained graph different characteristics for filaments, we were able to exhibit an exponential tail in the distribution of their length and their curvature. The radii of filaments resulting from local measurements in the T-ReX algorithm are spread over few Mpc with a peak between 2 and 3 Mpc. All these results were shown to be in agreement with previous findings of filaments statistics.

3. The galaxies used to determine the graph can be associated consistently with each individual filament, to nodes, or to the background in a probabilistic setup. This association was compared to a physical web finder extracting multi-scale filaments and we find a very good agreement with the advantage of having at hand a method relying on a much sparser sampling of the matter distribution.

4. By studying the connectivity of clusters in the cosmic web, we confirmed the mass-connectivity relation exhibited by previous works. We also showed that the cosmic environment around clusters significantly impact their shapes and the way they accrete matter such that, at fixed mass, halos in nodes (with high connectivities) have a larger accretion rate than those standing in filaments (weakly-connected). These results led us to study multiple scenarios of mass assembly his-

tory for galaxy clusters and how they connect to the cosmic web. In particular, old relaxed groups and clusters have a small connectivity while young, fast-accreting and unrelaxed galaxy clusters have more surrounding filaments on average.

We sketched an analysis of the filaments properties traced by the distribution of galaxies in hydrodynamical simulations. The natural next step will be to study such characteristics as extracted from actual galaxy surveys like the Sloan Digital Sky Surveys [York et al., 2000] or the Dark Energy Survey [Abbott et al., 2016], in a similar manner to [Malavasi et al., 2020b; Rost et al., 2020]. Samples of reliably detected filaments with their characteristics (size, radius, curvature, etc.) have already proven to be crucial for the study of the distribution of baryons in filaments [e.g. Tanimura et al., 2020a] or study their properties in other observables like Sunayev Zel'dovich effect or X-rays [e.g. Tanimura et al., 2020c,b]. However, constructing samples of reliable filaments necessitates to overcome some observational complications. The two main difficulties being that (i) most of the large-scale galaxy surveys are photometric observations with relatively large errors on the estimated redshift; and (ii) observations are carried out in redshift-space inducing some distortions of the spatial distribution of galaxies (see Sect. 2.2.2). Both of these issues create elongations of the distribution along the line-of-sight (LoS) that could be handled in the T-ReX formalism by allowing the spherically Gaussian components composing the graph nodes to be elongated in the LoS direction. This proposed modification could result in a built-in approach to reduce the effect redshift-space distortions and the blurring induced by photometric redshift measurements produced on the extracted filamentary pattern. Note however that this would come with an increase of the computational cost since the responsibilities from the Expectation-Maximisation procedure will require the full computation of Mahalanobis distances $(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\mathsf{T} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$ instead of isotropic L_2 norm which can easily be optimised.

In addition to the previously-mentioned interests, the detection of reliable filaments and their characteristics is key to improve our understanding of the structure formation. As we have seen, the connectivity in simulations is strongly correlated with the physical properties of clusters and has also been shown to depend on the underlying cosmology and on the redshift of the data distribution in Codis et al. [2018]. A further analysis of how such an information perform and how it can be combined with existing probes to constrain cosmological models is also an interesting perspective that matches the scientific question of the next Chapter 6.

Finally, in the T-ReX formalism, the filamentary structure is modelled as a graph, and we exploit only partially the graph information to constrain the overall smoothness of the resulting structure. However, the graph encodes much more information at the nodes and edges level. Assigning for instance local physical properties to graph nodes (like densities, masses, etc.) may offer the opportunity to link the topology of the cosmic web to the physics of tracers, similarly to Coutinho et al. [2016]. Additionally, the statistics of edges in the MST was found sensitive to cosmology in previous works [Hong & Dey, 2015; Hong et al., 2016; Naidoo et al., 2020], showing the ability of graph-derived statistics to encode a significant amount of information about the underlying cosmological model.

Chapter 6.

Constraining cosmological parameters with cosmic environments

"It should be quite straightforward."
J. Kuruvilla

Chapter 5 investigated a possible definition of the cosmic environments that are nodes and filaments from a distribution of matter tracers (galaxies or halos) and studies their physical properties through a graph framework. In this chapter, we make use of cosmic environments defined physically through the eigenvalues of the Hessian matrix of the gravitational potential and traced in simulations by the full dark matter distribution. Focusing on the two-point statistics derived from the environments and carried out in both real and redshift spaces, we present the first quantitative measurement of the informative power of the different cosmic environments about the underlying cosmological parameters as compared to the analysis of the matter power spectrum.

6.1 Context and introduction

6.1.1 The matter power spectrum as a cosmological probe

The most basic statistics that one can build from centred cosmological fields such as the overdensity field δ or convergence maps are based on the two-point correlation function or its

Fourier-equivalent, the power spectrum. This latter summary statistics quantitatively expresses the covariance between Fourier modes of the field (see Sect. 1.3.2). Beyond its simplicity, the two-point statistics fully encodes the information of Gaussian random fields, particularly well describing some cosmological fields, like the early distribution of matter. Theoretical predictions, allowed for instance by linear perturbation theory, depicts the late-time matter power spectrum as depending on the initial one with great accuracy at large scales ($k < 0.15\ h$/Mpc, see Fig. 2.2 and related discussion). In particular, the linear theory permits to describe the impact of each cosmological parameter of the ΛCDM model on the late-time matter power spectrum based on the one from the initial density field at the linear scales. As such, the interpretability of the two-point statistics is well-known and well-understood at linear scales [Heath, 1977; Peebles, 1980], even for non-Gaussian overdensity fields, making the matter power-spectrum a keystone for the statistical analysis of large-scale structures since their first observations in the early 80s.

6.1.2 The cosmic environments as an alternative probe

When the field is non-Gaussian, such as the late-time matter distribution in the Universe, the two-point statistics is not carrying all the information about the underlying field. Even though still informative, the matter power spectrum is subject to degeneracies among parameters of the ΛCDM model which prevent it from fully constraining different values. For instance, changes in the matter density Ω_m, in the normalisation of the power spectrum σ_8 and in the summed neutrino mass M_ν are known to produce similar effects on the matter power spectrum over a wide range of scales.

In Sect. 2.3.1, we discussed several statistics derived from topological definitions of the matter distribution or based on machine-learning data compression methods to incorporate the information that leaked into higher-than-two order moments. Alternatively, cosmic environments also exhibit particular dependencies with cosmological parameters. The hierarchical formation of structures makes nodes a wealthy source of information [White & Frenk, 1991] depending on the matter and dark energy contents of the Universe but also on the amplitude of the initial density fluctuations [for a review, see Allen et al., 2011]. Their number counts, shapes, mass profiles and evolution with redshift have been shown particularly efficient in partially breaking the degeneracy between Ω_m and σ_8 occurring in a matter power spectrum analysis [Bahcall et al., 1997; Bahcall & Fan, 1998; Holder et al., 2001]. The always larger and more complete availability of cluster samples obtained from different observables like optical, X-rays and millimetre wavelengths enabled to considerably improve the cosmological constraints derived from these statistics [see e.g. Mantz et al., 2015; Salvati et al., 2018; DES Collaboration et al., 2020; Corasaniti et al., 2021]. In Costanzi et al. [2013] is also shown that varying M_ν induces different abundances of massive clusters at fixed primordial conditions and Villaescusa-Navarro et al. [2014] exhibit a scale-dependence of the bias for those massive tracers in Universe with massive neutrinos. Voids, on their side, are objects of low-density making them the component of the cosmic web the least affected by the non-linear collapse of matter. This property is particularly interesting to probe the accelerated expansion of the Universe and study the dark energy [Lee & Park, 2009; Lavaux & Wandelt, 2012; Pisani et al., 2015]. The measure of their sizes, shapes, counts and corresponding evolution with redshift are key quantities related to the underlying cosmology [van de Weygaert & Platen, 2011; Hamaus et al., 2014, 2015]. Numerous works also point out the interest of studying the effect M_ν in voids since the large thermal velocities of neutrinos, coupled with the volume-dominating property of voids, make neutrinos contribute to a large extent to the overall mass

voids enclose. In particular, voids have been measured smaller and denser when M_ν increases in massive neutrino simulations [Villaescusa-Navarro et al., 2013; Massara et al., 2015; Kreisch et al., 2019]. The constraints brought by these two extreme environments that are voids and nodes are combined by Bayer et al. [2021] who show that the combination of information provided by the halo mass function and the void size function leads to considerable improvement over the matter power spectrum constraints in real space. Kreisch et al. [2021] recently shown that using the same statistics derived from halos in the simulations also yields sizeable gains.

In an attempt to incorporate local information about the underlying pattern into the two-point analysis, several works rely on a weighted version of the matter clustering in which a mark is assigned to each source, based for instance on the local luminosity [Beisbart & Kerscher, 2000; Sheth et al., 2005] or density [White, 2016]. This latter version, the marked-by-density power spectrum up-weights the low-density parts of the field and is of particular interest for discriminating between several cosmologies [Valogiannis & Bean, 2018; Armijo et al., 2018] and constraining cosmological parameters in real-space [Massara et al., 2021].

In this picture, the different components of the cosmic web, from its densest to its emptiest parts are a promising way to provide complementary information about the underlying cosmological model. They show their own sensitivities to some cosmological parameters whose effects are limiting the constraints obtained from the matter power spectrum. We hence seek to quantify the information brought by the combination of low, intermediary and high density environments represented respectively by voids, walls, filaments and nodes that should be able to break some degeneracies among parameters allowing to improve the constraints over a direct analysis of the matter clustering.

In this chapter, we undertake, using large N-body simulations, a quantitative analysis of the cosmological information content of all cosmic environments at linear and non-linear scales (up to $k \sim 0.5\ h$/Mpc) and in both real and redshift spaces. After introducing the N-body simulations from the Quijote suite and how the environments are theoretically and practically defined through the local tidal anisotropies, we explain in what manner the Fisher formalism can be used to assess the constraining power of a statistical representation of an observable.
Equipped with these methodological aspects, we show that the information carried by the two-point statistics of the cosmic environments is superior to a matter power spectrum analysis both in real and redshift spaces leading to a sizeable gain in the constraints put on cosmological parameters of the underlying model. In particular, the combination of the different environmental sensitivities breaks some key degeneracies in several planes and mostly between matter-related parameters like M_ν–σ_8 or M_ν–Ω_m.
After drawing these conclusions, we expose some caveats of the proposed study and discuss their potential impact on the presented results together with how they may limit the perspective of applying such an analysis to observational data.

6.2 Data & Methodology

Table 6.1. Specification of the simulations from the Quijote suite with their denomination and the number of realisations. Underlined values are those changing from the fiducial setup. 2LPT stands for 2nd order Lagrangian Perturbation Theory and ZA for Zel'dovich approximation.

Name	$\Omega_{\rm m}$	$\Omega_{\rm b}$	h	$n_{\rm s}$	σ_8	M_ν	ICs	# of real.
Fiducial	0.3175	0.049	0.6711	0.9624	0.834	0	2LPT	15000
Ω_m^+	0.3275	0.049	0.6711	0.9624	0.834	0	2LPT	500
Ω_m^-	0.3075	0.049	0.6711	0.9624	0.834	0	2LPT	500
Ω_b^+	0.3175	0.051	0.6711	0.9624	0.834	0	2LPT	500
Ω_b^-	0.3175	0.047	0.6711	0.9624	0.834	0	2LPT	500
h^+	0.3175	0.049	0.6911	0.9624	0.834	0	2LPT	500
h^-	0.3175	0.049	0.6511	0.9624	0.834	0	2LPT	500
$n_{\rm s}^+$	0.3175	0.049	0.6711	0.9824	0.834	0	2LPT	500
$n_{\rm s}^-$	0.3175	0.049	0.6711	0.9424	0.834	0	2LPT	500
σ_8^+	0.3175	0.049	0.6711	0.9624	0.849	0	2LPT	500
σ_8^-	0.3175	0.049	0.6711	0.9624	0.819	0	2LPT	500
M_ν^0	0.3175	0.049	0.6711	0.9624	0.834	0	ZA	500
M_ν^+	0.3175	0.049	0.6711	0.9624	0.834	0.1	ZA	500
M_ν^{++}	0.3175	0.049	0.6711	0.9624	0.834	0.2	ZA	500
M_ν^{+++}	0.3175	0.049	0.6711	0.9624	0.834	0.4	ZA	500

6.2.1 The Quijote suite of simulations

Quijote [Villaescusa-Navarro et al., 2020] is a publicly available[1] large suite of N-body simulations. With 44 100 simulations spanning more than a thousand cosmological models, each with multiple realisations, it is the ideal dataset to perform statistical cosmological analyses as it allows to build accurate covariance matrices and compute derivatives for any cosmological representation. Each simulation consists of a set of 512^3 particles (and 512^3 neutrinos in massive neutrinos cases) that are evolved forward in time from $z = 127$ to $z = 0$ using a tree-PM Gadget-3 code [Springel, 2005] in a $L = 1$ Gpc/h size box. The fiducial cosmology is a flat ΛCDM cosmology with parameters consistent with Planck Collaboration VI et al. [2020]: $\Omega_{\rm m} = 0.3175$, $\Omega_{\rm b} = 0.049$, $h = 0.6711$, $n_{\rm s} = 0.9624$ and $\sigma_8 = 0.834$. With these parameters, and assuming a zero mass for neutrinos ($M_\nu = 0$), $15,000$ random realisations are computed. The Quijote suite then provides 500 realisations by varying individually each parameter, fixing the others at their fiducial values. The stepsizes are: $\mathrm{d}\Omega_{\rm m} = 0.010$, $\mathrm{d}\Omega_{\rm b} = 0.002$, $\mathrm{d}h = 0.020$, $\mathrm{d}n_{\rm s} = 0.020$, and $\mathrm{d}\sigma_8 = 0.015$. Additionally, 500 realisations using several sum of neutrinos mass are also computed, with $M_\nu = \sum m_\nu = \{0.1, 0.2, 0.4\}$ eV, that we will refer to as M_ν^+, M_ν^{++}, and M_ν^{+++} cosmologies respectively. All these information about the Quijote suite of simulations are summarised in Table 6.1.

In the present analysis, we aim at studying quantitatively the cosmological constraints obtained from the two-point summary statistics derived in the different cosmic web environments in both real and redshift spaces. To do so, we mimick RSDs (see Sect. 2.2.2) in every simulations by displacing all particles (dark matter particles and neutrinos if any) through Eq. (2.1) along the third Cartesian axis of the box hence assuming the plane parallel approximation.

[1] *https://quijote-simulations.readthedocs.io/en/latest/*

6.2.2 Cosmic web segmentation

In this theoretical work, we make use of the T-web algorithm introduced in Hahn et al. [2007], and later extended by Forero-Romero et al. [2009] to identify cosmic environments through the tidal tensor $\boldsymbol{T}$. Using prescriptions originating from the linear growth of perturbations in the Zel'dovich approximation [Zel'dovich, 1970], this web finder defines environments in a physical way based on the local level of tidal anisotropy. From the discrete set of particle positions, we first rely on a B-spline interpolation scheme [Hockney & Eastwood, 1981; Sefusatti et al., 2016] to estimate the density field $\rho(\boldsymbol{x})$ on an $N_{\rm g}^3$ regular grid. For our purpose, we adopt an interpolation at the order four, namely the Piecewise-Cubic Spline (PCS) scheme, in which the mass of a particle is spread over the $4^3 = 64$ closest cells. By noting $d = N_{\rm g}\|\boldsymbol{x} - \boldsymbol{x}_{\rm p}\|_2/L$ with $\boldsymbol{x}$ the center of a grid cell, $\boldsymbol{x}_{\rm p}$ the particle position and L the size of the box length (assuming a cubic box), PCS weights are given by

$$\begin{cases} (4 - 6d^2 + 3d^3)/6 & \text{if } d \in [0, 1[\,, \\ (2-d)^3/6 & \text{if } d \in [1, 2[\,, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

This choice of interpolation order represents a good trade-off between the accuracy of the reconstructed field and its computational time.

From ρ, one can derive the gravitational potential Φ by solving the Poisson equation

$$\Delta\Phi(\boldsymbol{x}) = 4\pi G\rho(\boldsymbol{x}), \tag{6.2}$$

where Δ is the Laplacian operator and G the gravitational constant. It is convenient to write this equation in terms of the reduced potential $\Phi_r(\boldsymbol{x}) = \Phi(\boldsymbol{x})/4\pi G\bar{\rho}$ so that the Eq. (6.2) satisfies $\Delta\Phi_r(\boldsymbol{x}) = \delta(\boldsymbol{x})$, with $\delta(\boldsymbol{x}) = \rho(\boldsymbol{x})/\bar{\rho} - 1$ the overdensity. Solving this reduced version of the Poisson equation in Fourier space using a discrete approximation of the Laplacian operator (in our case, a 7-point approximation) holds an estimate of $\Phi(\boldsymbol{x})$ on the grid. From the gravitational potential, the tidal tensor can be obtained in each grid cell $\boldsymbol{x}$ as

$$\boldsymbol{T}_{i,j}(\boldsymbol{x}) = \frac{\partial^2\Phi(\boldsymbol{x})}{\partial\boldsymbol{x}_i\partial\boldsymbol{x}_j}, \tag{6.3}$$

leading to the field of eigenvalues $\lambda_1(\boldsymbol{x}) \leq \lambda_2(\boldsymbol{x}) \leq \lambda_3(\boldsymbol{x})$. The cosmic environment associated with a grid cell $\boldsymbol{x}$ is obtained depending on the number of eigenvalues below a given threshold $\lambda_{\rm th}$ as defined in Table 6.2.

The T-web algorithm hence segments the density field at the cell level. To build individual overdensity fields for each environment, we simply propagate the classes at the particle level by assigning the same environment signature to all hosted particles in a given cell. From these four disjoint sets of particles, we build four corresponding overdensity fields $\{\delta_{\rm V}, \delta_{\rm W}, \delta_{\rm F}, \delta_{\rm N}\}$ estimated with the PCS interpolation scheme. The full density field δ is hence decomposed into the four environmental fields and respects the linear combination

$$\delta = f_{\rm V}\,\delta_{\rm V} + f_{\rm W}\,\delta_{\rm W} + f_{\rm F}\,\delta_{\rm F} + f_{\rm N}\,\delta_{\rm N}, \tag{6.4}$$

where f_α denotes the mass fraction of the environment α, namely N_α/N, with N the total number of particles[2]. In Fig. 6.1 is shown this decomposition with the contribution of each

[2]Note that we express here the mass fractions in terms of number of particles since they all have the same mass in the N-body simulations.

Table 6.2. Cosmic web classification rules in the cell $\boldsymbol{x}$ depending on the eigenvalues $\lambda_1 \leq \lambda_2 \leq \lambda_3$ of the tidal tensor.

Environment	Condition
Void	$\lambda_1, \lambda_2, \lambda_3 < \lambda_{\rm th}$
Wall	$\lambda_1, \lambda_2 < \lambda_{\rm th}, \lambda_3 > \lambda_{\rm th}$
Filament	$\lambda_1 < \lambda_{\rm th}, \lambda_2, \lambda_3 > \lambda_{\rm th}$
Node	$\lambda_1, \lambda_2, \lambda_3 > \lambda_{\rm th}$

environments to the overall matter density fields displayed in the top panel for a thin 2.77 Mpc/h depth slice. As expected, nodes describe a discrete set of dense objects found at the extremities of filaments and intersection of walls while voids show large low-density areas covering most of the surface.

In our implementation of the T-web formalism, the potential is being smoothed at a scale of $\sigma_{\mathcal{N}}$ Mpc/h before the classification, resulting in three parameters for the full segmentation procedure that are: $N_{\rm g}$, the total number of grid cells, $\lambda_{\rm th}$ the threshold for the eigenvalues of the tidal tensor and $\sigma_{\mathcal{N}}$. After assessment of the effect of each parameter on the classification in a physically reasonable range (from 1 to 3 Mpc/h), we concluded that the smoothing scale and the number of cells were not of great impact on both the mass and volume fractions of the resulting environments. We consequently adopt for this work $\sigma_{\mathcal{N}} = 2$ Mpc/h and $N_{\rm g} = 360$ leading to a grid size of 2.77 Mpc/h. The $\lambda_{\rm th}$ parameter, however, has a significant effect on the classification, both in terms of volume and mass fractions in the different environments [Forero-Romero et al., 2009]. The impact of this parameter is illustrated in Fig. 6.2 in which are displayed the averaged mass fractions $\langle f_\alpha \rangle$ for each cosmic environment as drawn by the T-web algorithm with three values of $\lambda_{\rm th}$ that are $\{\lambda_{\rm th}^-, \lambda_{\rm th}^{\rm fid}, \lambda_{\rm th}^+\} = \{0.2, 0.3, 0.4\}$. The fiducial value of 0.3 is the one corresponding roughly to the threshold at which voids percolate for our cosmological volume of 1 (Gpc/h)^3 [Forero-Romero et al., 2009]. It is also the value adopted in many other works for classification by means of the T-web prescriptions such as Martizzi et al. [2019]. The left (resp. right) panel of Fig. 6.2 focuses on the real-space (resp. redshift-space) case where an increasing value of $\lambda_{\rm th}$ from 0.2 to 0.4 is attributing more particles in voids and less in filaments and nodes. Comparing the two panels also illustrates that more mass is being associated with filaments and less in nodes in the redshift space. This is mainly due to the Finger-of-God distortions squashing clustered regions and breaking their isotropic nature making them appearing elongated in the line-of-sight direction (see Sect. 2.2.2). Quantitatively, varying the threshold in eigenvalues from 0.2 to 0.4 yields a factor of two between the obtained mass fractions in voids in both real and redshift spaces. Consequently, in order to derive constraints from environments that are robust to uncertainties in the identification of the environments and also indirectly to changes in the definitions offered by the variety of methods [see e.g. Libeskind et al., 2017], we embed $\lambda_{\rm th}$ as a nuisance parameter in the formalism such that all presented constraints are marginalised over it.

6.3 Environments sensitivity to cosmology

6.3.1 Cosmic fractions as a function of cosmological parameters

The broad range of densities probed by the different environments at $z = 0$ hinted by Fig. 6.1 and pointed out more quantitatively by other works [Forero-Romero et al., 2009; Cautun

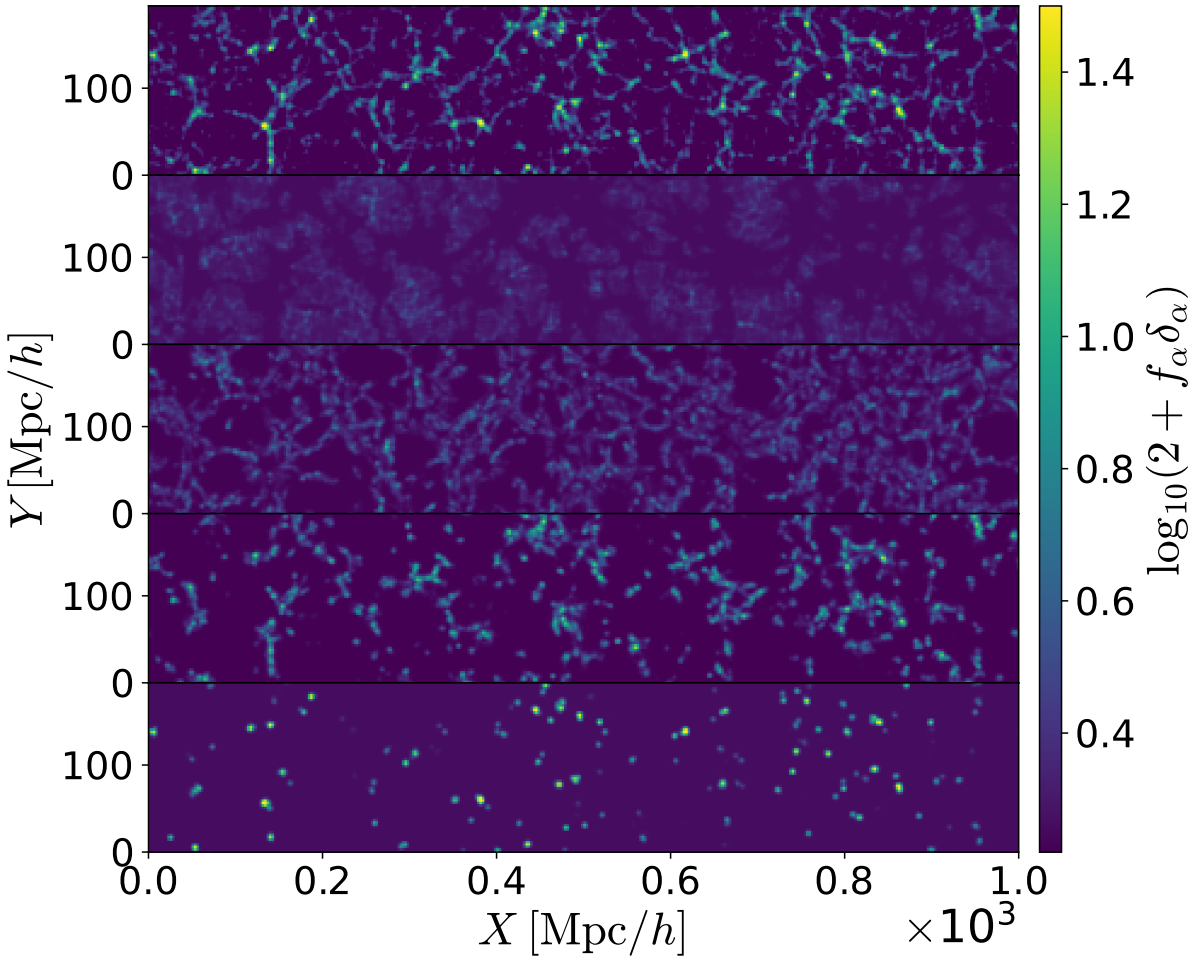

Fig. 6.1. A 2.77 Mpc/h depth slices showing the full field δ in the top panel and the different fields $\{\delta_{\rm V}, \delta_{\rm F}, \delta_{\rm W}, \delta_{\rm N}\}$ computed from the T-web classification of particles described in Sect. 6.2.2, all estimated with the PCS interpolation scheme. Environmental fields have been normalised such that $\delta = \sum_\alpha f_\alpha \delta_\alpha$.

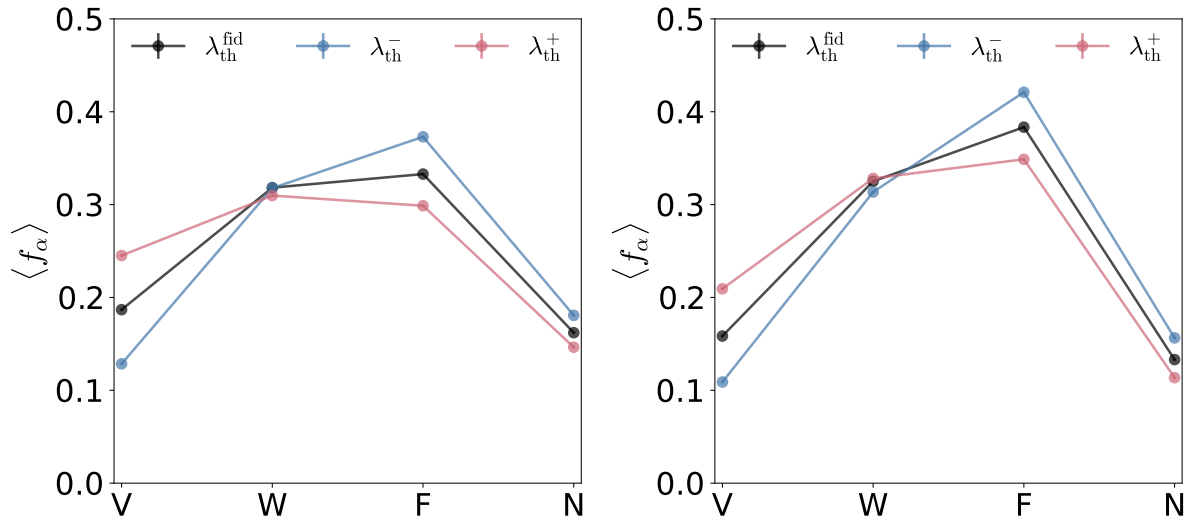

Fig. 6.2. Averaged mass fractions $\langle f_\alpha \rangle$ in the T-web environments for distinct values of $\{\lambda_{\rm th}^-, \lambda_{\rm th}^{\rm fid}, \lambda_{\rm th}^+\} = \{0.2, 0.3, 0.4\}$ in real (left panel) and redshift (right panel) spaces.

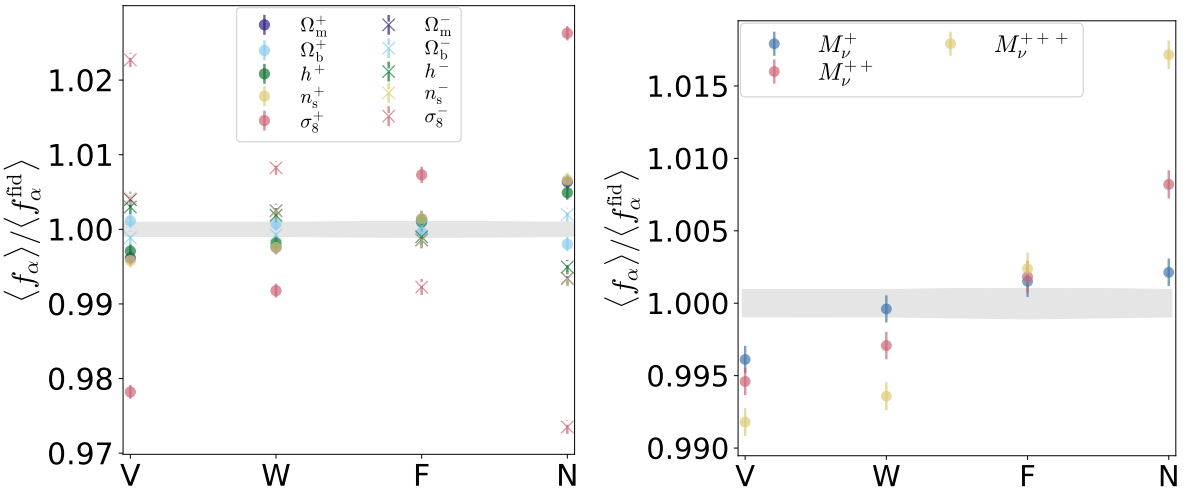

Fig. 6.3. Ratio of environmental mass fractions between different cosmologies and the fiducial ones for an eigenvalue threshold of $\lambda_{\rm th}^{\rm fid}$. Points are centred on the average over the $N_{\rm deriv} = 500$ realisations for each cosmology and error bars show the $\pm 3\sigma$ interval. The grey shaded area corresponds to the $\pm 3\sigma$ interval of the fiducial cosmology fractions obtained from the $N_{\rm fid} = 7000$ realisations.

et al., 2014; Libeskind et al., 2017] reflects their different gravitational histories (see also Sect. 2.4.1). Voids are for instance environments of low-density spanning most of the volume while nodes are highly non-linear and dense objects enclosing a similar value of the mass as voids but in only few percents of the volume. Filaments span a broad range of densities from long and tenuous arteries to small and dense bridges linking clusters together. Figure 6.3 shows the ratio between the average mass fractions in each environment when a cosmological parameter varies and the average obtained with the fiducial cosmology. The error bars (represented as the bars around points and crosses and as the grey shaded area for fiducial simulations) are the 3σ confidence intervals. Many parameters are causing sizeable changes in these proportions and, unsurprisingly, parameters related to matter density, like σ_8 and $\Omega_{\rm m}$, have the largest impact. The most important variations are induced by σ_8, for which an increase (resp. decrease) is leading to a larger (resp. smaller) mass fraction in dense environments. This is in agreement with the definition of σ_8 that is measuring how matter clusters at a scale of 8 Mpc/h. The impact of neutrino mass, even though smaller in comparison, is still significant and fraction ratios lie outside the 3σ confidence regions. When M_ν increases, it basically leads to make dense environments more massive, similarly to σ_8. All these different effects observed in the mass fractions already suggest that each cosmological parameter has a different impact on the environments that will be even more refined when inspecting the clustering statistics in each environment through the power spectra.

6.3.2 Power spectra in cosmic environments

The auto power spectrum $P^{\alpha\alpha}(k)$ is defined as the covariance of Fourier modes (see Sect. 1.3.2) of the density field δ_α, with $\alpha \in \{\mathrm{V}, \mathrm{W}, \mathrm{F}, \mathrm{N}\}$ denoting a given environment. More generally, for two overdensity fields δ_α and δ_β, the cross power spectrum is given by

$$P^{\alpha\beta}(k)\delta_{\rm D}(\boldsymbol{k}_1 + \boldsymbol{k}_2) = \frac{1}{(2\pi)^3}\langle \tilde{\delta}_\alpha(\boldsymbol{k}_1)\tilde{\delta}_\beta(\boldsymbol{k}_2)\rangle, \tag{6.5}$$

with $k = \|\boldsymbol{k}_1\|_2$, and $\tilde{\delta}$ referring to the Fourier transform of δ. In redshift space, the peculiar velocities of sources is causing a dependence of the power-spectrum with the line-of-sight inducing a breaking of the density field isotropy property which consequently alters the spatial pattern (see Sect. 2.2.2). This effect can be taken into account using a multipole expansion of the power spectrum expressed as

$$P_\ell^{\mathrm{s},\alpha\beta}(k) = \frac{2\ell+1}{2}\int_{-1}^{1} P^{\alpha\beta}(k,\mu)\mathcal{L}_\ell(\mu)\mathrm{d}\mu, \tag{6.6}$$

with $\mu = \boldsymbol{k}\cdot\hat{\boldsymbol{n}}/k$, the angle with the line of sight, $P^{\alpha\beta}(k,\mu)$ the 2D power spectrum obtained by binning both in k and μ, and $\mathcal{L}_\ell$ the Legendre polynomials. In this work, we rely on the three first non-zero multipoles of the power spectrum in redshift-space, $P_{\ell=0}^{\mathrm{s},\alpha\beta}(k)$, $P_{\ell=2}^{\mathrm{s},\alpha\beta}(k)$, $P_{\ell=4}^{\mathrm{s},\alpha\beta}(k)$, respectively called monopole, quadrupole and hexadecapole. These $\ell \leq 4$ orders are the only non-vanishing moments in the linear approximation of the distortions and encode the full 2D information at linear scales [Kaiser, 1987]. The corresponding Legendre polynomials are

$$\mathcal{L}_\ell(\mu) = \begin{cases} 1 & \text{if } \ell = 0, \\ \left(3\mu^2 - 1\right)/2 & \text{if } \ell = 2, \\ \left(35\mu^4 - 30\mu^2 + 3\right)/8 & \text{if } \ell = 4. \end{cases} \tag{6.7}$$

Note that Eq. (6.6) generalises Eq. (6.5) and, under isotropy conditions for the overdensity fields, all poles $P_\ell(k)$ with $\ell > 0$ are exactly 0. To avoid confusion, we refer to the real-space monopole as simply $P(k)$ and leave the subscript ℓ for redshift-space spectra, in addition to the superscript s. In the present study, in expressions (6.5) and (6.6) is in fact not directly appearing the overdensity field δ but its deconvolved evaluation. Indeed, since δ is estimated by the PCS smoothing interpolation scheme, it deforms the shape of the estimated power spectrum [Jing, 2005]. To correct for this effect, we first deconvolve the fields δ_α by applying the window function in Fourier space

$$W(k) = \left[\prod_i \left(1 - \frac{4}{3}s_i + \frac{2}{5}s_i^2 - \frac{4}{315}s_i^3\right)\right]^{-1}, \tag{6.8}$$

with $s_i = \sin\left(\pi k_i / 2k_{\mathrm{Nyq}}\right)$ and $k_{\mathrm{Nyq}} = \pi N_{\mathrm{g}}/L$ the Nyquist frequency. Additionally, to avoid any bias induced by aliasing effects, we restrict the analysis to the modes of $P(k)$ below half of the Nyquist frequency, $k_{\mathrm{Nyq}}/2 = 0.57\ h/\mathrm{Mpc}$. In our case, $k_{\mathrm{max}} = 0.5\ h/\mathrm{Mpc}$, allowing us to take into account both large-scales and non-linear ones in a robust manner. Finally, because of the discrete nature of the input, namely the dark matter particles, we also subtract the shot noise from power spectra estimated (6.5) and (6.6). Even though the number of particles is very large and we expect the shot noise contribution to be small at the scales of interest, auto-spectra $P^{\alpha\alpha}$ (including also P^{mm}) are subtracted by the quantity $1/\bar{n}_\alpha$ where $\bar{n}_\alpha = N_\alpha/V$.

6.4 Constraining power of cosmic environments

6.4.1 Fisher formalism for information content quantification

Considering a set of model parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ (in our case, cosmological parameters), we assume that the vector $\boldsymbol{s}(\boldsymbol{X}) \in \mathbb{R}^n$ is a statistic built from an observable (here, the binned power

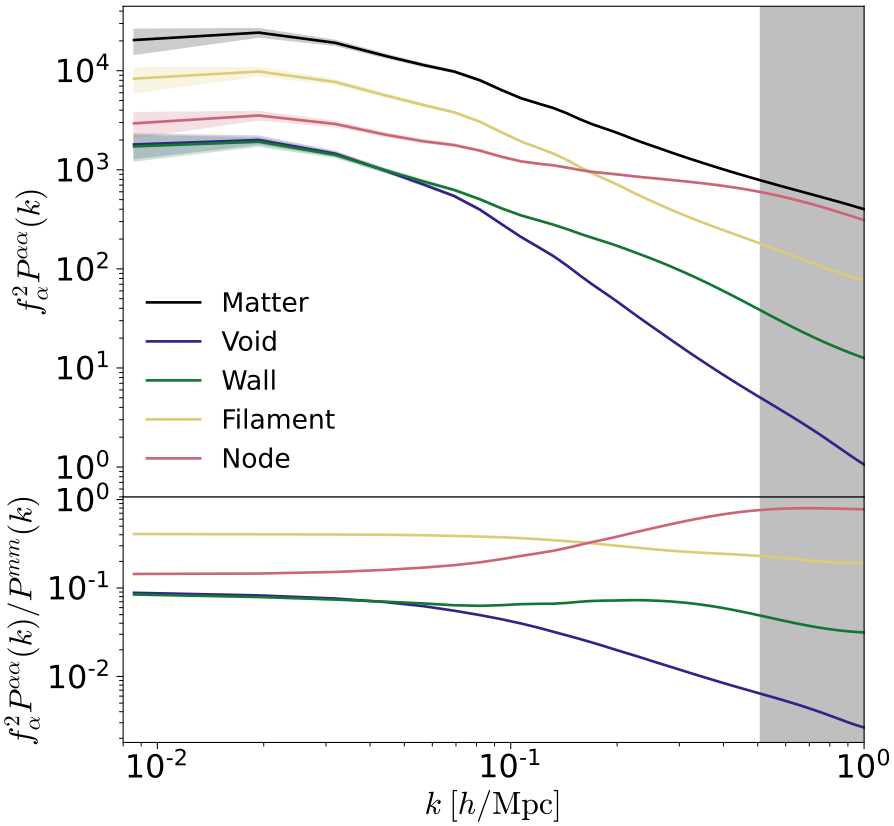

Fig. 6.4. The top panel shows the normalised power spectra $f_\alpha^2 P^{\alpha\alpha}(k)$ obtained in real space for each environment $\alpha \in \{\mathrm{V}, \mathrm{W}, \mathrm{F}, \mathrm{N}\}$ compared to the one from all dark-matter particles in black. Plain lines are the averages over the $N_{\mathrm{fid}} = 7,000$ fiducial simulations and the shaded areas are the 1-σ confidence intervals. The grey area depicts the range of $k > k_{\mathrm{max}}$ excluded from the analysis. In the bottom panel are found the ratios between the averaged normalised spectra in the environments and the matter one (black from the top panel).

spectra drawn from the overdensity fields) following a Gaussian distribution $\boldsymbol{s} \sim \mathcal{N}(\bar{\boldsymbol{s}}, \boldsymbol{\Sigma})$. Its log-likelihood can hence be written

$$\log p(\boldsymbol{s} \,|\, \boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{s} - \bar{\boldsymbol{s}})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{s} - \bar{\boldsymbol{s}}) - \frac{1}{2} \log |\boldsymbol{\Sigma}| + \mathrm{cst}, \tag{6.9}$$

where the constant comes from the normalisation of the distribution. A general way to quantify the information carried by $\boldsymbol{s}$ on $\boldsymbol{\theta}$ is to use the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$. From the Fréchet-Darmois-Cramér-Rao inequality, its inverse $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$ corresponds to a lower-bound on the variance of any unbiased estimator drawn from $\boldsymbol{s}$ hence assessing the efficiency of the representation. Elements of the Fisher matrix are defined as the variance of the derivative of the log-likelihood, namely

$$\boldsymbol{I}(\boldsymbol{\theta})_{i,j} = \mathbb{E}_\theta \left[\left(\frac{\partial \log p(\boldsymbol{s} \,|\, \boldsymbol{\theta})}{\partial \theta_i} \right)^{\mathsf{T}} \left(\frac{\partial \log p(\boldsymbol{s} \,|\, \boldsymbol{\theta})}{\partial \theta_j} \right) \right], \tag{6.10}$$

which can also be written in terms of the 2$^{\mathrm{nd}}$ derivative of the log-likelihood under some smoothness constraints (which are fulfilled in the Gaussian case),

$$\boldsymbol{I}(\boldsymbol{\theta})_{i,j} = -\mathbb{E}_\theta \left[\frac{\partial^2 \log p(\boldsymbol{s} \,|\, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right], \tag{6.11}$$

where $\mathbb{E}_\theta$ is the expectation taken over the distribution $p(\boldsymbol{s} \mid \boldsymbol{\theta})$. This latter equation is intuitively explaining how the amount of information is measured. A sharp log-likelihood around $\boldsymbol{\theta}$ implies a huge increase with small changes of the parameters, making the representation $\boldsymbol{s}$ very sensitive to variations $\mathrm{d}\theta$. On the other hand, a weakly-curved log-likelihood with a locally flat behaviour advocates for a poor description since its sensitivity with changes in the parameters is small. Under the Gaussian assumption described above and by further considering a covariance matrix $\boldsymbol{\Sigma}$ independant from cosmological parameters $\boldsymbol{\theta}$, mainly because this contribution is expected to be small and source of underestimation of errors [Carron, 2013; Kodwani et al., 2019], it yields

$$\boldsymbol{I}(\boldsymbol{\theta})_{i,j} = \left(\frac{\partial \bar{\boldsymbol{s}}}{\partial \theta_i}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \left(\frac{\partial \bar{\boldsymbol{s}}}{\partial \theta_j}\right). \tag{6.12}$$

The non-linear operation of the inversion to compute the precision matrix $\boldsymbol{\Sigma}^{-1}$ is actually leading to a biased estimate, even though the covariance may be computed using the classical unbiased estimation. Still under the previously-established Gaussian assumptions, the unbiased estimate of the precision matrix is given by [Kaufman, 1964; Hartlap et al., 2007]

$$\boldsymbol{\Sigma}^{-1} = \frac{N_{\mathrm{fid}} - n - 2}{N_{\mathrm{fid}} - 1} \hat{\boldsymbol{\Sigma}}^{-1}, \tag{6.13}$$

where N_{fid} is the number simulations at the fiducial cosmology, n is the length of the summary statistics vector $\boldsymbol{s}$ and $\hat{\boldsymbol{\Sigma}} = (\boldsymbol{s} - \bar{\boldsymbol{s}})\,(\boldsymbol{s} - \bar{\boldsymbol{s}})^{\mathsf{T}} \,/\, (N_{\mathrm{fid}} - 1)$ is the unbiased estimate of the covariance matrix.

The partial derivatives of the summary statistics with respect to parameters of the model can be computed using the variations of cosmologies provided by the Quijote suite of simulations. Considering the set of studied cosmological parameters $\{\Omega_{\mathrm{m}}, \Omega_{\mathrm{b}}, h, n_{\mathrm{s}}, \sigma_8\}$, we can estimate numerically

$$\frac{\partial \bar{\boldsymbol{s}}}{\partial \theta_i} \simeq \frac{\bar{\boldsymbol{s}}(\theta_i + \mathrm{d}\theta_i) - \bar{\boldsymbol{s}}(\theta_i - \mathrm{d}\theta_i)}{2\mathrm{d}\theta_i}. \tag{6.14}$$

In the case of massive neutrino simulations, $M_\nu \geq 0$ with a fiducial value at 0.0 eV. For this parameter, we thus cannot rely on Eq. (6.14) and instead estimate the derivative using the four-point forward approximation

$$\frac{\partial \bar{\boldsymbol{s}}}{\partial M_\nu} \simeq \frac{\bar{\boldsymbol{s}}(4M_\nu^+) - 12\bar{\boldsymbol{s}}(2M_\nu^+) + 32\bar{\boldsymbol{s}}(M_\nu^+) - 21\bar{\boldsymbol{s}}(M_\nu = 0.0)}{12M_\nu^+}. \tag{6.15}$$

Because massive neutrino simulations in Quijote are initialised using the Zel'dovich approximation and fiducial ones using the 2$^{\mathrm{nd}}$ order Lagrangian perturbation theory, the quantity $\bar{\boldsymbol{s}}(M_\nu = 0.0)$ is computed using the fiducial simulations initialised with the Zel'dovich approximation also. In all the presented results, if not mentioned otherwise, the numerical estimation of derivatives and covariances have been respectively made with $N_{\mathrm{deriv}} = 500$ and $N_{\mathrm{fid}} = 7000$ realisations. In Sect. 6.4.4, we discuss the impact of these numbers and assess the numerical stability of the results.

6.4.2 Real-space auto-spectra

We first study the constraining power of summary statistics derived from the set of power spectra in distinct environments P^{VV}, P^{WW}, P^{FF}, P^{NN} and their combination, P^{comb}, for cosmological models in real space. The two key ingredients of the Fisher-based quantification of information appear in Eq. (6.12) as:

(i) The partial derivatives of the statistics with respect to the cosmological parameters. Intuitively, a summary presenting large variations with parameters of the model carries more information than if it barely varies;

(ii) The covariance matrix $\mathbf{\Sigma}$ measuring the relations between elements of the summary statistics. Naturally, having elements varying the same way limits the overall informative power of the summary.

Figure 6.5 shows the first constituent, the derivatives $\partial P^{\alpha\alpha}(k)/\partial\theta_i$ while Fig. 6.6 plots a proxy of the second ingredient through the normalised version of the covariance matrix, namely the correlation matrix $\boldsymbol{C}$, whose elements are defined as

$$C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \tag{6.16}$$

The first striking observation when inspecting the correlation matrix for $P^{\rm mm}$ is that it quickly becomes highly non-diagonal with correlation coefficients $C_{ij} = 0.5$ at scales of $\sim 0.3\ h$/Mpc. Such high couplings between scales are expected at low redshifts, Fourier modes being more and more correlated with time, as a result of the non-linear evolution of the matter distribution [Blot et al., 2015]. These non-diagonal terms are intrinsically reducing the representative power of the matter power spectrum to constrain cosmological parameters, independently of how it varies with these latter. This insufficiency of $P^{\rm mm}(k)$ is also marked by the shape of the derivatives, quite monotonous with similar structures for most parameters, except $\Omega_{\rm m}$, $\Omega_{\rm b}$ and M_ν where the wiggles are signatures of an impact on the baryonic acoustic oscillations. On the opposite side, spectra drawn from cosmic environments are showing different patterns in the derivatives. Taking the example of the $\Omega_{\rm m}$ parameter in the top left panel of Fig. 6.5, the change of sign occurs at different scales, which seem to follow the order of average density, namely from void to node. This pattern is also observed for other parameters like $\Omega_{\rm b}$, h, $n_{\rm s}$ or M_ν. The correlation coefficients of the environmental statistics visible in the bottom panel of Fig. 6.6 also display less off-diagonal cross-correlation coefficients of high values, at the exception of the node-node case. Indeed, $P^{\rm NN}$ Fourier modes are highly correlated with values $C_{k_1k_2} \sim 0.5$ for $k_1, k_2 \sim 0.2\ h$/Mpc, which is a signature of the highly non-linear environment it represents.

These overall observations suggest that the environments bring different information on the set of cosmological parameters, and that, when combined together, they may break degeneracies and allow to put tighter constraints on the underlying cosmological model. One way to quantify this gain is the Fisher formalism in which the marginalised 1σ confidence ellipses can be obtained from the information matrices (6.12). These latter are shown in Fig. 6.7 when using the statistics of the matter power spectrum or the one from each environment either individually or combined all together. Table 6.3 delivers the complementary information of the marginalised σ_{θ_i} constraints obtained in the different cases and defined as

$$\sigma_{\theta_i} = \frac{1}{\sqrt{[\boldsymbol{I}(\boldsymbol{\theta})^{-1}]_{ii}}}, \tag{6.17}$$

where I_{ii} is the ith diagonal element of the Fisher information matrix. As already hinted by the analysis of the partial derivatives and the correlation matrix, the marginalised errors obtained from the matter-matter power spectrum $P^{\rm mm}$ are high. The corner plot especially shows us the degeneracies among parameters in almost all panels, observed for instance in the M_ν–σ_8 panel for which the matter power spectrum behaves similarly when varying either M_ν or σ_8, making it difficult for the $P^{\rm mm}$ statistics to disentangle the two effects. When inspecting the

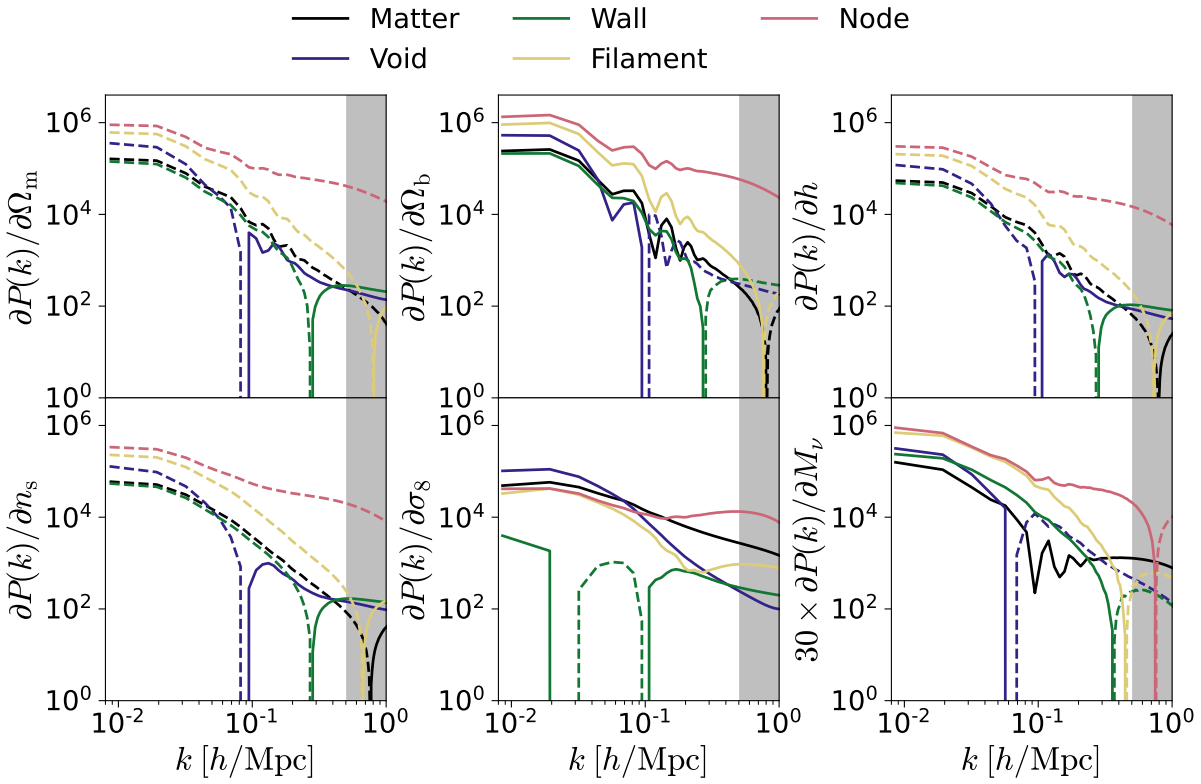

Fig. 6.5. Partial derivatives $\partial P^{\alpha\alpha}(k)/\partial\theta_i$ for the different environmental auto-spectra and for each studied parameters $\{\Omega_{\rm m}, \Omega_{\rm b}, h, n_{\rm s}, \sigma_8, M_\nu\}$. Dashed (resp. plain) lines correspond to negative (resp. positive) values of the derivative. The grey area depicts the range of $k > k_{\rm max}$ excluded from the analysis.

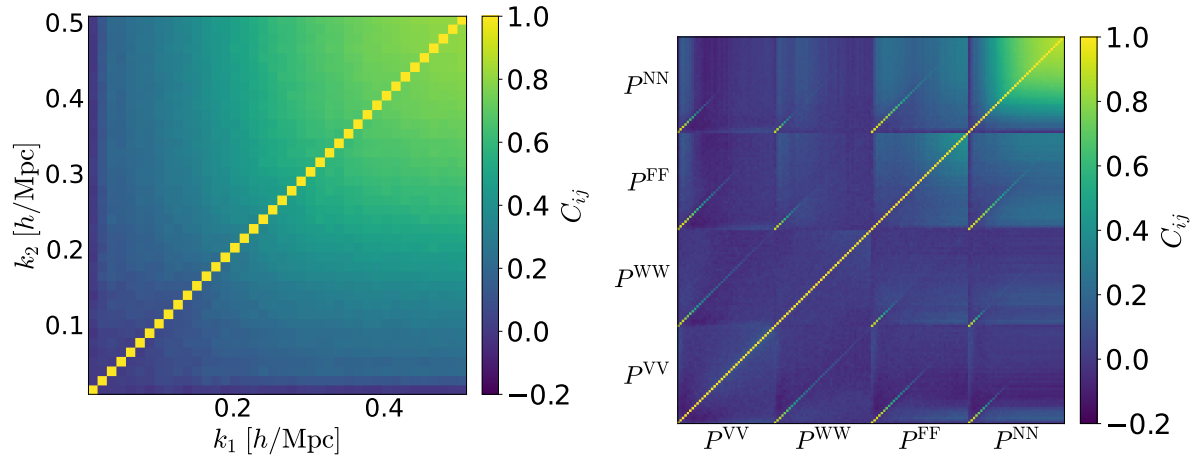

Fig. 6.6. *(left)* Correlation coefficients C_{ij} for the matter power spectrum in real space. *(right)* Same for $P_\ell^{\alpha\alpha}(k)$ coefficients extracted from the several environments. Each sub-matrix goes from $k = 0.1\ h/{\rm Mpc}$ to $k = k_{\rm max} = 0.5\ h/{\rm Mpc}$.

ellipses obtained from the individual spectra from the environments, we clearly distinguish different orientations for several parameters, such as the void and filament environments in the M_ν–Ω_m, M_ν–σ_8 or M_ν–Ω_b projected spaces. This observation is emphasised by Fig. 6.8 in which we show a zoom over the M_ν–Ω_m and M_ν–σ_8 planes. This suggest a different type of information delivered by the two environments, which, when combined all together, tightens up the constraints as quantitatively shown in Table 6.3 with improvement factors from 4.5 to 17.1 for all the five cosmological parameters considered and 15 for the sum of neutrino mass. Unsurprisingly, the constraints put by individual environments are, to a lesser extent, also better than a matter power spectrum analysis in real space. As discussed previously, we however see the broad ellipses and hence poor constraints obtained from the node environment due to the high level of correlations between the P^NN coefficients, which, together with the parameter degeneracies, lead to high errors on cosmological parameters. Focusing on M_ν, we also see that the best constraints are obtained by the void environment, as theoretically expected and stated in previous works [Pisani et al., 2015; Massara et al., 2015; Kreisch et al., 2019].

In White [2016] is exposed a way to build an estimate of the matter power spectrum weighted by the local density as

$$m(\boldsymbol{x}, R, p, \rho_\star) = \left[\frac{1+\rho_\star}{\rho_\star + \rho_R(\boldsymbol{x})}\right]^p, \tag{6.18}$$

where ρ_R is the density field smoothed by a top-hat filter of radius R, p the factor of enhancement of low (resp. high) density areas if $p > 0$ (resp. $p < 0$) and $\rho_\star$ a density parameter to tune. Confronting results from Table 6.3 to those obtained with the marked power spectrum analysis of Massara et al. [2021] in real-space shows that the combination of auto-spectra in cosmic environments yields similar constraints for all the studied parameters of the simulation. Similarly, Bayer et al. [2021] constrains the same set of cosmological parameters using the real-space matter power spectrum from Quijote together with the void size and halo mass functions (respectively VSF and HMF). They report 1σ marginalised errors of $\{0.0063, 0.037, 0.23, 0.10, 0.0069, 0.096\}$ on the set of parameters $\{\Omega_\mathrm{m}, \Omega_\mathrm{b}, h, n_\mathrm{s}, \sigma_8, M_\nu\}$ meaning that the combination of auto-spectra in environments performs better in most cases with improvement factors of $\{0.5, 4.1, 3.0, 3.4, 3.5, 1.6\}$ compared to the VSF+HMF statistics in real space. It is also noteworthy that λ_th is set free in the analysis and Fig. 6.7 shows that this parameter is well-constrained by most environments. Quantitatively, voids, walls, filaments, nodes and their combination are respectively leading to marginalised error over the threshold of $\sigma_{\lambda_\mathrm{th}} = \{0.0046, 0.0090, 0.0071, 0.0100, 0.0006\}$, indicating that the results are robust to changes in this threshold. Even though λ_th may influence the identified cosmic structures, it does not much affect the constraints. This is also encouraging in the sense that it leaves room to other definitions of cosmic environments to be applied and still ending up with similar results.

These obtained constraints are nonetheless dependant on the maximum scale involved for the power spectra coefficients k_max. Figure 6.9 illustrates this dependency by showing the evolution of σ_{θ_i} for all the parameters and derived statistics with the value of k_max. The first conclusion that we can draw is that the information extracted from P^mm is saturating when k_max increases. This saturation when going to smaller scales, pointed out by previous analyses [Takahashi et al., 2010; Blot et al., 2015; Chan & Blot, 2017], is mostly induced by the degeneracies among parameters for this precise summary statistics which do not lead to any further improvement on the constraints at mildly non-linear scales when $k_\mathrm{max} > 0.25$ h/Mpc. The individually smaller errors obtained in the environments are not observed at all

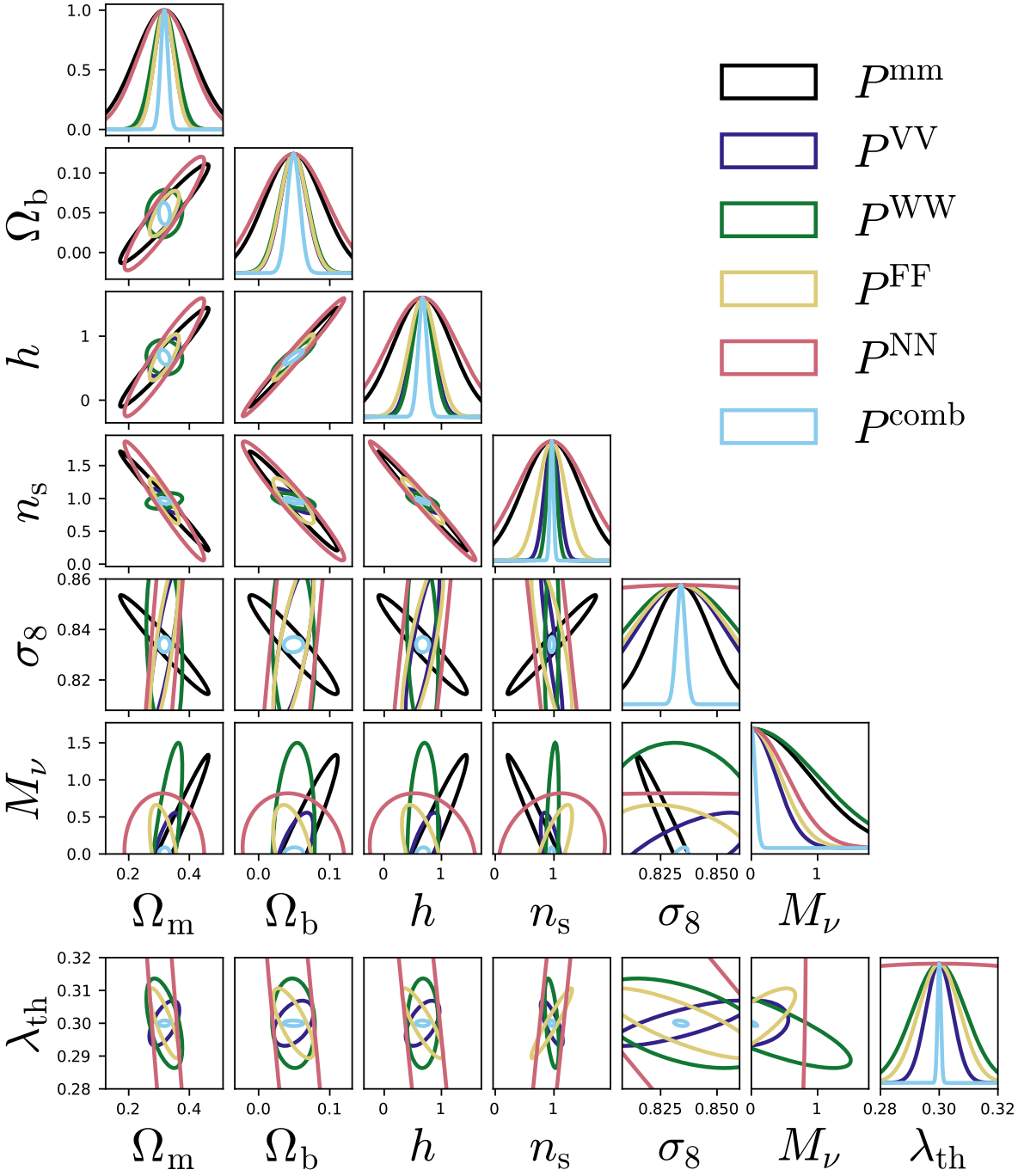

Fig. 6.7. 1σ confidence ellipses for all the pairs of cosmological $(\Omega_{\rm m}, \Omega_{\rm b}, h, n_{\rm s}, \sigma_8, M_\nu)$ and nuisance $(\lambda_{\rm th})$ parameters obtained from the matter power spectrum, or the one from the different environments and their combination in real space.

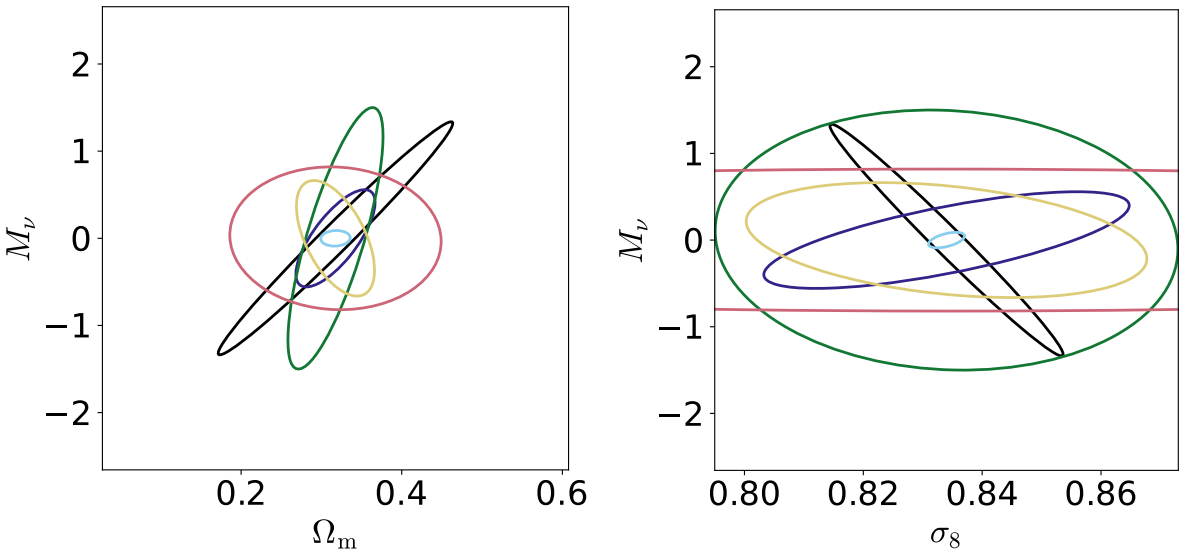

Fig. 6.8. Zoomed confidence ellipses obtained with P^{mm}, P^{VV}, P^{WW}, P^{FF}, P^{NN} and their combination, P^{comb}, in the M_ν–Ω_m (left panel) and M_ν–σ_8 (right panel) planes in real space.

Table 6.3. Marginalised 1-σ constraints obtained from the power spectrum in different environments for all cosmological parameters.

Statistics	$\sigma_{\Omega_\mathrm{m}}$	$\sigma_{\Omega_\mathrm{b}}$	σ_h	σ_{n_s}	σ_{σ_8}	σ_{M_ν}
P^{mm}	0.0967	0.0412	0.5128	0.4998	0.013	0.8857
P^{VV}	0.0327 (3.0)	0.0184 (2.2)	0.1960 (2.6)	0.1244 (4.0)	0.0204 (0.6)	0.3695 (2.4)
P^{WW}	0.0392 (2.5)	0.0199 (2.1)	0.1800 (2.8)	0.0796 (6.3)	0.0258 (0.5)	0.9938 (0.9)
P^{FF}	0.0322 (3.0)	0.0190 (2.2)	0.2399 (2.1)	0.0223 (2.2)	0.0225 (0.6)	0.4392 (2.0)
P^{NN}	0.0872 (1.1)	0.0473 (0.9)	0.6117 (0.8)	0.6005 (0.8)	0.1208 (0.1)	0.5427 (1.6)
P^{comb}	0.0122 (**8.0**)	0.0091 (**4.5**)	0.0766 (**6.7**)	0.0292 (**17.1**)	0.0020 (**6.5**)	0.0592 (**15.0**)

scales. In particular, when restricting the analysis to $k_{\mathrm{max}} < 0.2\ h$/Mpc, environments are not individually constraining better the set of cosmological parameters, even though their combination still lead to an improvement over the matter-matter analysis, for all considered values of k_{max}. It is however interesting to note the better performance of the mass-dominating environment, the filaments, that are, at large scales $k_{\mathrm{max}} < 0.2\ h$/Mpc, bringing the tightest constraints for almost all parameters. When going to smaller scales, filaments start to saturate and the volume-dominating environments, namely voids and walls are providing similar, if not tighter, constraints for most of the studied parameters.

Another quantity of interest to measure the power of a representation is given by the signal-to-noise ratio (hereafter SNR) that describes the reachable accuracy of the statistics measurement given the covariance matrix. In general, the SNR of a summary statistic $\boldsymbol{s} \in \mathbb{R}^n$ is defined as

$$\mathrm{SNR}(\boldsymbol{s}) = \sqrt{\boldsymbol{s}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{s}}, \tag{6.19}$$

with $\boldsymbol{\Sigma}^{-1}$ the corresponding precision matrix defined by Eq. (6.13). Figure 6.10 shows the evolution of the SNR for the matter and environmental power spectra as a function of the

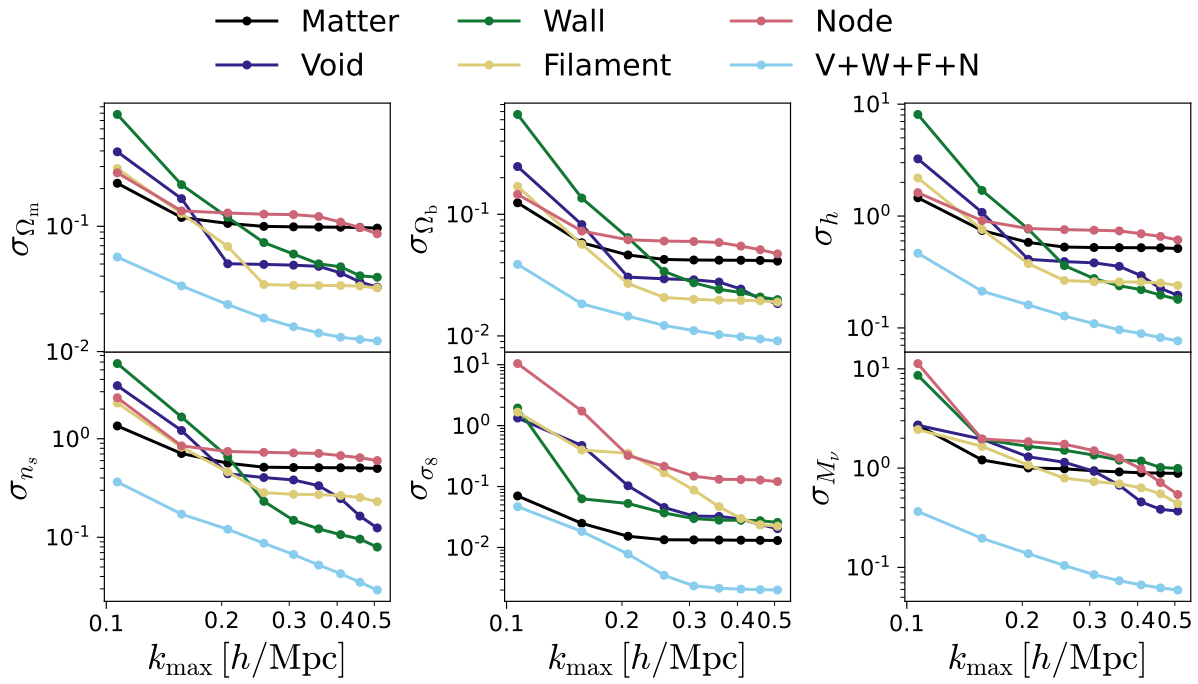

Fig. 6.9. Evolution of the marginalised constraint σ_{θ_i} put on cosmological parameters $\{\Omega_\mathrm{m}, \Omega_\mathrm{b}, h, n_\mathrm{s}, \sigma_8\}$ and the sum of neutrino mass M_ν in real space with the maximum scale used for the Fisher analysis, namely k_max.

maximum scale k_max. We again observe the flattening when k_max approaches non-linear scales at $0.25\ h$/Mpc for $\mathrm{SNR}(P^\mathrm{mm})$. Although environment-dependent spectra also suggest such a plateau, it happens at much lower scales and higher values of the SNR, except for nodes again which saturate at the lowest value among all of the studied statistics. The shapes of the SNR evolution with k_max from environments combination also suggest that there is room left for a further increase when going to even smaller scales, where the matter analysis will not be able to improve. Quantitatively, the SNR obtained from the combination of environments is eight times higher than the one from the matter auto-spectrum at $k_\mathrm{max} = 0.5\ h$/Mpc and explains also partly the better constraints put on cosmological parameters from Fig. 6.9 and Table 6.3.

6.4.3 Redshift-space auto-spectra

The real-space results demonstrate the power of the cosmic web in breaking degeneracies among cosmological parameters to bring, at all scales, significant improvements over the matter power spectrum constraints. This latter shows quick saturation at mildly non-linear scales which limits the efficiency of this statistics in constraining parameters of the model, even when pushing the analysis to small scales. Observations however usually rely on sources mapped in the sky and for which the redshift is used as a measure of distance. The so-called redshift space is well-known to have a significant impact on the matter clustering statistics. The two main differences between the real monopole $P(k)$ and its redshift-space counterpart $P^{\mathrm{s,mm}}_{\ell=0}(k)$ are (see Fig. 2.3 for an illustration): (i) a power boost at large scales due to the coherent motion of matter escaping from voids and moving towards dense regions [Kaiser, 1987]; (ii) a decrease of power at small scales where Finger-of-God effect is spreading particles around initially residing in a spherical overdensity [Jackson, 1972]. These two effects are clearly observed when comparing the redshift and real space matter power spectra shown in Fig. 6.11. Spectra de-

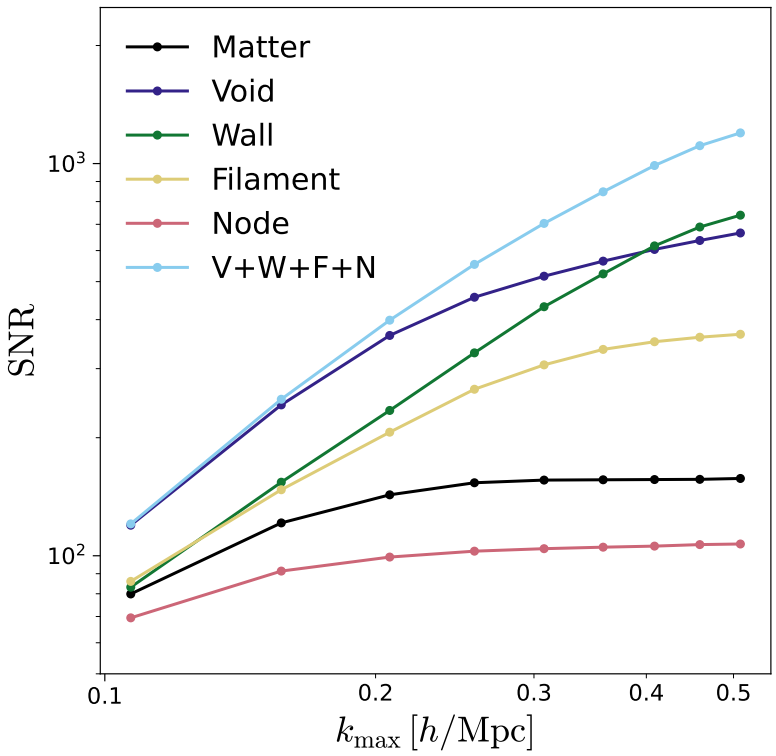

Fig. 6.10. Evolution of the SNR with the maximum scale for the power spectra used in the Fisher analysis in real space, namely $k_{\rm max}$.

rived from cosmic environments in redshift space, obtained by applying the same procedure as in real space with identical parameters, are shifted at large scales. This shift appears as a power boost for filaments and walls and as a decrease of power for voids and nodes with respect to their real-space counterparts. These two latter environments are the most subject to the decrease of power at all scales. Unexpectedly, the node environment is the most affected by RSDs at small scales, with a considerable power decrease due to the Finger-of-God effect, while the damping is much less marked and occurs at smaller scales for filaments and walls. It is also particularly emphasised for nodes since they may also be more difficult to detect by the classification procedure since they now appear as elongated structures in the line-of-sight (see for instance Fig. 2.4 and corresponding discussion). The individual impact of each cosmological parameter on the matter power spectrum is also different in redshift space. As an example, due to the distortions in the line-of-sight direction, a higher value of σ_8 is not implying a simple shift of the spectrum as in real space. Instead, it suppresses additional power at small-scales (large k) due to the Finger-of-God distortions inside non-linear clustered structures. It has been shown that the effect of massive neutrinos can be mimicked by a decrease of σ_8 on the redshift-space monopole [see e.g. Villaescusa-Navarro et al., 2018; Hahn et al., 2020], which in turn has a similar impact as a shift of $\Omega_{\rm m}$. This is illustrated in the left panel of Fig. 6.12 showing the residuals $P^{\rm s,mm}_{\ell=0}(k)^{\theta_i}/P^{\rm s,mm}_{\ell=0}(k)^{\rm fid} - 1$ with θ_i being either σ_8^-, $\Omega_{\rm m}^-$ or M_ν^+. The shape dependencies of the matter power spectrum for variations of σ_8 and M_ν have been shown similar at scales $k > 0.1\ h$/Mpc in Villaescusa-Navarro et al. [2018] and it is possible to find a value of σ_8 fitting at some percent levels the spectrum obtained for M_ν^+. Monopoles obtained in the cosmic environments however show various dependencies for changes in σ_8 and M_ν, as delineated in the right panel of Fig. 6.12. Consequently, a change in σ_8 cannot reproduce the effect of massive neutrinos in all environments, inducing a breaking of degen-

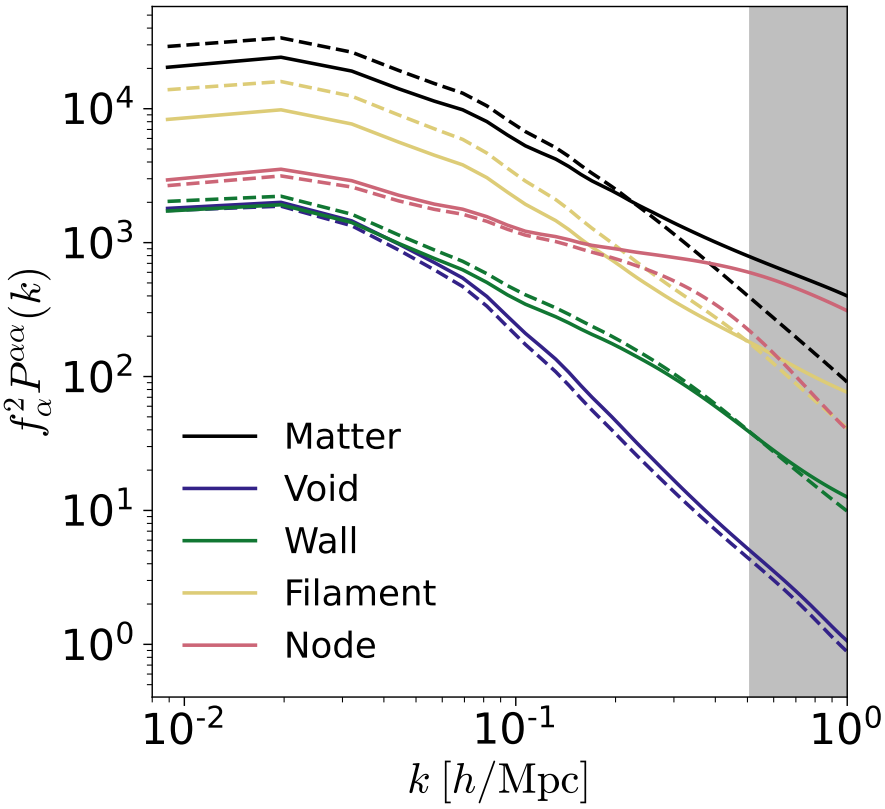

Fig. 6.11. Monopoles in real (plain lines) and redshift (dashed lines) spaces averaged over the $N_{\rm fid}$ realisations for the matter and the different environments. Monopoles of the environments are individually normalised by their mass fractions f_α to see each contribution to the matter monopole.

eracy between the two parameters. The differences between real and redshift monopoles can be appreciated when comparing the derivatives respectively in Fig. 6.5 and in the top panel of Fig. 6.13. Notable changes are observed in the σ_8 and M_ν panels in which the matter spectrum $P_{\ell=0}^{\rm s,mm}(k)$ is showing a broader range of amplitudes suggesting more changes with these parameters in redshift than in real space and hence different constraints obtained for it. The monopoles of the environments, on their side, are still showing different shapes from each other, highlighting different variations with cosmological parameters. These various effects illustrate the ability of statistics built from the combination of the spectra from different environments to break degeneracies between cosmological parameters and subsequently improve the constraints over the sole information from the matter power spectrum. The second ingredient of the Fisher analysis, the correlation coefficients between wavenumbers of the matter and environmental monopoles C_{k_1,k_2} defined in Eq. (6.16), are respectively shown in the bottom-left block-submatrices of the left and right panels of Fig. 6.14. Compared to the real-space matrices obtained in Fig. 6.6, the redshift-space case show much less high-amplitudes off-diagonal terms. Taking the same example as in real-space, $C_{k_1,k_2} \sim 0.20$ for $k_1, k_2 \sim 0.30$ h/Mpc. This phenomenon is also visible for the environment-dependent spectra, and more particularly in the case of nodes where the correlation between modes is also reduced.

All the discussed effects on the matter power spectrum $P_{\ell=0}^{\rm s,mm}$ in redshift space are leading to tighter constraints on cosmological parameters shown in the first row of Table 6.4. Almost all parameters are getting considerable gain over the monopole of the real-space analysis, except σ_{σ_8} which remains roughly at the same value. As we have been arguing in the previous

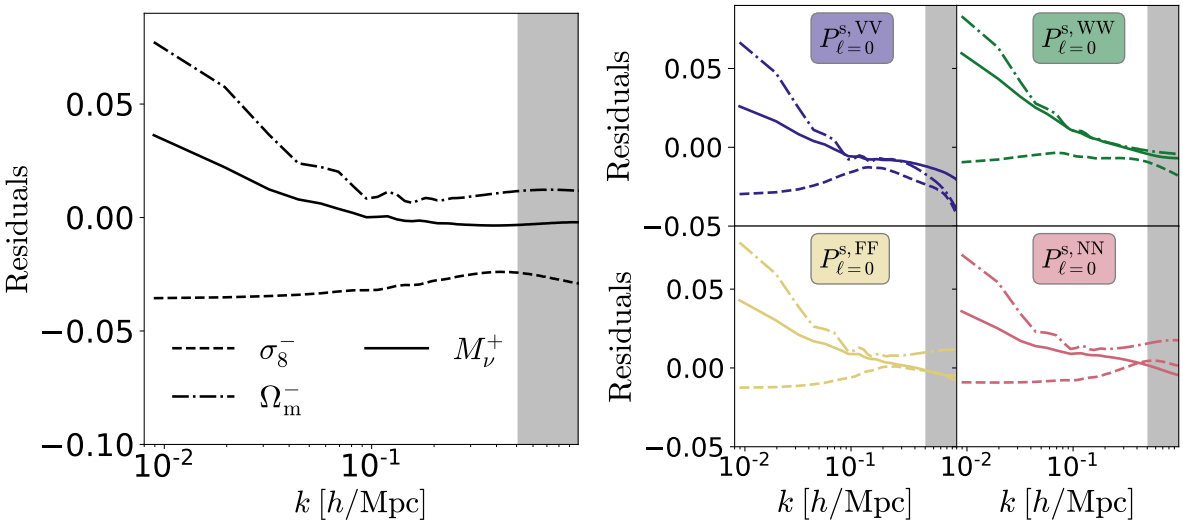

Fig. 6.12. A closer look at the impact of σ_8, Ω_m and M_ν on the matter power spectrum and those derived from cosmic environments. Are shown the residuals of matter (left panel) and spectra in cosmic environments (right panel) $P_{\ell=0}^{\mathrm{s},\alpha\alpha}(k)^{\theta_i}/P_{\ell=0}^{\mathrm{s},\alpha\alpha}(k)^{\mathrm{fid}} - 1$ with $\theta_i = \sigma_8^-$ (plain line), M_ν^+ (dashed line) or Ω_m^- (dotted dashed line).

Table 6.4. Marginalised 1-σ constraints obtained from the analysis of power spectra monopoles and quadrupoles computed in the different environments for all cosmological parameters.

Statistics	$\sigma_{\Omega_\mathrm{m}}$	$\sigma_{\Omega_\mathrm{b}}$	σ_h	σ_{n_s}	σ_{σ_8}	σ_{M_ν}
$P_{\ell=0}^{\mathrm{s,mm}}$	0.0079	0.0144	0.1485	0.0794	0.0123	0.3876
$P_{\ell=0}^{\mathrm{s,VV}}$	0.0307 (0.3)	0.0192 (0.7)	0.1998 (0.7)	0.1128 (0.7)	0.0158 (0.8)	0.3289 (1.2)
$P_{\ell=0}^{\mathrm{s,WW}}$	0.0260 (0.3)	0.0224 (0.6)	0.2164 (0.7)	0.1786 (0.4)	0.0332 (0.4)	0.9826 (0.4)
$P_{\ell=0}^{\mathrm{s,FF}}$	0.0187 (0.4)	0.0182 (0.8)	0.2175 (0.7)	0.1740 (0.5)	0.0291 (0.4)	0.3401 (1.1)
$P_{\ell=0}^{\mathrm{s,NN}}$	0.0340 (0.2)	0.0272 (0.5)	0.3596 (0.4)	0.3944 (0.2)	0.0963 (0.1)	1.1548 (0.3)
$P_{\ell=0}^{\mathrm{s,comb}}$	0.0051 (**1.6**)	0.0104 (**1.4**)	0.0839 (**1.8**)	0.0348 (**2.3**)	0.0033 (**3.8**)	0.0709 (**5.5**)
$P_{\ell=\{0,2\}}^{\mathrm{s,mm}}$	0.0048	0.0133	0.1391	0.0716	0.0020	0.0838
$P_{\ell=\{0,2\}}^{\mathrm{s,VV}}$	0.021 (0.2)	0.0143 (0.9)	0.1316 (1.1)	0.0801 (0.9)	0.0093 (0.2)	0.0984 (0.9)
$P_{\ell=\{0,2\}}^{\mathrm{s,WW}}$	0.0130 (0.4)	0.0213 (0.6)	0.1921 (0.7)	0.0766 (0.9)	0.0146 (0.1)	0.1449 (0.6)
$P_{\ell=\{0,2\}}^{\mathrm{s,FF}}$	0.0093 (0.5)	0.0158 (0.8)	0.1844 (0.8)	0.1283 (0.6)	0.0084 (0.2)	0.1825 (0.5)
$P_{\ell=\{0,2\}}^{\mathrm{s,NN}}$	0.0126 (0.4)	0.0210 (0.6)	0.2378 (0.6)	0.1553 (0.5)	0.0140 (0.1)	0.5322 (0.2)
$P_{\ell=\{0,2\}}^{\mathrm{s,comb}}$	0.0034 (**1.4**)	0.0097 (**1.4**)	0.0761 (**1.8**)	0.0306 (**2.3**)	0.0020 (**1.0**)	0.0349 (**2.4**)

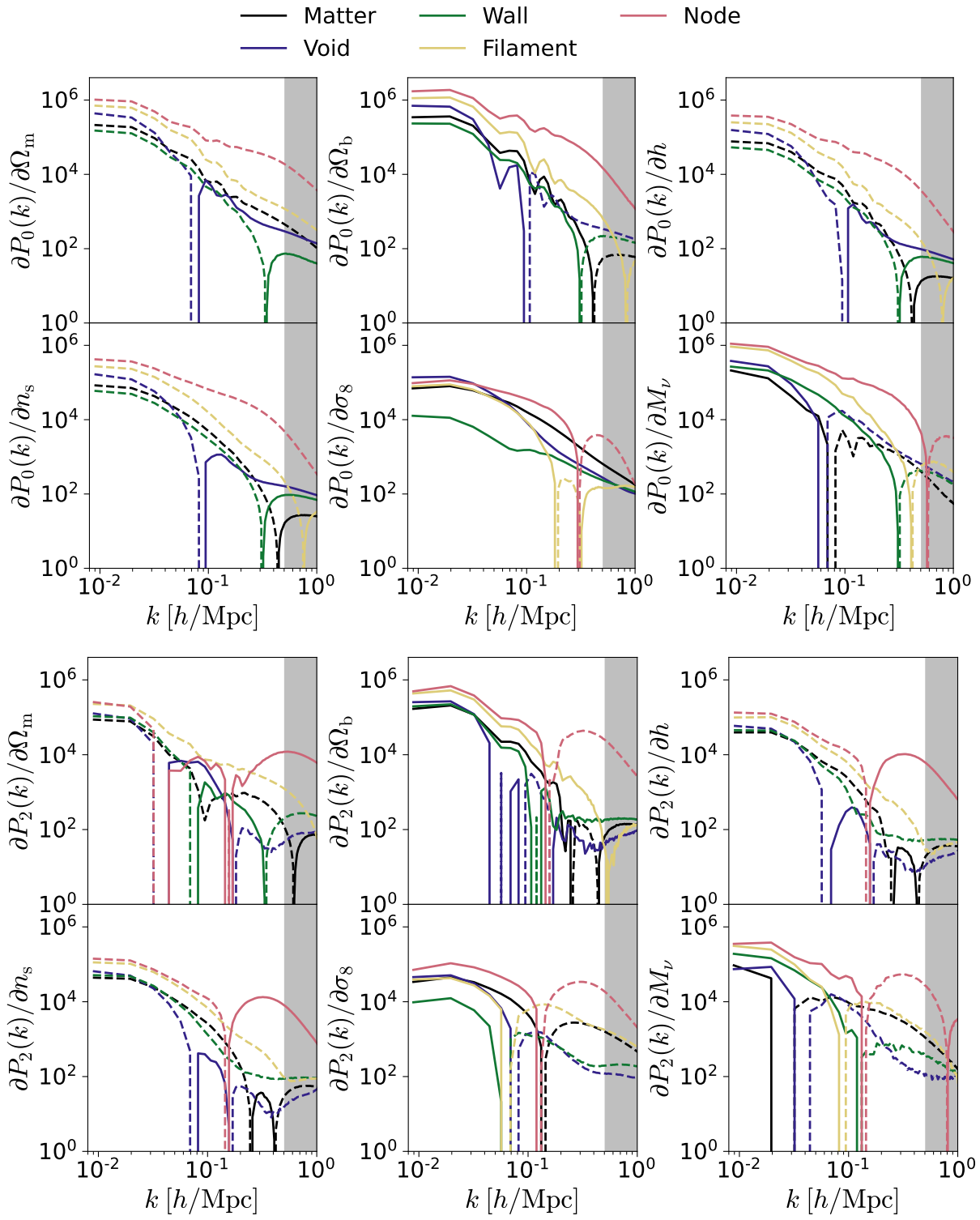

Fig. 6.13. Partial derivatives $\partial P^{\alpha\alpha}(k)/\partial\theta_i$ for the different environmental auto-spectra and for each studied parameters $\{\Omega_{\rm m}, \Omega_{\rm b}, h, n_{\rm s}, \sigma_8, M_\nu\}$. Dashed (resp. plain) lines correspond to negative (resp. positive) values of the derivative. The grey area depicts the range of $k > k_{\rm max}$ excluded from the analysis.

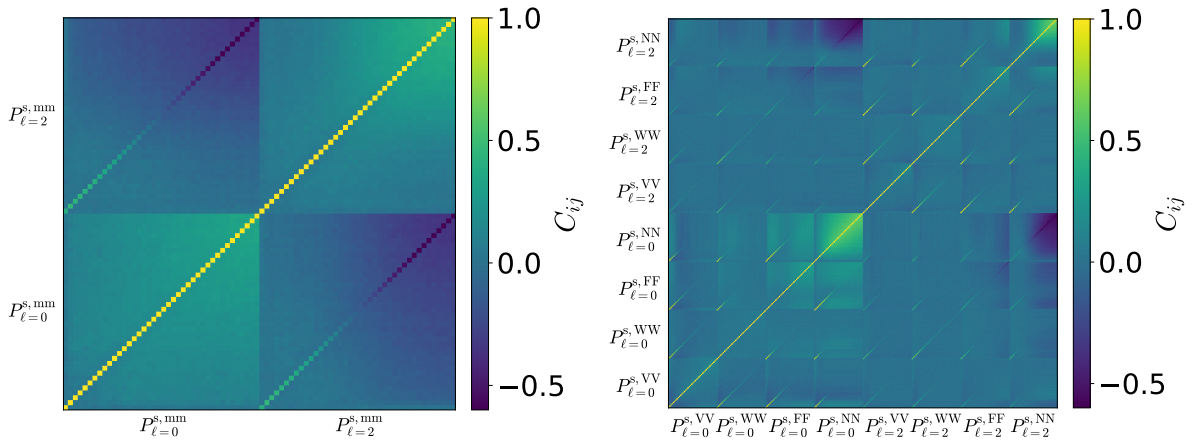

Fig. 6.14. *(left)* Correlation coefficients C_{ij} for matter-matter power spectrum monopole and quadrupole. *(right)* Same for $P_\ell^{s,\alpha\alpha}(k)$ coefficients extracted from the several environments. Each sub-matrix goes from $k = 0.1$ h/Mpc to $k = k_{\rm max} = 0.5$ h/Mpc.

paragraph, the various sensitivity of cosmic environments to some cosmological parameters are still leading to a sizeable decrease of the marginalised errors σ_{θ_i} when all of them are combined together. Even though they perform individually worse than the matter two-point analysis, their combination is allowing, by breaking some degeneracies, to bring improvement factors of 1.4 to 3.8 on cosmological parameters and 5.5 on M_ν. Adding the next non-zero multipole order in the matter analysis, namely the quadrupole $P^{\rm s,mm}_{\ell=2}$, is not changing significantly the constraints on $\Omega_{\rm b}, h$ and $n_{\rm s}$ but improves drastically those on $\Omega_{\rm m}, \sigma_8$ and M_ν respectively with factors 1.6, 6.2 and 4.6, as shown in the bottom-half part of Table 6.4. These tighter constraints obtained by the matter power spectrum in redshift space are explained by the breaking of degeneracies in the M_ν–σ_8 plane allowed by the quadrupole. Note however that the absolute value of these constraints are optimistic since our analysis rely on dark matter particles. Using tracers like halos or galaxies would considerably reduce the constraining power of the two-point statistics [see for instance Table 2 of Hahn et al., 2020]. This improvement is highlighted by much more features observed in the derivatives of the quadrupole with parameters like σ_8 and M_ν of the bottom part of Fig. 6.13. In this case, the 320 Fourier coefficients resulting from the combination of monopoles and quadrupoles in cosmic environments lead to improvement factors of 1.0 to 2.3 for cosmological parameters and 2.4 for the summed neutrino mass. In particular, we denote no further improvement of the error on σ_8, already well-constrained by the $P^{\rm s,mm}_{\ell=0} + P^{\rm s,mm}_{\ell=2}$ statistics. This is also illustrated in the $1\sigma_{\theta_i}$ marginalised confidence ellipses presented in Fig. 6.15 where the ellipses from environmental combination are superimposed on the matter ones for almost all paired parameters with σ_8. We still observe, however, some different orientations for ellipses in several planes like the M_ν–σ_8 or $\Omega_{\rm m}$–$n_{\rm s}$ which are at the origin of the non-negligible improvements for the constraints on M_ν and $n_{\rm s}$ over the matter monopole and quadrupole analysis. Exactly as for the real-space case, these gains are observed for all considered $k_{\rm max} \in [0.11, 0.5]$ h/Mpc scales, as shown in Fig. 6.16 and especially marked by a larger improvement factor when considering only the largest scales $k_{\rm max} < 0.20$ h/Mpc. It is also interesting to note the saturation of the void environment at roughly 0.20 h/Mpc which then sees its constraining power improving again at scales of 0.35 h/Mpc. Overall, while the matter power spectrum shows a plateau for all parameters, the combination of environments still suggest further improvement when going to smaller scales for some parameters like $\Omega_{\rm m}, \Omega_{\rm b}$, or h. Finally, in Fig. 6.17 are shown the

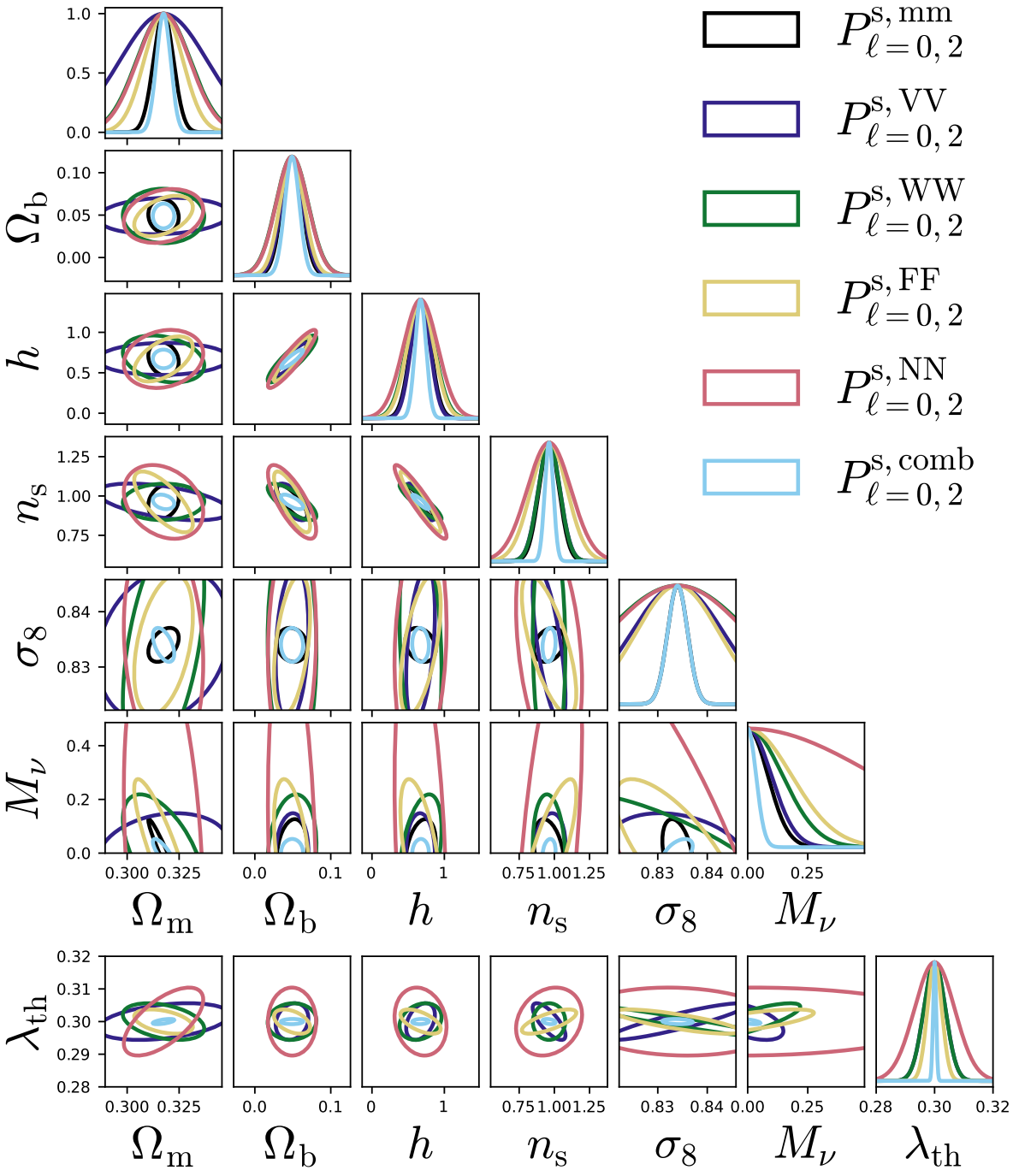

Fig. 6.15. 1σ confidence ellipses for all the pairs of cosmological $(\Omega_{\mathrm{m}}, \Omega_{\mathrm{b}}, h, n_{\mathrm{s}}, \sigma_8, M_\nu)$ and nuisance (λ_{th}) parameters obtained from the monopoles and quadrupoles from the matter power spectrum, or the one from the different environments and their combination in redshift space.

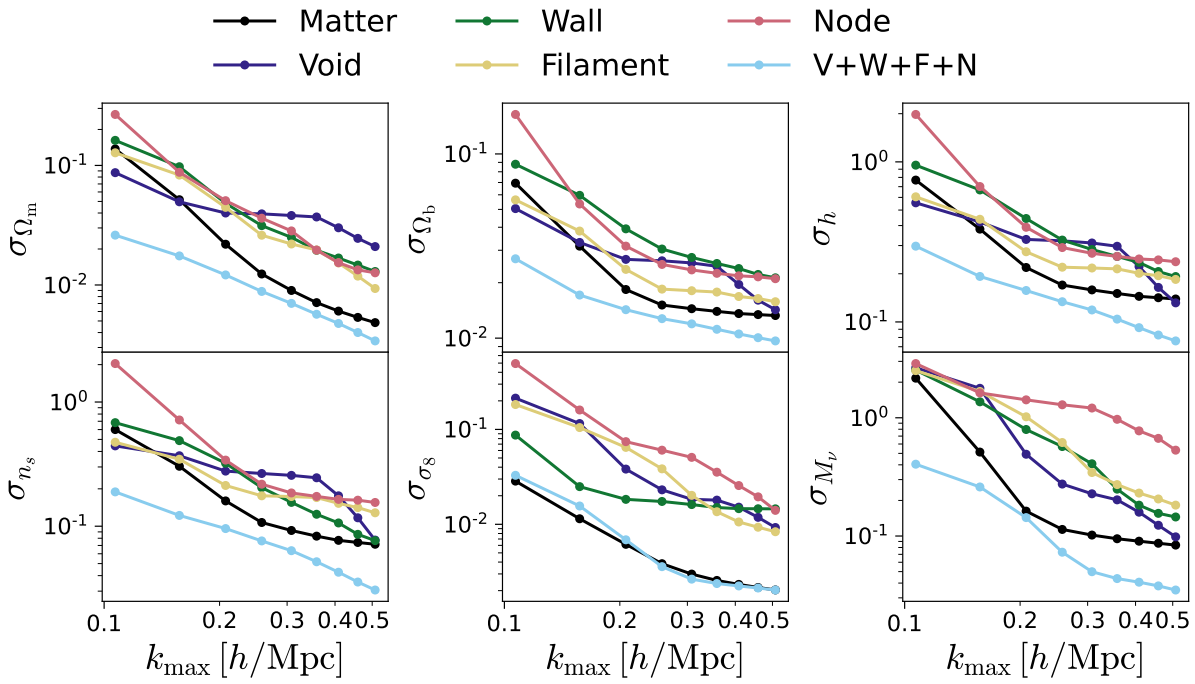

Fig. 6.16. Evolution of the marginalised constraint σ_θ put on cosmological parameters $\{\Omega_{\rm m}, \Omega_{\rm b}, h, n_{\rm s}, \sigma_8\}$ and the sum of neutrino mass M_ν in redshift space with the maximum scale used for the Fisher analysis, namely $k_{\rm max}$.

evolution of the SNRs with the maximum scale $k_{\rm max}$ for the monopole-only analysis in dotted line and the one adding the quadrupole in solid line. The first remarkable feature is the saturation of information for $P^{\rm s,mm}_{\ell=0}$ which slowly increases from 0.20 to 0.50 h/Mpc while increases much faster when adding the quadrupole. This is coherent with the observed constraints by the Fisher analysis. On the other hand, the environmental SNR are not changing much when adding the quadrupole, except for nodes which show a regain at small scales. Quantitatively, the SNR of the combined environmental clustering statistics is twice the one of the matter at all scales when using both monopoles and quadrupoles. Finally, adding the information of the next non-zero multipole $\ell = 4$ in the analysis does not improve further the constraints at the considered scale of 0.5 h/Mpc, both for the matter power spectrum and the ones obtained in the cosmic environments.

6.4.4 Stability and convergence analysis

In this Fisher forecast, we resort to numerical computations of the precision matrices defined in Eq. (6.13) but also of the derivatives from Eq. (6.14) and (6.15). To avoid biased results induced by a non-convergence of these quantities, it is essential to check the stability of the derived constraints under reduction of both $N_{\rm fid}$, the number of simulations used to compute the covariances and $N_{\rm deriv}$, the number of simulations for the derivatives. We focus here on the convergence of the constraints σ_{θ_i} derived from the combination of power spectra in all environments yielding the maximum length among all the studied statistical summaries with $n = 160$ in real space and $n = 320$ when combining the monopole and quadrupole in redshift space. Note that the convergence of the matter power spectrum is already studied in Villaescusa-Navarro et al. [2020]. Figures 6.18 and 6.19 show how the marginalised constraints σ_{θ_i} individually behave respectively in real and redshift space when varying $N_{\rm fid}$ and $N_{\rm deriv}$.

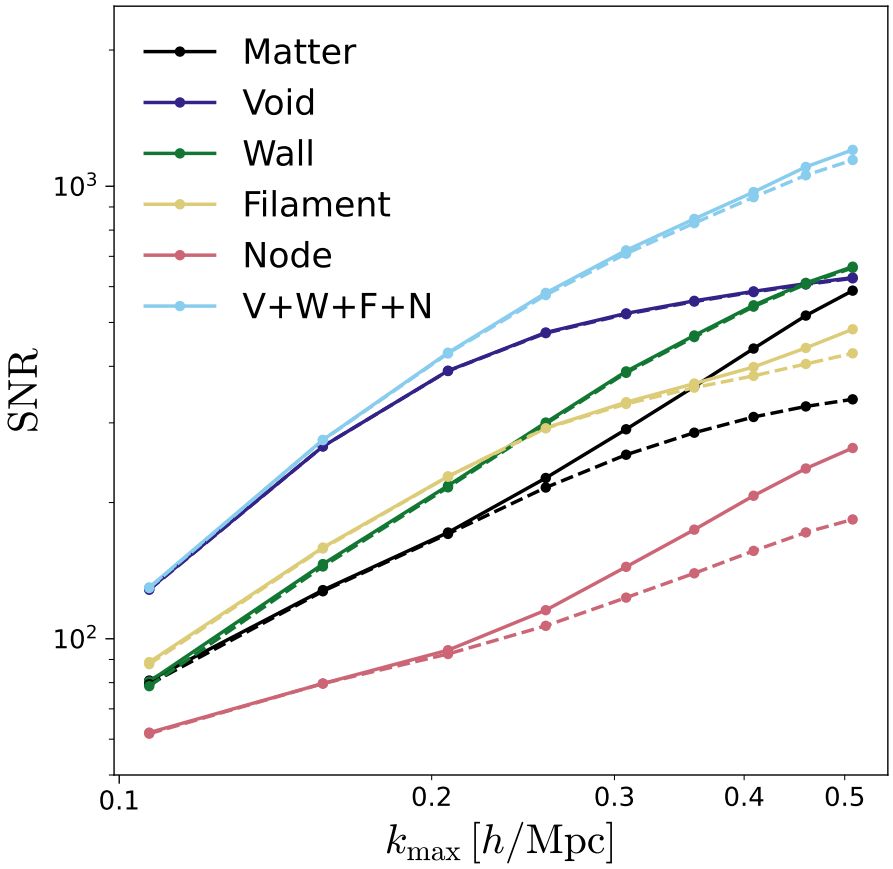

Fig. 6.17. Evolution of the SNR with the maximum scale for the power spectra used in the Fisher analysis in redshift space, namely $k_{\rm max}$, for the monopoles $P^{\rm s}_{\ell=0}$ only (dashed lines) and monopoles plus quadrupoles $P^{\rm s}_{\ell=0} + P^{\rm s}_{\ell=2}$ (plain lines).

We see that, for all parameters, convergence at a $\pm 1\%$ level is obtained when $N_{\rm fid} \sim 5000$ and $N_{\rm deriv} \sim 300$ in real space while these values are respectively shifted to 5250 and 350 in redshift space. These results show the good convergence properties of the analysed statistics, excluding any bias induced by numerical instabilities in the computation of Fisher constraints.

6.5 Conclusion and perspectives

In this work, we have carried out the first quantitative analysis of the cosmological information content of all cosmic web environments based on the two-point clustering statistics in Fourier space. The auto-spectra derived from environments were computed from individual density fields identified with the T-web formalism which relies on the local curvature of the gravitational potential to depict nodes, filaments, walls and voids. Using the large suite of Quijote simulations, we were able to carry out a Fisher analysis by computing the partial derivatives and the covariance matrix of the extracted statistics in a statistically robust manner in the non-linear regime with $k_{\rm max} \approx 0.5\ h$/Mpc. We compared the constraints put by the environments upon the matter power spectrum analysis on the set of five cosmological parameters $\{\Omega_{\rm m}, \Omega_{\rm b}, h, n_{\rm s}, \sigma_8\}$ and the summed neutrino mass M_ν.

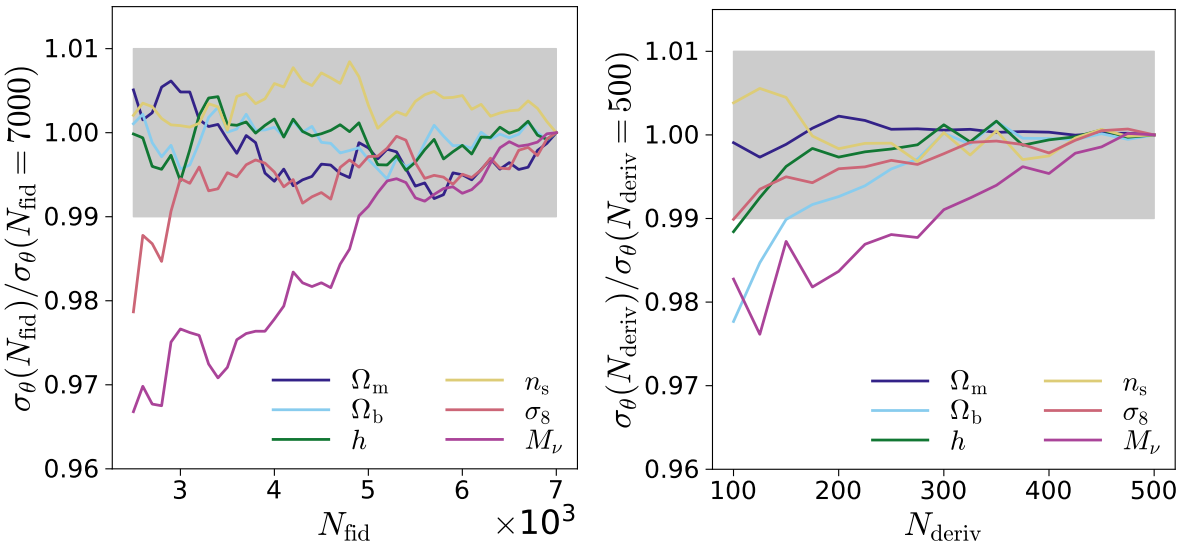

Fig. 6.18. Convergence analysis of the numerical precision matrix as a function of $N_{\rm fid}$ with fixed $N_{\rm deriv} = 500$ (left panel) and derivatives as a function of $N_{\rm deriv}$ with fixed $N_{\rm fid} = 7000$ (right panel) for the combined spectra statistics $P^{\rm comb}$ in real space. The grey area shows the $\pm 1\%$ variations in the marginalised constraints.

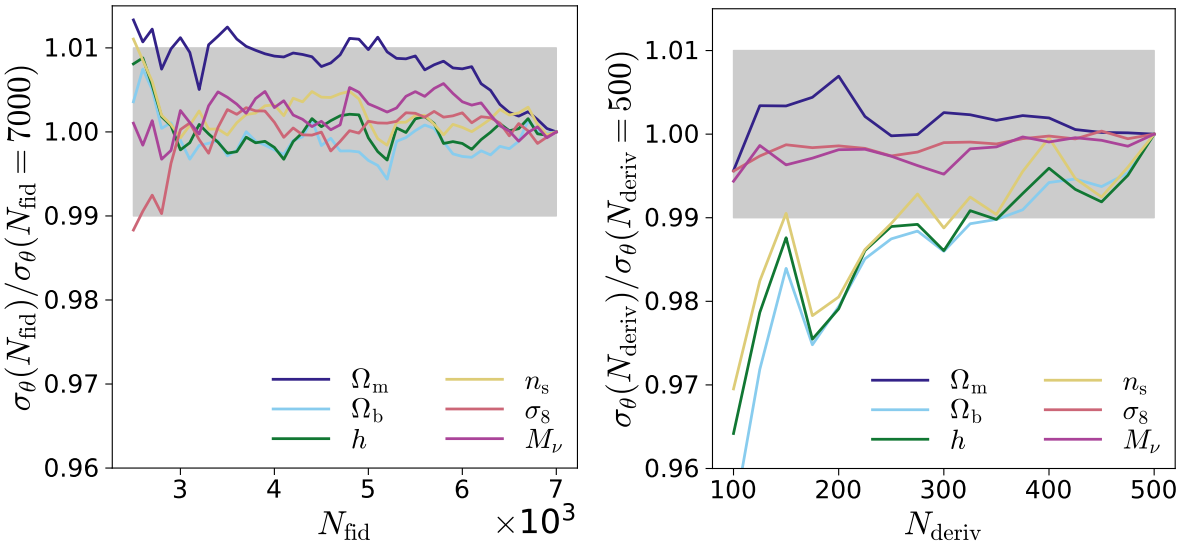

Fig. 6.19. Convergence analysis of the numerical precision matrix as a function of $N_{\rm fid}$ with fixed $N_{\rm deriv} = 500$ (left panel) and derivatives as a function of $N_{\rm deriv}$ with fixed $N_{\rm fid} = 7000$ (right panel) for the combined spectra statistics $P^{\rm s,comb}_{\ell=\{0,2\}}$ in redshift space. The grey area shows the $\pm 1\%$ variations in the marginalised constraints.

In particular, we find that

- The power spectra computed in the cosmic web environments show different shape dependencies when varying some cosmological parameters such as M_ν, Ω_m and σ_8 in real and redshift spaces that are not simple translations of the matter power spectrum. These variations originate from the intrinsic differences in densities and hence evolution histories of each environment, where the observed structures at $z = 0$ are imprinted differently depending on the cosmology.
- From a quantitative point of view, the combination of power spectra in the environments at linear and non-linear scales ($k_\mathrm{max} = 0.5\ h$/Mpc) leads to the possibility of breaking some key degeneracies between parameters of the model which consequently allows to tighten the constraints with improvement factors of $\{8.0, 4.5, 6.7, 17.1, 6.5, 15\}$ respectively on parameters $\{\Omega_\mathrm{m}, \Omega_\mathrm{b}, h, n_\mathrm{s}, \sigma_8, M_\nu\}$ over the power spectrum in real space. Redshift-space constraints, although already narrowed down when using the joint analysis of the matter monopoles and quadrupoles, are getting even tighter when combining the environmental auto-spectra allowing factors of improvements up to 2.3 on n_s and 2.4 on the sum of neutrino mass.
- The constraints obtained from the combination of two-point statistics in Fourier space in the cosmic environments are superior for the whole range of maximum scales analysed $k_\mathrm{max} \in [0.11, 0.5]\ h$/Mpc and this statistics yields a 8 (resp. 2) times higher signal-to-noise ratio than the matter power spectrum in real space (resp. redshift space).

The matter component used in this analysis, the cold dark matter, is however not a direct observable. Observational analyses rely on statistics derived from biased and visible tracers of the matter, such as galaxies. The first issue of handling such biased tracers is that they may trace differently the cosmic environments. For instance, using the halos of the Quijote simulations with the same mass threshold as used in Hahn et al. [2020], namely $M > 3.2 \times 10^{13} M_\odot/h$, would considerably alter the detection of low-density environments like voids or walls that we do not expect to host halos of such mass [Cautun et al., 2014]. It is then necessary to infer the underlying dark matter clustering taking this bias effect into account (see Sect. 2.2.2). Bias usually comes in the analysis as an additional nuisance parameter that needs to be inferred from the data itself, thus contributing to increase of the size of the parameter space and may also lead to additional degeneracies among parameters reducing consequently the constraining power of the statistics. In our case, assuming a linear biasing scheme, we would need to take an individual bias term into account for each environment $\{b_\mathrm{V}, b_\mathrm{W}, b_\mathrm{F}, b_\mathrm{N}\}$.

It is also well-known that adding constraints on the cosmological parameters from the independant cosmic microwave background measurements as a prior in such analyses of the matter power spectrum can considerably reduce the obtained errors. In this work, we however did not consider additional priors given that the aim was to carry a theoretical quantification of the information content of cosmic environments only. Adding such priors could however be of interest when targeting real observations, together with matter tracers, which will be the purpose of future works.

The estimates of the matter power spectrum weighted by the local density proposed by White [2016] yields a non-linear transform of the overdensity field that we exposed in Eq.

(6.18). Such a weighted version of the power spectrum has been shown to contain more information than the standard matter power spectrum in real-space by Massara et al. [2021]. It is however well-established that cosmic environments are complex objects and cannot be uniquely described in terms of density thresholds [see e.g. Cautun et al., 2014; Libeskind et al., 2017]. In this context, it would be interesting to link the presently exposed analysis of the environments with a mark depending on both the density and the level of local tidal anisotropy. This would have the vast advantage of relieving the analysis from detecting the environments with a pre-fixed definition.

In conclusion, we have shown that there is significantly more information contained in the density field when analysing independently the cosmic environments and then using their combination rather than directly relying on the matter density field summarised by solely its power spectrum, as it is traditionally done. The sizeable improvements in the constraints on all cosmological parameters brought by our environment-dependent analysis, even in the ideal case addressed in the present study, opens up the possibility to take advantage of the spectra and cross spectra in environments for the optimal exploitation of future large galaxy surveys such as DESI [Levi et al., 2013] or Euclid [Laureijs et al., 2011].

Conclusion

"On n'aura pas de poste permanent, mais on aura bien rigolé."
V. Bonjean

The initial matter distribution, a nearly-Gaussian density field, evolved through billions of years under the effect of gravity eventually drawing the complex picture of the cosmic web that astronomers, astrophysicists and cosmologists are together trying to observe, describe and model. Throughout this thesis, we explored multiple facets of the cosmic web with, as underlying guideline, the extraction of relevant physical or cosmological information from cosmological datasets of the large-scale structures.

Naturally, such a question is tightly coupling multiple research fields among which are data science and statistics. This is hence by first taking a step back from the cosmological questions that we proposed to build representations of generic point-cloud datasets. By resorting to a statistical physics reformulation of the clustering, we were able to improve our understanding of how the structure of a dataset is driving the partitioning of datapoints into multiple groups. The identification and follow-up of successive phase transitions occurring during a simulated annealing allowed us to extract information about the number, sizes, and hierarchical embeddings of the multiple sub-structures in the data. We showed, for different regularised models, that the quantity at the origin of the transitions can explicitly be derived and exactly computed for the first transition and approximated for the following ones.

Building upon this clustering procedure, we have established a graph regularised model to represent continuously-structured datasets. By assuming that the data are living on a latent one-dimensional manifold, and modelling this latter as a graph, we built a new framework for the estimation of principal graphs. Taking advantage of probabilistic foundations, the algorithm we proposed was shown particularly efficient to unveil the pattern in complex datasets with multiple scales and high levels of noise and outliers. Beyond its mathematically appealing properties, such as a guaranteed convergence towards a local maximum during the optimisation, it is also fast, making it suitable for the description of large datasets.

Such multi-dimensional datasets in which a representation must be learnt to allow further scientific analyses can arise in various fields, ranging from biology to cosmology. In our case, we used the previously-exposed principal graph method, T-ReX, to carry out an in-depth analysis of the filamentary structure of the cosmic web as depicted by galaxies in hydrodynamical simulations. We exposed how the graph representation of the interconnected network of galaxies can be used to define individual filaments. By focusing on a set of three simulations, we established the main properties of the filaments statistics such as their length and curvature distributions with their characteristic exponential tails showing only rare long and highly-curved filaments. In this intricated cosmic network, a special focus was put on filaments themselves but also in the way they interact with nodes of the cosmic web. We showed that these anchors can be readily expressed in the T-ReX formalism as dense bifurcations, leading to the identification of the environments of galaxies in agreement with the physical

definition of locally isotropic regions in the matter density field. The connectivity of nodes, quantifying their local embedding in the cosmic web, was also shown to be particularly sensitive to the assembly history of clusters. In particular, we provided evidences for a scenario in which early-formed and unrelaxed clusters are, on average, more connected to the filamentary pattern than old and relaxed clusters.

All these results emphasise the importance of identifying reliably the different environments to make the best use of cosmological observables in order to understand the formation of structures at all scales, from galaxies to clusters and how the cosmic web is shaping these elements of the Universe. By undertaking a thorough quantitative analysis of the two-point information in the cosmic web environments, we showed that the combined use of nodes, filaments, walls and voids efficiently breaks degeneracies between cosmological parameters to which these environments have different sensitivities. We thus tightened the constraints on cosmological parameters, over the use of the full matter power spectrum, by up to one order magnitude. We thus pictured the cosmic environments as a gold mine of information about the Universe itself, able to bring tremendous knowledge about the underlying cosmological model.

This thesis was conducted in the perspective of optimising the modelling, the characterisation and the utilisation of large-scale structure datasets for cosmological analyses. In this context, more and more importance is particularly given to topological and geometrical representations of the cosmic web enabling new insights in the study of the structure formation and evolution at multiple scales. This growing interest is obviously fuelled by the considerable advances from the observational, theoretical and statistical communities. On the observational side, the next generation of large-scale surveys will map the large-scale matter distribution at all wavelengths from the optical to the radio with Euclid, the Vera Rubin Observatory, DESI and SKA. These new data, their cross-correlations and combinations, will not only allow us to constrain the cosmological parameters of the standard cosmological model with accuracies orders of magnitudes better than what is currently achieved but they will also allow the testing of alternative theoretical models proposed to explain the Universe formation and evolution. It is paramount that the scientific exploitation of these surveys is accompanied with an extensive effort aiming at devising optimised methods to extract meaningful information and hence address the cosmological questions.

The work presented in the thesis mainly targeted the topological representation of the cosmic web as a graph structure, focusing of the global characteristics of the matter distribution and at present time. Graphs built from the spatial proximity of galaxies are a promising way to build new summary statistics based on quantities derived from graph theory such as the degree (the number of edges connecting a given galaxy), the assortativity (the correlation between linked galaxies) or the shortest paths (set of shortest geodesic paths on the graph linking two galaxies) that have already been shown to encode cosmological information [Hong & Dey, 2015; Hong et al., 2016; Naidoo et al., 2020]. Studying the correlations between the physical properties of graph nodes (galaxies) and graph properties (degree, assortativity, length, etc.) also offer a unique opportunity to link the topology of the cosmic web to the physics of tracers.

Another avenue of research interesting to follow is the study of both the evolution and dynamics of the large-scale filamentary pattern as the temporal morphism of the extracted graph structures. Indeed, whilst the proposed work, as well as the current literature gives importance to the pattern at present time, or studies it independently at several time snapshots, only few works address the time evolution as a set of successive events. This could in particular allow the interpretation of the topological modifications of the graph structure, i.e. the local variations in the number of branches and the related size of cycles in the topology, as thermo-

dynamic events of the cosmic pattern. By building a model bringing key information of the time evolution of the filamentary pattern as viewed by the successive merging of primordial filaments, it would provide a perfect playground for the testing of hypotheses, such as the cosmic web detachment and other environmental quenching effects argued by several works [see e.g. Aragon Calvo et al., 2019].

The complexity of datasets encountered in cosmology, involving a wide range of scales, linear and non-linear physics, makes it an adapted field to apply and develop machine learning methods, which were shown particularly efficient in solving problems intractable so far based on such large and complex data in reasonable times. The outstanding results obtained by the first applications of deep learning methods in astrophysics are however often limited by the difficulties in interpreting the obtained models. It is only with the joint effort of various communities, notably from computer science and statistical physics that these problems could be readily tackled. In particular, the structure of the data and how this latter is embedded in space (symmetries, lower-dimensional manifolds, etc.) is expected to play a major role in the learning dynamics [see e.g. Goldt et al., 2020]. The work presented in this thesis explored this aspect in the simple context of clustering. An exciting perspective would be to develop more fundamental relations between the data structure and the features that are learnt in the context of simple neural networks model to better understand what makes them particularly successful for certain tasks.

On the application side of the machine learning, many recent works investigate the possibility to recover the early and/or late time large-scale dark matter distribution from the observation of sparse tracers like halos, galaxies or clusters at redshift $z = 0$ [see for instance Jasche & Wandelt, 2013; Leclercq et al., 2015; Schmittfull et al., 2017]. This mapping problem is made complex by the non-linear relation between the non-Gaussian density field and its sparse discrete representation by the distribution of galaxies. For the purpose of inferring non-linear relations between two or three dimensional fields, machine learning have already been shown particularly powerful. The recent availability of large amount of simulations targetting the specific needs to perform statistical analyses and machine learning applications like Quijote [Villaescusa-Navarro et al., 2020] makes it now possible to explore deep learning methods for such cosmological investigations. Coupled to the previous point on their interpretability, such models, if well-understood, could help improving our knowledge on the interplay between dark and baryonic matter but more so on the evolution of the Universe.

Bibliography

[Abbott et al. 2016] Abbott, T., Abdalla, F.˜B., Aleksi, J., et al.: The Dark Energy Survey: more than dark energy - an overview. *Mon. Not. R. Astron. Soc.* 460 (2016), Nr. August, S. 1270–1299. *http://dx.doi.org/10.1093/mnras/stw641* 5, 18, 23, 117

[Abbott et al. 2019] Abbott, T. M. C., Allam, S., Andersen, P., et al.: First Cosmology Results using Type Ia Supernovae from the Dark Energy Survey: Constraints on Cosmological Parameters. *Astrophys. J.* 872 (2019), Nr. 2, S. L30. *http://dx.doi.org/10.3847/2041-8213/ab04fa* 19

[Abell et al. 1989] Abell, George O., Corwin, Harold G., Jr., Olowin, Ronald P.: A Catalog of Rich Clusters of Galaxies. *Astrophys. J.* 70 (1989). *http://dx.doi.org/10.1086/191333* 94

[Abolfathi et al. 2018] Abolfathi, Bela, Aguado, D. S., Aguilar, Gabriela, et al.: The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. *Astrophys. J. Suppl. Ser.* 235 (2018), Nr. 2, S. 42. *http://dx.doi.org/10.3847/1538-4365/aa9e8a* 19

[Ade et al. 2019] Ade, Peter, Aguirre, James, Ahmed, Zeeshan, et al.: The Simons Observatory: Science goals and forecasts. *J. Cosmol. Astropart. Phys.* 2019 (2019), Nr. 2. *http://dx.doi.org/10.1088/1475-7516/2019/02/056* 5

[Adler & Taylor 2007] Adler, R. J., Taylor, J. E.: *Random Fields and Geometry*. Springer-Verlag New York, 2007. – 454 S. *http://dx.doi.org/10.1007/978-0-387-48116-6*. *http://dx.doi.org/10.1007/978-0-387-48116-6* 17

[Advani et al. 2020] Advani, Madhu S., Saxe, Andrew M., Sompolinsky, Haim: High-dimensional dynamics of generalization error in neural networks. *Neural Networks* 132 (2020), S. 428–446. *http://dx.doi.org/10.1016/j.neunet.2020.08.022* 41

[Aghanim et al. 2015] Aghanim, N., Hurier, G., Diego, J. M., et al.: The Good, the Bad, and the Ugly: Statistical quality assessment of SZ detections. *Astron. Astrophys.* 580 (2015), 1–15. *http://dx.doi.org/10.1051/0004-6361/201424963* 44

[Ahmed et al. 2015] Ahmed, Mahmuda, Karagiorgou, Sophia, Pfoser, Dieter, Wenk, Carola: A comparison and evaluation of map construction algorithms using vehicle tracking data. *Geoinformatica* 19 (2015), 601–632. *https://doi.org/10.1007/s10707-014-0222-6* 66

[Akaho & Kappen 2000] Akaho, S., Kappen, H. J.: Nonmonotonic Generalization Bias of Gaussian Mixture Models. *Neural Comput.* 12 (2000), Nr. 6, 1411–1427. *http://dx.doi.org/10.1162/089976600300015439* 49

[Akaike 1974] Akaike, H.: A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* 19 (1974), Nr. 6, S. 716–723. *http://dx.doi.org/10.1109/TAC.1974.1100705* 55

[Albergante et al. 2020] Albergante, Luca, Mirkes, Evgeny M., Chen, Huidong, et al.: Robust and scalable learning of complex dataset topologies via elpigraph. *Entropy* 22 (2020). *https://www.mdpi.com/1099-4300/22/3/296* 67, 73, 77, 78, 81, 85

[Alcock & Paczynski 1979] Alcock, C., Paczynski, B.: An evolution free test for non-zero cosmological constant. *Nature* 281 (1979), 358. *http://dx.doi.org/10.1038/281358a0* 25

[Allen et al. 2011] Allen, Steven W., Evrard, August E., Mantz, Adam B.: Cosmological parameters from observations of galaxy clusters. *Annu. Rev. Astron. Astrophys.* 49 (2011), S. 409–470. *http://dx.doi.org/10.1146/annurev-astro-081710-102514* 120

[Allys et al. 2020] Allys, E., Marchand, T., Cardoso, J. F., et al.: New interpretable statistics for large-scale structure analysis and generation. *Phys. Rev. D* 102 (2020), Nr. 10. *http://dx.doi.org/10.1103/PhysRevD.102.103506* 6, 29, 40

[Alpaslan et al. 2016] Alpaslan, Mehmet, Grootes, Meiert W., Marcum, Pamela M., et al.: Galaxy And Mass Assembly (GAMA): Stellar mass growth of spiral galaxies in the cosmic web. *Mon. Not. R. Astron. Soc.* 457 (2016), S. 2287–2300 94, 95

[Alpaslan et al. 2014a] Alpaslan, Mehmet, Robotham, Aaron S., Driver, Simon, et al.: Galaxy and mass assembly (GAMA): The large-scale structure of galaxies and comparison to mock universes. *Mon. Not. R. Astron. Soc.* 438 (2014), Nr. 1, S. 177–194. *http://dx.doi.org/10.1093/mnras/stt2136* 32, 34, 36, 95

[Alpaslan et al. 2014b] Alpaslan, Mehmet, Robotham, Aaron S., Obreschkow, Danail, et al.: Galaxy and Mass Assembly (GAMA): Fine filaments of galaxies detected within voids. *Mon. Not. R. Astron. Soc. Lett.* 440 (2014), Nr. 1, S. 1–6. *http://dx.doi.org/10.1093/mnrasl/slu019* 32, 34, 71, 95

[Alpert et al. 1999] Alpert, Charles J., Kahng, Andrew B., Yao, So Z.: Spectral partitioning with multiple eigenvectors. *Discret. Appl. Math.* 90 (1999), Nr. 1-3, S. 3–26. *http://dx.doi.org/10.1016/S0166-218X(98)00083-3* 69

[Amit et al. 1985] Amit, Daniel J., Gutfreund, Hanoch, Sompolinsky, H.: Spin-glass models of neural networks. *Phys. Rev. A* 32 (1985), Nr. 2, 1007—-1018. *http://dx.doi.org/10.1103/PhysRevA.32.1007* 41

[Angulo et al. 2012] Angulo, R. E., Springel, V., White, S. D., et al.: Scaling relations for galaxy clusters in the millennium-XXL simulation. *Mon. Not. R. Astron. Soc.* 426 (2012), Nr. 3, 2046–2062. *http://dx.doi.org/10.1111/j.1365-2966.2012.21830.x* 23

[Aragon-Calvo 2019] Aragon-Calvo, M. A.: Classifying the large-scale structure of the universe with deep neural networks. *Mon. Not. R. Astron. Soc.* 484 (2019), Nr. 4, S. 5771–5784. *http://dx.doi.org/10.1093/mnras/stz393* 34, 81

[Aragon-Calvo et al. 2010] Aragon-Calvo, MA, Weygaert, Rien Van D., Jones, Bernard J T.: Multiscale Phenomenology of the Cosmic Web. *Mon. Not. R. Astron. Soc.* 408 (2010), Nr. November, S. 2163–2187 111, 112

[Aragon-Calvo et al. 2007] Aragon-Calvo, Miguel A., Jones, Bernard J. T., Weygaert, Rien van d., et al.: The Multiscale Morphology Filter: Identifying and Extracting Spatial Patterns in the Galaxy Distribution. *Astron. Astrophys.* 474 (2007), 315–338. *http://dx.doi.org/10.1051/0004-6361:20077880* 32, 33, 34

[Aragon Calvo et al. 2019] Aragon Calvo, Miguel A., Neyrinck, Mark C., Silk, Joseph: Galaxy Quenching from Cosmic Web Detachment. *Open J. Astrophys.* 2 (2019), Nr. 1, S. 7. *http://dx.doi.org/10.21105/astro.1607.07881* 29, 33, 149

[Aragón-Calvo et al. 2010] Aragón-Calvo, Miguel A., Platen, Erwin, Van De Weygaert, Rien, Szalay, Alexander S.: The spine of the cosmic web. *Astrophys. J.* 723 (2010), Nr. 1, S. 364–382. *http://dx.doi.org/10.1088/0004-637X/723/1/364* 33, 34, 81

[Arjovsky et al. 2017] Arjovsky, Martin, Chintala, Soumith, Bottou, Léon: Wasserstein GAN. In: *Int. Conf. Mach. Learn.*, 2017 *http://arxiv.org/abs/1701.07875*, 214—-223 40

[Armijo et al. 2018] Armijo, Joaquín, Cai, Yan C., Padilla, Nelson, et al.: Testing modified gravity using a marked correlation function. *Mon. Not. R. Astron. Soc.* 478 (2018), Nr. 3, 3627–3632. *http://dx.doi.org/10.1093/MNRAS/STY1335* 121

[Aurenhammer et al. 2013] Aurenhammer, Franz, Klein, Rolf, Lee, Der-Tsai: *Voronoi Diagrams and Delaunay Triangulations*. World Scientific Publishing Co., Inc., 2013. – 348 S. *http://dx.doi.org/10.5555/2563475*. *http://dx.doi.org/10.5555/2563475* 71

[Bahcall & Fan 1998] Bahcall, N. A., Fan, X.: The Most Massive Distant Clusters: Determining Ωm and σ8. *Am. Astron. Soc.* 504 (1998), Nr. 1, 1–6. *http://dx.doi.org/10.1086/306088* 29, 120

[Bahcall et al. 1997] Bahcall, Neta A., Fan, Xiaohui, Cen, Renyue: Constraining Ω with Cluster Evolution. *Astrophys. J.* 485 (1997), Nr. 2, S. L53–L56. *http://dx.doi.org/10.1086/310814* 29, 120

[Bahri et al. 2020] Bahri, Yasaman, Kadmon, Jonathan, Pennington, Jeffrey, et al.: Statistical Mechanics of Deep Learning. *Annu. Rev. Condens. Matter Phys.* 11 (2020), S. 501–528. *http://dx.doi.org/10.1146/annurev-conmatphys-031119-050745* 62

[Bair 2013] Bair, Eric: Semi-supervised clustering methods. *Wiley Interdiscip. Rev. Comput. Stat.* 5 (2013), Nr. 5, 349–361. *http://dx.doi.org/10.1002/wics.1270* 44

[Baldry et al. 2006] Baldry, I. K., Balogh, M. L., Bower, R. G., et al.: Galaxy bimodality versus stellar mass and environment. *Mon. Not. R. Astron. Soc.* 373 (2006), Nr. 2, S. 469–483. *http://dx.doi.org/10.1111/j.1365-2966.2006.11081.x* 29, 94

[Barchi et al. 2016] Barchi, P., Costa, F. da, Sautter, R., et al.: Improving galaxy morphology with machine learning. *J. Comput. Interdiscip. Sci.* 7 (2016), Nr. 3, S. 1–7. *http://dx.doi.org/10.6062/jcis.2016.07.03.0114* 44

[Barchi et al. 2020] Barchi, P. H., Carvalho, R. R., Rosa, R. R., et al.: Machine and Deep Learning applied to galaxy morphology - A comparative study. *Astron. Comput.* 30 (2020), 100334. *http://dx.doi.org/10.1016/j.ascom.2019.100334* 40

[Bardeen et al. 1986] Bardeen, J.M., Bond, J.R., Kaiser, N., Szalay, A.S.: The Statistics of Peaks of Gaussian Random Fields. *Astrophys. J.* 304 (1986), 15. *http://dx.doi.org/10.1086/164143* 27

[Barrow et al. 1985] Barrow, John D., Bhavsar, Suketu P., Sonoda, D. H.: Minimal spanning trees, filaments and galaxy clustering. *Mon. Not. R. Astron. Soc.* 216 (1985), S. 17–35 32, 68, 71, 95, 97

[Bayer et al. 2021] Bayer, Adrian E., Villaescusa-Navarro, Francisco, Massara, Elena, et al.: Detecting Neutrino Mass by Combining Matter Clustering, Halos, and Voids. *Astrophys. J.* 919 (2021), Nr. 1, 24. *http://dx.doi.org/10.3847/1538-4357/ac0e91* 121, 132

[Beisbart & Kerscher 2000] Beisbart, Claus, Kerscher, Martin: Luminosity- and Morphology-dependent Clustering of Galaxies. *Astrophys. J.* 545 (2000), Nr. 1, 6–25. *http://dx.doi.org/10.1086/317788* 121

[Belkin & Niyogi 2001] Belkin, Mikhail, Niyogi, Partha: Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In: *Adv. Neural Inf. Process. Syst. 14*, Press, MIT, 2001, S. 585—-591 69, 74

[Belkin & Niyogi 2003] Belkin, Mikhail, Niyogi, Partha: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15 (2003), Nr. 6, 1373–1396. *http://dx.doi.org/10.1162/089976603321780317* 66, 69, 79

[Bentley 1975] Bentley, Jon L.: Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* 18 (1975), Nr. 9, 509–517. *http://dx.doi.org/10.1145/361002.361007* 84

[Berkolaiko 2017] Berkolaiko, Gregory: An elementary introduction to quantum graphs. *Contemp. Math.* 700 (2017), 41–72. *http://dx.doi.org/10.1090/conm/700/14182* 68

[Bernardeau et al. 2002] Bernardeau, F., Colombi, S., Gaztañaga, E., Scoccimarro, R.: Large-scale structure of the Universe and cosmological pertubation theory. *Phys. Rep.* 367 (2002), Nr. 1-3, 1–248. *http://dx.doi.org/10.1016/S0370-1573(02)00135-7* 11, 14

[Bhavsar & Splinter 1996] Bhavsar, Suketu P., Splinter, Randall J.: The superiority of the minimal spanning tree in percolation analyses of cosmological data sets. *Mon. Not. R. Astron. Soc.* 282 (1996), Nr. 4, S. 1461–1466. *http://dx.doi.org/10.1093/mnras/282.4.1461* 95

[Bishop 1994] Bishop, Christopher M.: Graph mixture density networks. *Ast. Univ.* (1994), Nr. NCRG/94/004. *https://research.aston.ac.uk/en/publications/mixture-density-networks* 45

[Bishop 2006] Bishop, Christopher M.: *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006 44

[Blot et al. 2015] Blot, L., Corasaniti, P. S., Alimi, J. M., et al.: Matter power spectrum covariance matrix from the DEUS-PUR ΛCDM simulations: Mass resolution and non-Gaussian errors. *Mon. Not. R. Astron. Soc.* 446 (2015), Nr. 2, S. 1756–1764. *http://dx.doi.org/10.1093/mnras/stu2190* 130, 132

[Bond et al. 1996] Bond, J R., Kofman, Lev, Pogosyan, Dmitry: How filaments of galaxies are woven into the cosmic web. *Nature* 380 (1996), S. 603–606. *http://dx.doi.org/10.1038/380603a0* 5, 19, 22

[Bond et al. 2010] Bond, Nicholas A., Strauss, Michael A., Cen, Renyue: Crawling the cosmic network: Identifying and quantifying filamentary structure. *Mon. Not. R. Astron. Soc.* 409 (2010), Nr. 1, S. 156–168. *http://dx.doi.org/10.1111/j.1365-2966.2010.17307.x* 108, 109

[Bondy & Murty 2008] Bondy, A., Murty, U. S. R.: *Graph Theory*. 1. Springer London, 2008. – 655 S. *http://dx.doi.org/10.2307/3617646*. *http://dx.doi.org/10.2307/3617646* 68

[Bonjean et al. 2020] Bonjean, V., Aghanim, N., Douspis, M., et al.: Filament profiles from WISExSCOS galaxies as probes of the impact of environmental effects. *Astron. Astrophys.* 638 (2020), A75. *http://dx.doi.org/10.1051/0004-6361/201937313* 30, 94

[Bonjean et al. 2019] Bonjean, V., Aghanim, N., Salome, P., et al.: Star formation rates and stellar masses from machine learning. *Astron. Astrophys.* 622 (2019), 1–12. *http://dx.doi.org/10.1051/0004-6361/201833972* 40, 109

[Bonjean et al. 2018] Bonjean, V., Aghanim, N., Salomé, P., et al.: Gas and galaxies in filament between clusters of galaxies: The study of A399-A401. *Astron. Astrophys.* 609 (2018), A49. *http://dx.doi.org/10.1051/0004-6361/201731699* 36

[Bonnaire et al. 2020] Bonnaire, Tony, Aghanim, Nabila, Decelle, Aurélien, Douspis, Marian: T-ReX : a graph-based filament detection method. *Astron. Astrophys.* 637 (2020), A18. *http://dx.doi.org/10.1051/0004-6361/201936859* 32, 34, 65, 71, 81, 86, 93, 95

[Bonnaire et al. 2021a] Bonnaire, Tony, Decelle, Aurélien, Aghanim, Nabila: Cascade of Phase Transitions for Multi-Scale Clustering. *Phys. Rev. E* 103 (2021), Nr. 1, 012105. *http://dx.doi.org/10.1103/PhysRevE.103.012105* 39

[Bonnaire et al. 2021b] Bonnaire, Tony, Decelle, Aurélien, Aghanim, Nabila: Regularization of Mixture Models for Robust Principal Graph Learning. *arXiv e-prints* (2021), 1–12. *http://arxiv.org/abs/2106.09035* 32, 65, 78, 95

[Borgatti et al. 2009] Borgatti, Stephen P., Mehra, Ajay, Brass, Daniel J., Labianca, Giuseppe: Network analysis in the social sciences. *Science (80-.).* 323 (2009), Nr. 5916, S. 892–895. *http://dx.doi.org/10.1126/science.1165821* 68

[Borůvka 1926] Borůvka, Otakar: O jisíém problému minimálním. *Práce Morav. přírodovědecké společnosti* 3 (1926), 37–58. *https://dml.cz/handle/10338.dmlcz/500114?show=full* 60, 70

[Bos et al. 2014] Bos, E. G., Van De Weygaert, Rien, Kitaura, Francisco, Cautun, Marius: Bayesian cosmic web reconstruction: BARCODE for clusters. *Proc. Int. Astron. Union* 11 (2014), Nr. 308, S. 271–288. *http://dx.doi.org/10.1017/S1743921316009996* 36

[Bourlard & Kamp 1998] Bourlard, Hervé, Kamp, Yves: Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybern.* 59 (1998), S. 291—-294. *http://dx.doi.org/10.1007/BF00332918* 62

[Brouwer & Haemers 2012] Brouwer, Andries E., Haemers, Willem H.: *Spectra of Graphs.* 2012. – 250 S. *http://dx.doi.org/10.1007/978-1-4614-1939-6*. *http://dx.doi.org/10.1007/978-1-4614-1939-6* 69

[Bubenik 2015] Bubenik, Peter: Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* 16 (2015), 77–102. *http://jmlr.org/papers/v16/bubenik15a.html* 29

[Buncher & Carrasco Kind 2020] Buncher, Brandon, Carrasco Kind, Matias: Probabilistic cosmic web classification using fast-generated training data. *Mon. Not. R. Astron. Soc.* 487 (2020), Nr. September, S. 5041–5060. *http://dx.doi.org/10.1093/mnras/staa2008* 33, 34

[Butcher & Oemler, A. 1984] Butcher, H., Oemler, A., Jr.: The evolution of galaxies in clusters. V. A study of populations since z = 0.5. *Astrophys. J.* 285 (1984), 426–438. *http://dx.doi.org/10.1086/162519* 29

[Cadiou et al. 2020] Cadiou, C., Pichon, C., Codis, S., et al.: When do cosmic peaks, filaments, or walls merge? A theory of critical events in a multiscale landscape. *Mon. Not. R. Astron. Soc.* 496 (2020), Nr. 4, 4787–4821. *http://dx.doi.org/10.1093/mnras/staa1853* 30

[Capparelli et al. 2018] Capparelli, Ludovico, Di Valentino, Eleonora, Melchiorri, Alessandro, Chluba, Jens: Impact of theoretical assumptions in the determination of the neutrino effective number from future CMB measurements. *Phys. Rev. D* 97 (2018), Nr. 6, S. 1–7. *http://dx.doi.org/10.1103/PhysRevD.97.063519* 19

[Carrasco et al. 2012] Carrasco, John Joseph M., Hertzberg, Mark P., Senatore, Leonardo: The effective field theory of cosmological large scale structures. *J. High Energy Phys.* 2012 (2012), Nr. 9, S. 1–41. *http://dx.doi.org/10.1007/JHEP09(2012)082* 23

[Carrasco Kind & Brunner 2013] Carrasco Kind, Matias, Brunner, Robert J.: TPZ: Photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Mon. Not. R. Astron. Soc.* 432 (2013), Nr. 2, S. 1483–1501. *http://dx.doi.org/10.1093/mnras/stt574* 6, 40

[Carron 2013] Carron, J.: On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices. *Astron. Astrophys.* 551 (2013), S. 10–14. *http://dx.doi.org/10.1051/0004-6361/201220538* 129

[Cautun et al. 2014] Cautun, Marius, Weygaert, Rien Van D., Jones, Bernard J T., Frenk, Carlos S.: Evolution of the cosmic web. *Mon. Not. R. Astron. Soc.* 441 (2014), S. 2923–2973. *http://dx.doi.org/10.1093/mnras/stu768* 29, 30, 31, 99, 109, 124, 145, 146

[Cautun et al. 2013] Cautun, Marius, Weygaert, Rien van d., Jones, Bernard J.: Nexus: Tracing the cosmic web connection. *Mon. Not. R. Astron. Soc.* 429 (2013), Nr. 2, S. 1286–1308. *http://dx.doi.org/10.1093/mnras/sts416* 31, 32, 34, 100, 105

[Cavendish 1974] Cavendish, James C.: Automatic triangulation of arbitrary planar domains for the finite element method. *Int. J. Numer. Methods Eng.* 8 (1974), Nr. 4, 679–696. *http://dx.doi.org/https://doi.org/10.1002/nme.1620080402* 71

[Cen & Ostriker 2006] Cen, Renyue, Ostriker, Jeremiah P.: Where Are the Baryons? II. Feedback Effects. *Astrophys. J.* 650 (2006), Nr. 2, 560–572. *http://dx.doi.org/10.1086/506505* 30

[Chabanier et al. 2019] Chabanier, Solene, Millea, Marius, Palanque-Delabrouille, Nathalie: Matter power spectrum: From Ly α forest to CMB scales. *Mon. Not. R. Astron. Soc.* 489 (2019), Nr. 2, S. 2247–2253. *http://dx.doi.org/10.1093/mnras/stz2310* 20

[Chan & Blot 2017] Chan, Kwan C., Blot, Linda: Assessment of the information content of the power spectrum and bispectrum. *Phys. Rev. D* 96 (2017), Nr. 2. *http://dx.doi.org/10.1103/PhysRevD.96.023528* 132

[Chazal et al. 2018] Chazal, F., Fasy, B., Lecci, B., et al.: Robust Topological Inference: Distance To a Measure and Kernel Distance. *J. Mach. Learn. Res.* 18 (2018), 1–40. *http://jmlr.org/papers/v18/15-484.html* 81

[Chazal et al. 2011] Chazal, Frédéric, Cohen-Steiner, David, Mérigot, Quentin: Geometric Inference for Probability Measures. *Found. Comput. Math.* 11 (2011), Nr. 6, S. 733–751. *http://dx.doi.org/10.1007/s10208-011-9098-0* 81

[Chazal & Michel 2017] Chazal, Frédéric, Michel, Bertrand: An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *arXiv e-prints* (2017), 1–38. *http://dx.doi.org/10.3389/frai.2021.667963* 29

[Chen et al. 2014] Chen, Yen-Chi, Genovese, Christopher R., Wasserman, Larry: Generalized Mode and Ridge Estimation. *arXiv e-prints* (2014), 1–12. *http://arxiv.org/abs/1406.1803* 32, 34, 73, 78, 81

[Chen et al. 2016] Chen, Yen C., Ho, Shirley, Brinkmann, Jon, et al.: Cosmic web reconstruction through density ridges: Catalogue. *Mon. Not. R. Astron. Soc.* 461 (2016), Nr. 4, S. 3896–3909. *http://dx.doi.org/10.1093/mnras/stw1554* 36, 94

[Chen et al. 2015] Chen, Yen C., Ho, Shirley, Freeman, Peter E., et al.: Cosmic web reconstruction through density ridges: Method and algorithm. *Mon. Not. R. Astron. Soc.* 454 (2015), Nr. 1, S. 1140–1156. *http://dx.doi.org/10.1093/mnras/stv1996* 34, 78, 81, 87, 104

[Cheng et al. 2020] Cheng, Sihao, Ting, Yuan-Sen, Ménard, Brice, Bruna, Joan: A new approach to observational cosmology using the scattering transform. *Mon. Not. R. Astron. Soc.* 499 (2020), 5902–5914. *http://dx.doi.org/10.1093/mnras/staa3165* 5, 6, 29, 40

[Choromanska et al. 2015] Choromanska, Anna, Henaff, Mikael, Mathieu, Michael, et al.: The loss surfaces of multilayer networks. *J. Mach. Learn. Res.* 38 (2015), S. 192–204 41

[Chung 1999] Chung, Fan R. K.: Lectures on Spectral Graph Theory. *ACM SIGACT News* 30 (1999), Nr. 2, 14. *http://dx.doi.org/10.1145/568547.568553* 69, 70, 74

[Codis et al. 2015] Codis, S., Gavazzi, R., Dubois, Y., et al.: Intrinsic alignment of simulated galaxies in the cosmic web: Implications for weak lensing surveys. *Mon. Not. R. Astron. Soc.* 448 (2015), Nr. 4, S. 3391–3404. *http://dx.doi.org/10.1093/mnras/stv231* 94

[Codis et al. 2012] Codis, Sandrine, Pichon, Christophe, Devriendt, Julien, et al.: Connecting the cosmic web to the spin of dark haloes: Implications for galaxy formation. *Mon. Not. R. Astron. Soc.* 427 (2012), Nr. 4, 3320–3336. *http://dx.doi.org/10.1111/j.1365-2966.2012.21636.x* 94

[Codis et al. 2018] Codis, Sandrine, Pogosyan, Dmitri, Pichon, Christophe: On the connectivity of the cosmic web: Theory and implications for cosmology and galaxy formation. *Mon. Not. R. Astron. Soc.* 479 (2018), Nr. 1, S. 973–993. *http://dx.doi.org/10.1093/mnras/sty1643* 30, 111, 112, 117

[Colberg 2007] Colberg, Jörg M.: Quantifying cosmic superstructures. *Mon. Not. R. Astron. Soc.* 375 (2007), Nr. 1, S. 337–347. *http://dx.doi.org/10.1111/j.1365-2966.2006.11312.x* 32, 68, 71, 95

[Colberg et al. 2005] Colberg, Jörg M., Krughoff, K. S., Connolly, Andrew J.: Intercluster filaments in a ACDM Universe. *Mon. Not. R. Astron. Soc.* 359 (2005), Nr. 1, S. 272–282. *http://dx.doi.org/10.1111/j.1365-2966.2005.08897.x* 97, 109

[Cole & Lacey 1996] Cole, Shaun, Lacey, Cedric: The structure of dark matter haloes in hierarchical clustering models. *Mon. Not. R. Astron. Soc.* 281 (1996), 716. *http://dx.doi.org/10.1093/mnras/281.2.716* 115

[Collaboration et al. 2009] Collaboration, LSST S., Abell, Paul A., Allison, Julius, et al.: LSST Science Book, Version 2.0. *arXiv e-prints* (2009), 1–596. *https://ui.adsabs.harvard.edu/abs/2009arXiv0912.0201L* 5, 23

[Colless et al. 2001] Colless, Matthew, Dalton, Gavin, Maddox, Steve, et al.: The 2dF Galaxy Redshift Survey: Spectra and redshifts. *Mon. Not. R. Astron. Soc.* 328 (2001), Nr. 4, S. 1039–1063. *http://dx.doi.org/10.1046/j.1365-8711.2001.04902.x* 5, 23, 94

[Contigiani et al. 2021] Contigiani, O, Bahé, Y M., Hoekstra, H: The mass–size relation of galaxy clusters. *Mon. Not. R. Astron. Soc.* 505 (2021), Nr. 2, 2932–2940. *http://dx.doi.org/10.1093/mnras/stab1463* 111

[Corasaniti et al. 2021] Corasaniti, Pier-Stefano, Sereno, Mauro, Ettori, Stefano: Cosmological Constraints from Galaxy Cluster Sparsity, Cluster Gas Mass Fraction, and Baryon Acoustic Oscillation Data. *Astrophys. J.* 911 (2021), Nr. 2, S. 82. *http://dx.doi.org/10.3847/1538-4357/abe9a4* 120

[Costanzi et al. 2013] Costanzi, Matteo, Villaescusa-Navarro, Francisco, Viel, Matteo, et al.: Cosmology with massive neutrinos III: The halo mass function and an application to galaxy clusters. *J. Cosmol. Astropart. Phys.* 2013 (2013), Nr. 12. *http://dx.doi.org/10.1088/1475-7516/2013/12/012* 120

[Coutinho et al. 2016] Coutinho, B. C., Hong, Sungryong, Albrecht, Kim, et al.: The Network Behind the Cosmic Web. *arXiv e-prints* (2016). *http://arxiv.org/abs/1604.03236* 68, 117

[Crain et al. 2015] Crain, Robert A., Schaye, Joop, Bower, Richard G., et al.: The EAGLE simulations of galaxy formation: Calibration of subgrid physics and model variations. *Mon. Not. R. Astron. Soc.* 450 (2015), Nr. 2, 1937–1961. *http://dx.doi.org/10.1093/mnras/stv725* 23

[Darragh Ford et al. 2019] Darragh Ford, E., Laigle, C., Gozaliasl, G., et al.: Group connectivity in COSMOS: A tracer of mass assembly history. *Mon. Not. R. Astron. Soc.* 489 (2019), Nr. 4, S. 5695–5708. *http://dx.doi.org/10.1093/mnras/stz2490* 30, 111, 112, 115

[D'Ascoli et al. 2020] D'Ascoli, Stéphane, Refinetti, Maria, Biroli, Giulio, Krzakala, Florent: Double trouble in double descent: Bias and variance(s) in the lazy regime. *Proc. 37th Int. Conf. Mach. Learn.* 119 (2020), 2280—-2290. *http://proceedings.mlr.press/v119/d-ascoli20a.html* 41

[Davis et al. 1985] Davis, M., Efstathiou, G., Frenk, C. S., White, S. D. M.: The evolution of large-scale structure in a universe dominated by cold dark matter. *Astrophys. J.* 292 (1985), 371–394. *http://dx.doi.org/10.1086/163168* 97, 111

[Decelle & Furtlehner 2021] Decelle, Aurélien, Furtlehner, Cyril: Restricted Boltzmann machine: Recent advances and mean-field theory. *Chinese Phys. B* 30 (2021), Nr. 4, S. 1–44. *http://dx.doi.org/10.1088/1674-1056/abd160* 41, 62

[Decelle et al. 2011] Decelle, Aurelien, Krzakala, Florent, Moore, Cristopher, Zdeborová, Lenka: Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* 84 (2011), Nr. 6, S. 066106. *http://dx.doi.org/10.1103/PhysRevE.84.066106* 41

[Dempster et al. 1977] Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc.* 39 (1977), 1–38. *https://www.jstor.org/stable/2984875* 46

[DES Collaboration et al. 2020] DES Collaboration, Abbott, T. M., Aguena, M., et al.: Dark Energy Survey Year 1 Results: Cosmological constraints from cluster abundances and weak lensing. *Phys. Rev. D* 102 (2020), Nr. 2, 1–35. *http://dx.doi.org/10.1103/PhysRevD.102.023509* 120

[Desjacques et al. 2019] Desjacques, Vincent, Jeong, Donghui, Schmidt, Fabian: Large-Scale Galaxy Bias. *arXiv e-prints* (2019). *https://arxiv.org/abs/1611.09787* 27

[Di Valentino et al. 2017] Di Valentino, Eleonora, Melchiorri, Alessandro, Linder, Eric V., Silk, Joseph: Constraining dark energy dynamics in extended parameter space. *Phys. Rev. D* 96 (2017), Nr. 2, S. 1–11. *http://dx.doi.org/10.1103/PhysRevD.96.023523* 19

[Di Valentino et al. 2012] Di Valentino, Eleonora, Melchiorri, Alessandro, Salvatelli, Valentina, Silvestri, Alessandra: Parametrized modified gravity and the CMB bispectrum. *Phys. Rev. D - Part. Fields, Gravit. Cosmol.* 86 (2012), Nr. 6, S. 1–8. *http://dx.doi.org/10.1103/PhysRevD.86.063517* 19

[Dietrich et al. 2012] Dietrich, Jörg P., Werner, Norbert, Clowe, Douglas, et al.: A filament of dark matter between two clusters of galaxies. *Nature* 487 (2012), Nr. 7406, S. 202–204. *http://dx.doi.org/10.1038/nature11224* 36

[Dolag et al. 2009] Dolag, K., Borgani, S., Murante, G., Springel, V.: Substructures in hydrodynamical cluster simulations. *Mon. Not. R. Astron. Soc.* 399 (2009), Nr. 2, 497–514. *http://dx.doi.org/10.1111/j.1365-2966.2009.15034.x* 63

[Donnari et al. 2019] Donnari, Martina, Pillepich, Annalisa, Nelson, Dylan, et al.: The star formation activity of Illustris TNG galaxies: Main sequence, UVJ diagram, quenched fractions, and systematics. *Mon. Not. R. Astron. Soc.* 485 (2019), Nr. 4, S. 4817–4840. *http://dx.doi.org/10.1093/mnras/stz712* 23

[Doroshkevich & Shandarin 1978] Doroshkevich, Shandarin: A statistical approach to the theory of galaxy formation. *Sov. Astron.* 22 (1978), 653–660. *https://ui.adsabs.harvard.edu/abs/1978SvA....22..653D* 5, 21, 22

[Dressler 1980] Dressler, A: Galaxy morphology in rich clusters: implications for the formation and evolution of galaxies. *Astrophys. J.* 236 (1980), 351–365. *http://dx.doi.org/10.1086/157753* 94

[Driver et al. 2009] Driver, Simon P., Norberg, P., Baldry, I. K., et al.: GAMA: Towards a physical understanding of galaxy formation. *Astron. Geophys.* 50 (2009), Nr. 5, 5.12–5.19. *http://dx.doi.org/10.1111/j.1468-4004.2009.50512.x* 5, 23

[Dubois et al. 2014] Dubois, Y, Pichon, C, Welker, C, et al.: Dancing in the dark : galactic properties trace spin swings along the cosmic web. *Mon. Not. R. Astron. Soc.* 444 (2014), Nr. 2, S. 1453–1468. *http://dx.doi.org/10.1093/mnras/stu1227* 5, 23

[Duchamp & Stuetzle 1996] Duchamp, Tom, Stuetzle, Werner: Extremal properties of principal curves in the plane. *Ann. Stat.* 24 (1996), Nr. 4, S. 1511–1520. *http://dx.doi.org/10.1214/aos/1032298280* 74

[Duque et al. 2021] Duque, Javier C., Migliaccio, Marina, Marinucci, Domenico, Vittorio, Nicola: A novel Cosmic Filament catalogue from SDSS data. *arXiv e-prints* (2021), 1–21. *http://arxiv.org/abs/2106.05253* 32, 34, 94

[Eckert et al. 2015] Eckert, Dominique, Jauzac, Mathilde, Shan, Huanyuan, et al.: Warm-hot baryons comprise 5-10 per cent of filaments in the cosmic web. *Nature* 528 (2015), 105–107. *https://doi.org/10.1038/nature16058* 36

[Edelsbrunner & Harer 2008] Edelsbrunner, Herbert, Harer, John: Persistent homology - a survey. Version: 2008. *https://www.bibsonomy.org/bibtex/27b57db8b855314c9acadd65091d14707/qmerigot*. In: *Surv. Discret. Comput. Geom. Twenty Years Later.* 2008, 257–282 29

[Edelsbrunner et al. 2002] Edelsbrunner, Herbert, Letscher, David, Zomorodian, Afra: Topological Persistence and Simplification. *Discret. Comput. Geom.* 28 (2002), 511. *http://dx.doi.org/10.1007/s00454-002-2885-2* 81, 90

[Efros & Leung 1999] Efros, Alexei A., Leung, Thomas K.: Texture synthesis by non-parametric sampling. In: *Proc. IEEE Int. Conf. Comput. Vis.* Bd. 2, 1999. *http://dx.doi.org/10.1109/iccv.1999.790383*, S. 1033–1038 16

[Efstathiou et al. 1985] Efstathiou, G., Davis, M., White, S.D.M., Frenk, C.S.: Numerical techniques for large cosmological N-body simulations. *Astrophys. J.* 57 (1985), 241–260. *http://dx.doi.org/10.1086/191003* 22

[Einasto et al. 1980] Einasto, J., Joeveer, M., Saar, E.: Structure of superclusters and supercluster formation. *Mon. Not. R. Astron. Soc.* 193 (1980), 353–375. *http://dx.doi.org/10.1093/mnras/193.2.353* 5

[Epps & Hudson 2017] Epps, D, Hudson, Michael J.: The Weak Lensing Masses of Filaments between Luminous Red Galaxies. *Mon. Not. R. Astron. Soc.* 468 (2017), Nr. March, 2605–2613. *http://dx.doi.org/10.1093/mnras/stx517* 36

[Estrada 2013] Estrada, Ernesto: Graph and Network Theory in Physics. A Short Introduction. *arXiv e-prints* (2013), 1–53. *https://arxiv.org/abs/1302.4378v2* 68

[Facco et al. 2017] Facco, Elena, D'Errico, Maria, Rodriguez, Alex, Laio, Alessandro: Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* 7 (2017), Nr. 1, 1–8. *http://dx.doi.org/10.1038/s41598-017-11873-y* 66

[Falck & Neyrinck 2015] Falck, B., Neyrinck, M. C.: The persistent percolation of single-stream voids. *Mon. Not. R. Astron. Soc.* 450 (2015), Nr. 3, S. 3239–3253. *http://dx.doi.org/10.1093/mnras/stv879* 33

[Falck et al. 2012] Falck, Bridget L., Neyrinck, Mark C., Szalay, Alexander S.: Origami: Delineating halos using phase-space folds. *Astrophys. J.* 754 (2012), Nr. 2. *http://dx.doi.org/10.1088/0004-637X/754/2/126* 33, 34

[Fischetti & Stringher 2019] Fischetti, Matteo, Stringher, Matteo: Embedded hyper-parameter tuning by Simulated Annealing. *arXiv e-prints* (2019). *http://arxiv.org/abs/1906.01504* 90

[Forero-Romero et al. 2009] Forero-Romero, J. E., Hoffman, Y., Gottlöber, S., et al.: A dynamical classification of the cosmic web. *Mon. Not. R. Astron. Soc.* 396 (2009), Nr. 3, S. 1815–1824. *http://dx.doi.org/10.1111/j.1365-2966.2009.14885.x* 32, 123, 124

[Friedmann 1922] Friedmann, A.: Uber die Krummung des Raumes. *Zeitschrift fur Phys.* 10 (1922), 377–386. *http://dx.doi.org/10.1007/BF01332580* 12

[Galárraga-Espinosa et al. 2021] Galárraga-Espinosa, D., Aghanim, N., Langer, M., Tanimura, H.: Properties of gas phases around cosmic filaments at z=0 in the IllustrisTNG simulation. *Astron. Astrophys.* 649 (2021), A117. *http://dx.doi.org/10.1051/0004-6361/202039781* 30, 63, 111

[Galárraga-Espinosa et al. 2020] Galárraga-Espinosa, Daniela, Aghanim, Nabila, Langer, Mathieu, et al.: Populations of filaments from the distribution of galaxies in numerical simulations. *Astron. Astrophys.* 641 (2020), A173. *http://dx.doi.org/10.1051/0004-6361/202037986* 31, 94, 108, 109, 111

[Ganeshaiah Veena et al. 2019] Ganeshaiah Veena, Punyakoti, Cautun, Marius, Tempel, Elmo, et al.: The Cosmic Ballet II: Spin alignment of galaxies and haloes with large-scale filaments in the EAGLE simulation. *Mon. Not. R. Astron. Soc.* 487 (2019), Nr. 2, S. 1607–1625. *http://dx.doi.org/10.1093/mnras/stz1343* 30, 94

[Ganeshaiah Veena et al. 2018] Ganeshaiah Veena, Punyakoti, Cautun, Marius, Weygaert, Rien van d., et al.: The Cosmic Ballet: Spin and shape alignments of haloes in the cosmic web. *Mon. Not. R. Astron. Soc.* 481 (2018), Nr. 1, 414–438. *http://dx.doi.org/10.1093/mnras/sty2270* 30, 94

[Gay et al. 2012] Gay, Christophe, Pichon, Christophe, Pogosyan, Dmitry: Non-Gaussian statistics of critical sets in 2D and 3D: Peaks, voids, saddles, genus, and skeleton. *Phys. Rev. D - Part. Fields, Gravit. Cosmol.* 85 (2012), Nr. 2. *http://dx.doi.org/10.1103/PhysRevD.85.023011* 30

[Genovese et al. 2014] Genovese, Christopher R., Perone-Pacifico, Marco, Verdinelli, Isabella, Wasserman, Larry: Nonparametric ridge estimation. *Ann. Stat.* 42 (2014), Nr. 4, S. 1511–1545. *http://dx.doi.org/10.1214/14-AOS1218* 32, 78

[Gerber & Whitaker 2013] Gerber, Samuel, Whitaker, Ross: Regularization-free principal curve estimation. *J. Mach. Learn. Res.* 14 (2013), Nr. 3, 1285–1302. *http://jmlr.org/papers/v14/gerber13a.html* 74, 86

[Goldt et al. 2019] Goldt, Sebastian, Advani, Madhu S., Saxe, Andrew M., et al.: Generalisation dynamics of online learning in over-parameterised neural networks. *arXiv e-prints* (2019), Nr. 1, 1–25. *http://arxiv.org/abs/1901.09085* 41

[Goldt et al. 2020] Goldt, Sebastian, Mézard, Marc, Krzakala, Florent, Zdeborová, Lenka: Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model. *Phys. Rev. X* 10 (2020), Nr. 4, S. 1–34. *http://dx.doi.org/10.1103/PhysRevX.10.041044* 62, 149

[González & Padilla 2010] González, Roberto E., Padilla, Nelson D.: Automated detection of filaments in the large-scale structure of the Universe. *Mon. Not. R. Astron. Soc.* 407 (2010), Nr. 3, S. 1449–1463. *http://dx.doi.org/10.1111/j.1365-2966.2010.17015.x* 32, 34

[Goodfellow et al. 2014] Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, et al.: Generative adversarial networks. In: *Proc. 27th Int. Conf. Neural Inf. Process. Syst. - Vol. 2*, 2014. – *https://arxiv.org/abs/1406.2661*, 2672–2680 40

[Gorban & Zinovyev 2005] Gorban, A, Zinovyev, A: Elastic Principal Graphs and Manifolds and their Practical Applications. *Computing* 75 (2005), S. 359–379. *http://dx.doi.org/10.1007/s00607-005-0122-6* 67, 74

[Gorban et al. 2016] Gorban, A. N., Mirkes, E. M., Zinovyev, A.: Robust principal graphs for data approximation. *arXiv e-prints* 2 (2016), Nr. 1, 1–16. *http://dx.doi.org/10.5445/KSP/1000058749/11* 67, 73

[Gorban & Zinovyev 2009] Gorban, Alexander N., Zinovyev, Andrei Y.: Principal graphs and manifolds. Version: 2009. *http://dx.doi.org/10.4018/978-1-60566-766-9.ch002*. In: *Handb. Res. Mach. Learn. Appl. trends algorithms, methods, Tech.* 2009, 28–59 67, 85

[Gott et al. 1986] Gott, J. R., Melott, Adrian L., Dickinson, Mark: The Sponge-Like Topology of Large Scale Structure in the Universe. *Astrophys. J.* 306 (1986), 341. *http://dx.doi.org/10.1086/164347* 28, 33

[Gouin et al. 2021] Gouin, C., Bonnaire, T., Aghanim, N: Shape and connectivity of groups and clusters: Impact of dynamical state and accretion history. *Astron. Astrophys.* 651 (2021), A56. *http://dx.doi.org/10.1051/0004-6361/202140327* 93, 111

[Gouin et al. 2017] Gouin, C., Gavazzi, R., Codis, S., et al.: Multipolar moments of weak lensing signal around clusters. Weighing filaments in harmonic space. *Astron. Astrophys.* 605 (2017), A27. *http://dx.doi.org/10.1051/0004-6361/201730727* 36

[Gouin et al. 2020] Gouin, Céline, Aghanim, N., Bonjean, V., Douspis, M.: Probing the azimuthal environment of galaxies around clusters from cluster core to cosmic filaments. *Astron. Astrophys.* 635 (2020), Nr. A195, A195. *http://dx.doi.org/10.1051/0004-6361/201937218* 30

[de Graaf et al. 2019] Graaf, Anna de, Cai, Yan-chuan, Heymans, Catherine, Peacock, John A.: Probing the missing baryons with the Sunyaev-Zel ' dovich effect from filaments. *Astron. Astrophys.* 624 (2019), A48. *http://dx.doi.org/10.1051/0004-6361/201935159* 18, 36

[Grira et al. 2004] Grira, Nizar, Crucianu, Michel, Boujemaa, Nozha: Unsupervised and Semi-supervised Clustering: a brief survey. In: *A Rev. Mach. Learn. Tech. Process. Multimed. Content*, 2004 *http://cedric.cnam.fr/{~}crucianm/src/BriefSurveyClustering.pdf* 44

[Hagen & Kahng 1992] Hagen, Lars, Kahng, Andrew B.: New Spectral Methods for Ratio Cut Partitioning and Clustering. *IEEE Trans. Comput. Des. Integr. Circuits Syst.* 11 (1992), Nr. 9, S. 1074–1085. *http://dx.doi.org/10.1109/43.159993* 69

[Haggar et al. 2020] Haggar, Roan, Gray, Meghan E., Pearce, Frazer R., et al.: THETHREE-HUNDRED project: Backsplash galaxies in simulations of clusters. *Mon. Not. R. Astron. Soc.* 492 (2020), Nr. 4, 6074–6085. *http://dx.doi.org/10.1093/MNRAS/STAA273* 114

[Hahn & Villaescusa-Navarro 2021] Hahn, Chang H., Villaescusa-Navarro, Francisco: Constraining Mν with the bispectrum II: The total information content of the galaxy bispectrum. *J. Cosmol. Astropart. Phys.* 2021 (2021), Nr. 4, 029. *http://dx.doi.org/10.1088/1475-7516/2021/04/029* 28

[Hahn et al. 2020] Hahn, Chang H., Villaescusa-Navarro, Francisco, Castorina, Emanuele, Scoccimarro, Roman: Constraining Mν with the bispectrum. Part I. Breaking parameter degeneracies. *J. Cosmol. Astropart. Phys.* 2020 (2020), Nr. 3, 0–32. *http://dx.doi.org/10.1088/1475-7516/2020/03/040* 5, 28, 136, 140, 145

[Hahn et al. 2007] Hahn, Oliver, Porciani, Cristiano, Carollo, C. M., Dekel, Avishai: Properties of dark matter haloes in clusters, filaments, sheets and voids. *Mon. Not. R. Astron. Soc.* 375 (2007), Nr. 2, S. 489–499. *http://dx.doi.org/10.1111/j.1365-2966.2006.11318.x* 30, 32, 34, 94, 123

[Hamaus et al. 2015] Hamaus, Nico, Sutter, P. M., Lavaux, Guilhem, Wandelt, Benjamin D.: Probing cosmology and gravity with redshift-space distortions around voids. *J. Cosmol. Astropart. Phys.* 2015 (2015), Nr. 11, S. 0–38. *http://dx.doi.org/10.1088/1475-7516/2015/11/036* 29, 120

[Hamaus et al. 2014] Hamaus, Nico, Wandelt, Benjamin D., Sutter, P. M., et al.: Cosmology with void-galaxy correlations. *Phys. Rev. Lett.* 112 (2014), Nr. 4, S. 1–5. *http://dx.doi.org/10.1103/PhysRevLett.112.041304* 29, 120

[Hamilton 1998] Hamilton, A. J. S.: Linear Redshift Distortions: A Review. Version: 1998. *http://dx.doi.org/10.1007/978-94-011-4960-0_17*. In: *Evol. Universe*. 1998, 185–275 25

[Hartlap et al. 2007] Hartlap, J., Simon, P., Schneider, P.: Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix. *Astron. Astrophys.* 464 (2007), Nr. 1, 399–404. *http://dx.doi.org/10.1051/0004-6361:20066170* 129

[Hastie et al. 2019] Hastie, Trevor, Montanari, Andrea, Rosset, Saharon, Tibshirani, Ryan J.: Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints* (2019). *http://arxiv.org/abs/1903.08560* 41

[Hastie & Stuetzle 1989] Hastie, Trevor, Stuetzle, Werner: Principal curves. *J. Am. Stat. Assoc.* 84 (1989), S. 502–516. *http://dx.doi.org/10.1080/01621459.1989.10478797* 66, 67, 74

[He et al. 2019] He, Siyu, Li, Yin, Feng, Yu, et al.: Learning to Predict the Cosmological Structure Formation. *Proc. Natl. Acad. Sci.* 116 (2019), Nr. 28, S. 13825–13832 40

[Heath 1977] Heath, D.J.: The growth of density perturbations in zero pressure Friedmann-Lema{\^\i}tre universes. *Mon. Not. R. Astron. Soc.* 179 (1977), Nr. 1, 351–358. *https://ui.adsabs.harvard.edu/abs/1977MNRAS.179..351H* 120

[Hébert-Dufresne et al. 2016] Hébert-Dufresne, Laurent, Grochow, Joshua A., Allard, Antoine: Multi-scale structure and topological anomaly detection via a new network statistic: The onion decomposition. *Sci. Rep.* 6 (2016), S. 1–8. *http://dx.doi.org/10.1038/srep31708* 97

[Heidenreich et al. 2013] Heidenreich, Nils-Bastian, Schindler, Anja, Sperlich, Stefan: Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *Adv. Stat. Anal.* 97 (2013), Nr. 4, S. 403–433. *http://dx.doi.org/10.1007/s10182-013-0216-y* 87

[Hein & Audibert 2005] Hein, Matthias, Audibert, Jean Y.: Intrinsic dimensionality estimation of submanifolds in Rˆd. In: *Proc. 22nd Int. Conf. Mach. Learn.*, 2005. *http://dx.doi.org/10.1145/1102351.1102388*, 289–296 66

[Hein et al. 2005] Hein, Matthias, Audibert, Jean Y., Von Luxburg, Ulrike: From graphs to manifolds - Weak and strong pointwise consistency of graph Laplacians. *Proc. 18th Conf. Learn. Theory* (2005), S. 470–485. *http://dx.doi.org/10.1007/11503415_32* 74

[Hein & Markus 2007] Hein, Matthias, Markus, Maier: Manifold Denoising. Version: 2007. *http://papers.nips.cc/paper/2997-manifold-denoising.pdf*. In: *Adv. Neural Inf. Process. Syst. 19*. MIT Press, 2007, 561—-568 69

[Hirschmann et al. 2014] Hirschmann, Michaela, Dolag, Klaus, Saro, Alexandro, et al.: Cosmological simulations of black hole growth: AGN luminosities and downsizing. *Mon. Not. R. Astron. Soc.* 442 (2014), Nr. 3, S. 2304–2324. *http://dx.doi.org/10.1093/mnras/stu1023* 108

[Ho 1995] Ho, Tin K.: Random decision forests. In: *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR* Bd. 1, 1995. *http://dx.doi.org/10.1109/ICDAR.1995.598994*, S. 278–282 40

[Hockney & Eastwood 1981] Hockney, R. W., Eastwood, J. W.: *Computer Simulation Using Particles*. McGraw-Hill, 1981. – 540 S. *https://ui.adsabs.harvard.edu/abs/1981csup.book.....H* 36, 123

[Holder et al. 2001] Holder, G., Haiman, Z., Mohr, J. J.: Constraints on Ωm, $\Omega\Lambda$, and σ8 from Galaxy Cluster Redshift Distributions. *Astrophys. J.* 560 (2001), Nr. 2, L111–L114. *http://dx.doi.org/10.1086/324309* 29, 120

[Hong et al. 2016] Hong, Sungryong, Coutinho, Bruno C., Dey, Arjun, et al.: Discriminating topology in galaxy distributions using network analysis. *Mon. Not. R. Astron. Soc.* 459 (2016), Nr. 3, S. 2690–2700. *http://dx.doi.org/10.1093/mnras/stw803* 117, 148

[Hong & Dey 2015] Hong, Sungryong, Dey, Arjun: Network analysis of cosmic structures: Network centrality and topological environment. *Mon. Not. R. Astron. Soc.* 450 (2015), Nr. 2, S. 1999–2015. *http://dx.doi.org/10.1093/mnras/stv722* 117, 148

[Hong et al. 2021] Hong, Sungwook E., Jeong, Donghui, Seong Hwang, Ho, Kim, Juhan: Revealing the Local Cosmic Web from Galaxies by Deep Learning. *Astrophys. J.* 913 (2021), Nr. 1, 76. *http://dx.doi.org/10.3847/1538-4357/abf040* 40

[Hopfield 1982] Hopfield, John J.: Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.* 79 (1982), Nr. April, 2554—-2558. *http://dx.doi.org/10.1073/pnas.79.8.2554* 41

[Huang et al. 2018] Huang, Jincai, Deng, M I N., Tang, Jianbo, et al.: Automatic Generation of Road Maps from Low Quality GPS Trajectory Data via Structure Learning. *IEEE Access* 6 (2018), S. 71965–71975. *http://dx.doi.org/10.1109/ACCESS.2018.2882581* 89

[Hubble 1929] Hubble, E.: A relation between distance and radial velocity among extra-galactic nebulae. *Proc. Natl. Acad. Sci.* 15 (1929), 168–173. http://dx.doi.org/10.1073/pnas.15.3.168 12

[Jackson 1972] Jackson, J.C.: A critique of Rees's theory of primordial gravitational radiation. *Mon. Not. R. Astron. Soc.* 156 (1972), Nr. 1, 74–92. http://dx.doi.org/10.1093/mnras/156.1.1P 25, 135

[Jain et al. 2000] Jain, A.K., Duin, R. P. W., Mao, Jianchang: Statistical pattern recognition: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2000), 4–37. http://dx.doi.org/10.1109/34.824819 45

[Jalalzai 2012] Jalalzai, Khalid: *Regularization of inverse problems in image processing*, Ecole Polytechnique X, Diss., 2012. https://pastel.archives-ouvertes.fr/pastel-00787790 42

[Jasche & Wandelt 2013] Jasche, Jens, Wandelt, Benjamin D.: Bayesian physical reconstruction of initial conditions from large-scale structure surveys. *Mon. Not. R. Astron. Soc.* 432 (2013), Nr. 2, S. 894–913. http://dx.doi.org/10.1093/mnras/stt449 22, 36, 149

[Jaynes & Rosenkrantz 1983] Jaynes, E. T., Rosenkrantz, Roger D.: *Papers on probability, statistics, and statistical physics.* 1983. – 189–191 S. http://dx.doi.org/10.1007/BF00046910. http://dx.doi.org/10.1007/BF00046910 50

[Jensen 1906] Jensen, J. L. W. V.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math.* 30 (1906), Nr. 2, 175–193. http://dx.doi.org/10.1007/BF02418571 47

[Jing 2005] Jing, Y. P.: Correcting for the Alias Effect When Measuring the Power Spectrum Using a Fast Fourier Transform. *Astrophys. J.* 620 (2005), Nr. 2, 559–563. http://dx.doi.org/10.1086/427087 36, 127

[Jing & Suto 2002] Jing, Y. P., Suto, Yasushi: Triaxial Modeling of Halo Density Profiles with High-Resolution N -Body Simulations. *Astrophys. J.* 574 (2002), Nr. 2, 538–553. http://dx.doi.org/10.1086/341065 112

[Joeveer et al. 1978] Joeveer, M., Einasto, J., Tago, E.: Spatial distribution of galaxies and of clusters of galaxies in the southern galactic hemisphere. *Mon. Not. R. Astron. Soc.* 185 (1978), 357–369. http://dx.doi.org/10.1093/mnras/185.2.357 5, 23

[Kaiser 1984] Kaiser, N.: On the spatial correlations of Abell clusters. *Astrophys. J.* 284 (1984), L9–L12. http://dx.doi.org/10.1086/184341 27

[Kaiser 1987] Kaiser, Nick: Clustering in real space and in redshift space. *Mon. Not. R. Astron. Soc.* 227 (1987), Nr. 1, 1–21. http://dx.doi.org/10.1093/mnras/227.1.1 25, 127, 135

[Kauffmann et al. 2004] Kauffmann, Guinevere, White, Simon D., Heckman, Timothy M., et al.: The environmental dependence of the relations between stellar mass, structure, star formation and nuclear activity in galaxies. *Mon. Not. R. Astron. Soc.* 353 (2004), Nr. 3, S. 713–731. http://dx.doi.org/10.1111/j.1365-2966.2004.08117.x 30, 94

[Kaufman 1964] Kaufman, GM: Some Bayesian moment formulae. *Cent. Oper. Res. Econom. Discuss. Pap.* (1964), Nr. 6710, S. 44–49 129

[Kegl et al. 2000] Kegl, B., Krzyzak, A., Linder, T., Zeger, K.: Learning and design of principal curves. *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2000), S. 281–297 67, 74

[Khintchine 1934] Khintchine, A: Korrelationstheorie der stationären stochastischen Prozesse. *Math. Ann.* 109 (1934), 604–615. *http://eudml.org/doc/159698* 17

[Kingma & Welling 2013] Kingma, Diederik P., Welling, Max: Auto-encoding variational bayes. In: *2nd Int. Conf. Learn. Represent.*, 2013 (ML). *https://arxiv.org/abs/1312.6114*, 1–14 40

[Kirkpatrick et al. 1983] Kirkpatrick, S, Gelatt, C D., Vecchi, M P.: Optimization by Simulated Annealing. *Science (80-.).* 220 (1983), Nr. 4598, S. 671—-680 51, 87

[Kitaura & Angulo 2012] Kitaura, F. S., Angulo, R. E.: Linearisation with Cosmological Perturbation Theory. *Mon. Not. R. Astron. Soc.* 425 (2012), 2443–2454. *http://dx.doi.org/10.1111/j.1365-2966.2012.21614.x* 34

[Kitaura 2013] Kitaura, Francisco-shu: The Initial Conditions of the Universe from Constrained Simulations. *Mon. Not. R. Astron. Soc.* 429 (2013), Nr. Feb, S. L84–L88. *http://dx.doi.org/10.1093/mnrasl/sls029* 32

[Kloppenburg & Tavan 1997] Kloppenburg, Martin, Tavan, Paul: Deterministic annealing for density estimation by multivariate normal mixtures. *Phys. Rev. E* 55 (1997), Nr. 3, 2089–2092. *http://dx.doi.org/10.1103/PhysRevE.55.R2089* 48

[Klypin & Shandarin 1983] Klypin, A.˜A., Shandarin, S.˜F.: Three-dimensional numerical model of the formation of large-scale structure in the Universe. *Mon. Not. R. Astron. Soc.* 204 (1983), S. 891–907. *http://dx.doi.org/10.1093/mnras/204.3.891* 21

[Klypin et al. 2016] Klypin, Anatoly, Yepes, Gustavo, Gottlöber, Stefan, et al.: Multidark simulations: The story of dark matter halo concentrations and density profiles. *Mon. Not. R. Astron. Soc.* 457 (2016), Nr. 4, S. 4340–4359. *http://dx.doi.org/10.1093/mnras/stw248* 23, 115

[Kodwani et al. 2019] Kodwani, Darsh, Alonso, David, Ferreira, Pedro G.: The effect on cosmological parameter estimation of a parameter-dependent covariance matrix. *Open J. Astrophys.* 2 (2019), Nr. 1, 3. *http://dx.doi.org/10.21105/astro.1811.11584* 129

[Kofman et al. 1992] Kofman, L., Pogosyan, D., Shandarin, S. F., Melott, A. L.: Coherent structures in the universe and the adhesion model. *Astrophys. J.* 393 (1992), Nr. 2, 437–449. *http://dx.doi.org/10.1086/171517* 15

[Kofman & Shandarin 1988] Kofman, L. A., Shandarin, S. F.: Theory of adhesion for the large-scale structure of the Universe. *Nature* 334 (1988), 129–131. *http://dx.doi.org/10.1038/334129a0* 15

[Kolatt et al. 1996] Kolatt, T., Dekel, A., Ganon, G., Willick, J. A.: Simulating Our Cosmological Neighborhood: Mock Catalogs for Velocity Analysis. *Astrophys. J.* 458 (1996), 419. *http://adsabs.harvard.edu/full/1996ApJ...458..419K* 22

[Komatsu et al. 2011] Komatsu, E., Smith, K. M., Dunkley, J., et al.: Seven-year wilkinson microwave anisotropy probe (WMAP*) observations: Cosmological interpretation. *Astrophys. Journal, Suppl. Ser.* 192 (2011), Nr. 2. *http://dx.doi.org/10.1088/0067-0049/192/2/18* 18, 108

[Koutrouli et al. 2020] Koutrouli, Mikaela, Karatzas, Evangelos, Paez-Espino, David, Pavlopoulos, Georgios A.: A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* 8 (2020), Nr. January, 1–26. *http://dx.doi.org/10.3389/fbioe.2020.00034* 68

[Kraljic et al. 2018] Kraljic, K., Arnouts, S., Pichon, C., et al.: Galaxy evolution in the metric of the cosmic web. *Mon. Not. R. Astron. Soc.* 474 (2018), Nr. 1, S. 547–571. *http://dx.doi.org/10.1093/MNRAS/STX2638* 94

[Kraljic et al. 2020] Kraljic, Katarina, Davé, Romeel, Pichon, Christophe: And yet it flips: connecting galactic spin and the cosmic web. *Mon. Not. R. Astron. Soc.* (2020), S. 237. *http://dx.doi.org/10.1093/mnras/staa250* 30, 36, 94

[Kratochvil et al. 2012] Kratochvil, Jan M., Lim, Eugene A., Wang, Sheng, et al.: Probing cosmology with weak lensing Minkowski functionals. *Phys. Rev. D - Part. Fields, Gravit. Cosmol.* 85 (2012), Nr. 10, S. 1–19. *http://dx.doi.org/10.1103/PhysRevD.85.103513* 28

[Kreisch et al. 2019] Kreisch, Christina D., Pisani, Alice, Carbone, Carmelita, et al.: Massive neutrinos leave fingerprints on cosmic voids. *Mon. Not. R. Astron. Soc.* 488 (2019), Nr. 3, 4413–4426. *http://dx.doi.org/10.1093/mnras/stz1944* 121, 132

[Kreisch et al. 2021] Kreisch, Christina D., Pisani, Alice, Villaescusa-Navarro, Francisco, et al.: The GIGANTES dataset: precision cosmology from voids in the machine learning era. *arXiv e-prints* (2021). *http://arxiv.org/abs/2107.02304* 121

[Kruskal 1956] Kruskal, Joseph B.: On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Am. Math. Soc.* 7 (1956), 48–50. *http://www.jstor.org/stable/2033241* 71

[Kuchment 2008] Kuchment, Peter: Quantum graphs: An introduction and a brief survey. In: *Proc. Symp. Pure Math.*, 2008. *http://dx.doi.org/10.1090/pspum/077/2459876*, 291–312 68

[Kuchner et al. 2020] Kuchner, Ulrike, Aragón-Salamanca, Alfonso, Pearce, Frazer R., et al.: Mapping and characterization of cosmic filaments in galaxy cluster outskirts: Strategies and forecasts for observations from simulations. *Mon. Not. R. Astron. Soc.* 494 (2020), Nr. 4, 5473–5491. *http://dx.doi.org/10.1093/mnras/staa1083* 114

[Kukačka et al. 2017] Kukačka, Jan, Golkov, Vladimir, Cremers, Daniel: Regularization for deep learning: A taxonomy. In: *arXiv e-prints*, 2017 *https://openreview.net/forum?id=SkHkeixAW*, 1–23 42

[Kullback & Leibler 1951] Kullback, S, Leibler, R.A: On information and sufficiency. *Ann. Math. Stat.* 22 (1951), Nr. 1, 79–86. *http://dx.doi.org/10.1214/aoms/1177729694* 47

[Kurlin 2015] Kurlin, V: A One-Dimensional Homologically Persistent Skeleton of an Unstructured Point Cloud in Any Metric Space. In: *Eurographics Assoc.* Bd. 34. Graz, Austria : Eurographics Association, 2015. *http://dx.doi.org/10.1111/cgf.12713*, 253–262 81, 83, 90

[Kuruvilla & Aghanim 2021] Kuruvilla, Joseph, Aghanim, Nabila: Information content in mean pairwise velocity and mean relative velocity between pairs in a triplet. *Astron. Astrophys.* 653 (2021), Nr. 2019, A130. *http://dx.doi.org/10.1051/0004-6361/202140552* 28

[Kuutma et al. 2017] Kuutma, Teet, Tamm, Antti, Tempel, Elmo: From voids to filaments : environmental transformations of galaxies in the SDSS. *Astron. Astrophys.* 600 (2017), S. L6. *http://dx.doi.org/10.1051/0004-6361/201730526* 30

[Laigle et al. 2018] Laigle, C., Pichon, C., Arnouts, S., et al.: COSMOS2015 photometric redshifts probe the impact of filaments on galaxy properties. *Mon. Not. R. Astron. Soc.* 474 (2018), Nr. 4, S. 5437–5458. *http://dx.doi.org/10.1093/MNRAS/STX3055* 30, 36

[de Lapparent et al. 1987] Lapparent, V de, Geller, M.˜J., Huchra, J.˜P.: A Slice of the Universe. *Astrophys. J.* 302 (1987), S. L1. *http://dx.doi.org/10.1086/184625* 5, 14

[Laureijs et al. 2011] Laureijs, R., Amiaux, J., Arduini, S., et al.: Euclid Definition Study Report. *arXiv e-prints* (2011). *https://ui.adsabs.harvard.edu/abs/2011arXiv1110.3193L* 5, 18, 23, 146

[Lavaux & Wandelt 2010] Lavaux, Guilhem, Wandelt, Benjamin D.: Precision cosmology with voids: Definition, methods, dynamics. *Mon. Not. R. Astron. Soc.* 403 (2010), Nr. 3, S. 1392–1408. *http://dx.doi.org/10.1111/j.1365-2966.2010.16197.x* 33

[Lavaux & Wandelt 2012] Lavaux, Guilhem, Wandelt, Benjamin D.: Precision cosmography with stacked voids. *Astrophys. J.* 754 (2012), Nr. 2. *http://dx.doi.org/10.1088/0004-637X/754/2/109* 29, 120

[Leclercq et al. 2013] Leclercq, Florent, Jasche, Jens, Gil-Marín, Héctor, Wandelt, Benjamin: One-point remapping of Lagrangian perturbation theory in the mildly non-linear regime of cosmic structure formation. *J. Cosmol. Astropart. Phys.* 2013 (2013), Nr. 11, 1–22. *http://dx.doi.org/10.1088/1475-7516/2013/11/048* 36

[Leclercq et al. 2015] Leclercq, Florent, Jasche, Jens, Lavaux, Guilhem, Wandelt, Benjamin: Probabilistic cartography of the large-scale structure. *arXiv e-prints* (2015), S. 1–4 149

[LeCun et al. 2015] LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey: Deep learning. *Nature* 521 (2015), 436–444. *http://dx.doi.org/10.1038/nature14539* 40

[LeCun et al. 1999] LeCun, Yann, Haffner, Patrick, Bottou, Léon, Bengio, Yoshua: Object Recognition with Gradient-Based Learning. In: *Shape, Contour Group. Comput. Vis.*, 1999. *http://dx.doi.org/10.5555/646469.691875*, 319 28, 41

[Lee & Park 2009] Lee, Jounghun, Park, Daeseong: Constraining the dark energy equation of state with cosmic voids. *Astrophys. J.* 696 (2009), Nr. 1, L10–L12. *http://dx.doi.org/10.1088/0004-637X/696/1/L10* 29, 120

[Lemaitre 1931] Lemaitre, A. G.: Expansion of the universe, A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae. *Mon. Not. R. Astron. Soc.* 91 (1931), Nr. 5, S. 483–490. *http://dx.doi.org/10.1093/mnras/91.5.483* 12

[Lesieur et al. 2016] Lesieur, Thibault, De Bacco, Caterina, Banks, Jess, et al.: Phase transitions and optimal algorithms in high-dimensional Gaussian mixture clustering. In: *2016 54th Annu. Allert. Conf. Commun. Control. Comput.*, 2016. *http://dx.doi.org/10.1109/ALLERTON.2016.7852287*, 601–608 41

[Levi et al. 2013] Levi, Michael, Bebek, Chris, Beers, Timothy, et al.: The DESI Experiment, a whitepaper for Snowmass 2013. *arXiv e-prints* (2013), 1–14. *http://arxiv.org/abs/1308.0847* 5, 24, 146

[Li & Barron 2000] Li, Jonathan Q., Barron, Andrew R.: Mixture Density Estimation. Version: 2000. *https://papers.nips.cc/paper/1999/hash/a0f3601dc682036423013a5d965db9aa-Abstract.html*. In: Solla, S. A. (Hrsg.), Leen, T. K. (Hrsg.), Muller, K. (Hrsg.): *Adv. Neural Inf. Process. Syst.* MIT Press, 2000, 279—-285 45

[Libeskind et al. 2017] Libeskind, Noam I., Weygaert, Rien van d., Cautun, Marius, et al.: Tracing the cosmic web. *Mon. Not. R. Astron. Soc.* 473 (2017), Nr. 1, S. 1195–1217. *http://dx.doi.org/10.1093/mnras/stx1976* 5, 32, 34, 36, 94, 100, 101, 124, 126, 146

[Linder 2005] Linder, Eric V.: Cosmic growth history and expansion history. *Phys. Rev. D - Part. Fields, Gravit. Cosmol.* 72 (2005), Nr. 4, 1–8. *http://dx.doi.org/10.1103/PhysRevD.72.043529* 25

[Little 1974] Little, W. A.: The existence of persistent states in the brain. *Math. Biosci.* 19 (1974), Nr. 1-2, S. 101–120. *http://dx.doi.org/10.1016/0025-5564(74)90031-5* 41

[Little & Shaw 1978] Little, W.A., Shaw, Gordon L.: Analytic study of the memory storage capacity of a neural network. *Math. Biosci.* 39 (1978), Nr. 3, 281–290. *http://dx.doi.org/10.1016/0025-5564(78)90058-5* 41

[Logan & Fotopoulou 2020] Logan, C. H., Fotopoulou, S.: Unsupervised star, galaxy, QSO classification: Application of HDBSCAN. *Astron. Astrophys.* 633 (2020), 1–25. *http://dx.doi.org/10.1051/0004-6361/201936648* 44

[López-Corredoira 2014] López-Corredoira, M.: Alcock-Paczyński cosmological test. *Astrophys. J.* 781 (2014), Nr. 2, 1–57. *http://dx.doi.org/10.1088/0004-637X/781/2/96* 25

[van der Maaten & Geoffrey 2008] Maaten, Laurens van d., Geoffrey, Hinton: Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9 (2008), Nr. 86, 2579–2605. *http://jmlr.org/papers/v9/vandermaaten08a.html* 66

[MacKay 2002] MacKay, David J. C.: *Information Theory, Inference & Learning Algorithms.* 2002. *http://dx.doi.org/10.5555/971143*. *http://dx.doi.org/10.5555/971143* 43

[MacQueen 1967] MacQueen, J. B.: Some methods for classification and analysis of multivariate observations. Version: 1967. *http://dx.doi.org/10.1007/s11665-016-2173-6*. In: Cam, L. M. L. (Hrsg.), Neyman, J. (Hrsg.): *Proc. fifth Berkeley Symp. Math. Stat. Probab.* Bd. 1. University of California Press, 1967, 281–297 43

[Maeder 2017] Maeder, Andre: an Alternative To the Λcdm Model: the Case of Scale Invariance. *Astrophys. J.* 834 (2017), Nr. 2, S. 194. *http://dx.doi.org/10.3847/1538-4357/834/2/194* 19

[Malavasi et al. 2017] Malavasi, N., Arnouts, S., Vibert, D., et al.: The VIMOS Public Extragalactic Redshift Survey (VIPERS): Galaxy segregation inside filaments at z = 0.7. *Mon. Not. R. Astron. Soc.* 465 (2017), Nr. 4, S. 3817–3822. *http://dx.doi.org/10.1093/mnras/stw2864* 30, 36, 94

[Malavasi et al. 2020a] Malavasi, Nicola, Aghanim, Nabila, Douspis, Marian, et al.: Characterising filaments in the SDSS volume from the galaxy distribution. *Astron. Astrophys.* 642 (2020), Nr. A19, A19. *http://dx.doi.org/10.1051/0004-6361/202037647* 31

[Malavasi et al. 2020b] Malavasi, Nicola, Aghanim, Nabila, Tanimura, Hideki, et al.: Like a spider in its web : a study of the Large Scale Structure around the Coma cluster. *Astron. Astrophys.* 634 (2020), A30. *https://arxiv.org/abs/1910.11879* 36, 94, 108, 112, 117

[Mallat 2012] Mallat, Stéphane: Group Invariant Scattering. *Commun. Pure Appl. Math.* 65 (2012), Nr. 10, S. 1331–1398. *http://dx.doi.org/10.1002/cpa.21413* 28

[Mantz et al. 2015] Mantz, Adam B., Von der Linden, Anja, Allen, Steven W., et al.: Weighing the giants - IV. cosmology and neutrino mass. *Mon. Not. R. Astron. Soc.* 446 (2015), Nr. 3, 2205–2225. *http://dx.doi.org/10.1093/mnras/stu2096* 120

[Mao et al. 2017] Mao, Qi, Li, Wang, Ivor W., Tsang, Sun, Yijun: Principal Graph and Structure Learning Based on Reversed Graph Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (2017), Nr. 11, 2227–2241. *http://dx.doi.org/10.1109/TPAMI.2016.2635657* 77, 82, 86, 90

[Mao et al. 2015] Mao, Qi, Yang, Le, Wang, Li, et al.: SimplePPT: A Simple Principal Tree Algorithm. *Proc. 2015 SIAM Int. Conf. Data Min.* (2015), 792–800. *http://dx.doi.org/10.1137/1.9781611974010.89* 67, 77, 78, 90

[Martinez et al. 2016] Martinez, H.J., Muriel, H, Coenda, V: Galaxies infalling into groups : filaments vs . isotropic infall. *Mon. Not. R. Astron. Soc.* 455 (2016), Nr. 1, 127–135. *http://dx.doi.org/10.1093/mnras/stv2295* 30

[Martinez et al. 1993] Martinez, Vicent J., Paredes, Silvestre, Saar, Enn: Wavelet analysis of the multifractal character of the galaxy distribution. *Mon. Not. R. Astron. Soc.* 260 (1993), Nr. 2, 365–375. *http://dx.doi.org/10.1093/mnras/260.2.365* 28

[Martizzi et al. 2019] Martizzi, Davide, Vogelsberger, Mark, Artale, Maria C., et al.: Baryons in the Cosmic Web of IllustrisTNG - I: Gas in knots, filaments, sheets, and voids. *Mon. Not. R. Astron. Soc.* 486 (2019), Nr. 3, S. 3766–3787. *http://dx.doi.org/10.1093/mnras/stz1106* 30, 63, 124

[Massara et al. 2021] Massara, Elena, Villaescusa-Navarro, Francisco, Ho, Shirley, et al.: Using the Marked Power Spectrum to Detect the Signature of Neutrinos in Large-Scale Structure. *Phys. Rev. Lett.* 126 (2021), Nr. 1, S. 1–5. *http://dx.doi.org/10.1103/PhysRevLett.126.011301* 121, 132, 146

[Massara et al. 2015] Massara, Elena, Villaescusa-Navarro, Francisco, Viel, Matteo, Sutter, P. M.: Voids in massive neutrino cosmologies. *J. Cosmol. Astropart. Phys.* 2015 (2015), Nr. 11. *http://dx.doi.org/10.1088/1475-7516/2015/11/018* 29, 121, 132

[McLachlan & Krishnan 1997] McLachlan, G.J., Krishnan, T.: *The EM algorithm and extensions.* Wiley, 1997. *http://dx.doi.org/10.1002/9780470191613.* *http://dx.doi.org/10.1002/9780470191613* 44, 47, 82

[Mecke et al. 1994] Mecke, K. R., Buchert, T., Wagner, H.: Robust morphological measures for large-scale structure in the Universe. *Astron. Astrophys.* 288 (1994), 697–704. *http://adsabs.harvard.edu/full/1994A{%}26A...288..697M* 28, 33

[Merlet & Zerubia 1996] Merlet, N., Zerubia, J.: New prospects in line detection by dynamic programming. *IEEE Trans. Pattern Anal. Mach. Intell.* 18 (1996), Nr. 4, 426–431. *http://dx.doi.org/10.1109/34.491623* 87

[Mezard & Montanari 2009] Mezard, Marc, Montanari, Andrea: *Information, Physics, and Computation.* Oxford University Press, Inc., 2009. *http://dx.doi.org/10.5555/1592967*. *http://dx.doi.org/10.5555/1592967* 41

[Mézard & Mora 2009] Mézard, Marc, Mora, Thierry: Constraint satisfaction problems and neural networks: A statistical physics perspective. *J. Physiol. Paris* 103 (2009), Nr. 1-2, S. 107–113. *http://dx.doi.org/10.1016/j.jphysparis.2009.05.013* 41

[Milgrom 1983] Milgrom, M: A modification of the Newtonian dynamics as a possible alternative to the hidden mass hypothesis. *Astrophys. J.* (1983), 270. *http://dx.doi.org/10.1086/161130* 19

[Mirza & Osindero 2014] Mirza, Mehdi, Osindero, Simon: Conditional Generative Adversarial Nets. *arXiv e-prints* (2014), 1–7. *http://arxiv.org/abs/1411.1784* 40

[Mo & White 1996] Mo, H. J., White, S. D.: An analytic model for the spatial clustering of dark matter haloes. *Mon. Not. R. Astron. Soc.* 282 (1996), Nr. 2, 347–361. *http://dx.doi.org/10.1093/mnras/282.2.347* 27

[Moccia et al. 2018] Moccia, Sara, Momi, Elena D., Hadji, Sara E., Mattos, Leonardo S.: Blood vessel segmentation algorithms – Review of methods , datasets and evaluation metrics. *Comput. Methods Programs Biomed.* 158 (2018), Nr. February, S. 71–91. *http://dx.doi.org/10.1016/j.cmpb.2018.02.001* 66

[More et al. 2011] More, Surhud, Kravtsov, Andrey V., Dalal, Neal, Gottlöber, Stefan: The overdensity and masses of the friends-of-friends halos and universality of halo mass function. *Astrophys. Journal, Suppl. Ser.* 195 (2011), Nr. 1. *http://dx.doi.org/10.1088/0067-0049/195/1/4* 23

[Mostoghiu et al. 2019] Mostoghiu, Robert, Knebe, Alexander, Cui, Weiguang, et al.: The Three Hundred Project: The evolution of galaxy cluster density profiles. *Mon. Not. R. Astron. Soc.* 483 (2019), Nr. 3, 3390–3403. *http://dx.doi.org/10.1093/mnras/sty3306* 115

[Mucesh et al. 2021] Mucesh, S, Hartley, W G., Palmese, A, et al.: A machine learning approach to galaxy properties: joint redshift–stellar mass probability distributions with Random Forest. *Mon. Not. R. Astron. Soc.* 502 (2021), Nr. 2, S. 2770–2786. *http://dx.doi.org/10.1093/mnras/stab164* 40

[Murtagh & Contreras 2012] Murtagh, Fionn, Contreras, Pedro: Algorithms for hierarchical clustering: an overview. *WIREs Data Min. Knowl. Discov.* 2 (2012), 86–97. *http://dx.doi.org/https://doi.org/10.1002/widm.53* 56

[Musso et al. 2018] Musso, M, Cadiou, C, Pichon, C, et al.: How does the cosmic web impact assembly bias ? *Mon. Not. R. Astron. Soc.* 476 (2018), Nr. 4, 4877–4906. *http://dx.doi.org/10.1093/mnras/sty191* 30

[Naidoo et al. 2020] Naidoo, Krishna, Whiteway, Lorne, Massara, Elena, et al.: Beyond two-point statistics : using the Minimum Spanning Tree as a tool for cosmology. *Mon. Not. R.*

Astron. Soc. 491 (2020), Nr. 2, 1709–1726. *http://dx.doi.org/10.1093/mnras/stz3075* 95, 117, 148

[Nascimento & De Carvalho 2011] Nascimento, Mariá C.V., De Carvalho, André C.: Spectral methods for graph clustering - A survey. *Eur. J. Oper. Res.* 211 (2011), Nr. 2, S. 221–231. *http://dx.doi.org/10.1016/j.ejor.2010.08.012* 69

[Nelson et al. 2019] Nelson, Dylan, Springel, Volker, Pillepich, Annalisa, et al.: The IllustrisTNG Simulations : Public Data Release. *Comput. Astrophys. Cosmol.* 6 (2019). *http://dx.doi.org/10.1186/s40668-019-0028-x* 5, 23, 108, 111

[Nicastro et al. 2018] Nicastro, F, Kaastra, J, Krongold, Y, et al.: Detection of the Missing Baryons in a Warm-Hot Intergalactic Medium. *Nature* 558 (2018), Nr. 7710, 406–409. *http://dx.doi.org/10.1038/s41586-018-0204-1* 36

[Nishimori 2001] Nishimori, Hidetoshi: *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, 2001. *http://dx.doi.org/10.1093/acprof:oso/9780198509417.001.0001*. *http://dx.doi.org/10.1093/acprof:oso/9780198509417.001.0001* 41

[Oemler 1974] Oemler, Augustus J.: The Systematic Properties of Clusters of Galaxies. Photometry of 15 Clusters. *Astrophys. J.* 194 (1974), 1–20. *http://dx.doi.org/10.1086/153216* 29

[Oliver et al. 1996] Oliver, Jonathan J., Baxter, Rohan A., Wallace, Chris S.: Unsupervised Learning Using MML. *Proc. 13th Int. Conf. Mach. Learn.* (1996), S. 364—-372 55

[Ozertem & Erdogmus 2011] Ozertem, Umut, Erdogmus, Deniz: Locally Defined Principal Curves and Surfaces. *J. Mach. Learn. Res.* 12 (2011), 1249–1286. *http://jmlr.org/papers/v12/ozertem11a.html* 32, 67, 78, 89

[Paranjape et al. 2018a] Paranjape, Aseem, Hahn, Oliver, Sheth, Ravi K.: Halo assembly bias and the tidal anisotropy of the local halo environment. *Mon. Not. R. Astron. Soc.* 476 (2018), Nr. 3, S. 3631–3647. *http://dx.doi.org/10.1093/mnras/sty496* 30

[Paranjape et al. 2018b] Paranjape, Aseem, Hahn, Oliver, Sheth, Ravi K.: The dependence of galaxy clustering on tidal environment in the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* 476 (2018), Nr. 4, S. 5442–5452. *http://dx.doi.org/10.1093/mnras/sty633* 27

[Park & Lee 2009] Park, Daeseong, Lee, Jounghun: The size distribution of void filaments in a ΛCDM cosmology. *Mon. Not. R. Astron. Soc.* 397 (2009), Nr. 4, 2163–2169. *http://dx.doi.org/10.1111/j.1365-2966.2009.15117.x* 95

[Parzen 1962] Parzen, Emanuel: On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* 33 (1962), 1065–1076. *http://dx.doi.org/10.1214/aoms/1177704472* 45

[Peacock 1998] Peacock, J. A.: *Cosmological Physics*. Cambridge University Press, 1998. *http://dx.doi.org/10.1017/CBO9780511804533*. *http://dx.doi.org/10.1017/CBO9780511804533* 17

[Pearce et al. 2001] Pearce, F. R., Jenkins, A., Frenk, C. S., et al.: Simulations of galaxy formation in a cosmological volume. *Mon. Not. R. Astron. Soc.* 326 (2001), Nr. 2, S. 649–666. *http://dx.doi.org/10.1046/j.1365-8711.2001.04616.x* 23

[Pearson & Coles 1995] Pearson, Russell C., Coles, Peter: Quantifying the geometry of large-scale structure. *Mon. Not. R. Astron. Soc.* 272 (1995), 231–240. *http://dx.doi.org/10.1093/mnras/272.1.231* 95

[Pedregosa et al. 2011] Pedregosa, Fabian, Weiss, Ron, Brucher, Matthieu: Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (2011), S. 2825–2830. *http://dx.doi.org/10.1007/s13398-014-0173-7.2* 48

[Peebles 1980] Peebles, P. J. E.: *The large-scale structure of the universe.* Princeton Univ. Press, 1980. – 448 S. *http://dx.doi.org/10.2307/j.ctvxrpz4n*. *http://dx.doi.org/10.2307/j.ctvxrpz4n* 11, 120

[Pereyra et al. 2020a] Pereyra, Luis A., Sgró, Mario A., Merchán, Manuel E., et al.: Detection and analysis of cluster-cluster filaments. *Mon. Not. R. Astron. Soc.* 499 (2020), S. 4876–4886. *http://dx.doi.org/10.1093/mnras/staa3112* 32, 71, 95

[Pereyra et al. 2020b] Pereyra, Luis A., Sgró, Mario A., Merchán, Manuel E., et al.: Semita: a novel cosmological filament finding algorithm. *Mon. Not. R. Astron. Soc.* 499 (2020), 4876–4886. *http://dx.doi.org/10.1093/mnras/staa3112* 32, 34, 95

[Peter et al. 2013] Peter, Annika H., Rocha, Miguel, Bullock, James S., Kaplinghat, Manoj: Cosmological simulations with self-interacting dark matter - II. Halo shapes versus observations. *Mon. Not. R. Astron. Soc.* 430 (2013), Nr. 1, S. 105–120. *http://dx.doi.org/10.1093/mnras/sts535* 30, 94

[Philcox 2021] Philcox, Oliver H.: A faster Fourier transform? Computing small-scale power spectra and bispectra for cosmological simulations in O(N2) time. *Mon. Not. R. Astron. Soc.* 501 (2021), Nr. 3, S. 4004–4034. *http://dx.doi.org/10.1093/mnras/staa3882* 6, 18, 28

[Philcox & Eisenstein 2020] Philcox, Oliver H., Eisenstein, Daniel J.: Computing the small-scale galaxy power spectrum and bispectrum in configuration space. *Mon. Not. R. Astron. Soc.* 492 (2020), Nr. 1, S. 1214–1242. *http://dx.doi.org/10.1093/mnras/stz3335* 6, 28

[Pichon et al. 2010] Pichon, Christophe, Gay, Christophe, Pogosyan, Dmitry, et al.: The skeleton: Connecting large scale structures to galaxy formation. In: *Invis. Universe* Bd. 1241, 2010. *http://dx.doi.org/10.1063/1.3462607*, 1108–1117 111

[Pisani et al. 2015] Pisani, Alice, Sutter, P. M., Hamaus, Nico, et al.: Counting voids to probe dark energy. *Phys. Rev. D - Part. Fields, Gravit. Cosmol.* 92 (2015), Nr. 8, S. 1–10. *http://dx.doi.org/10.1103/PhysRevD.92.083531* 29, 120, 132

[Planck Collaboration I et al. 2016] Planck Collaboration I, Adam, R., Ade, P. A., et al.: Planck 2015 results: I. Overview of products and scientific results. *Astron. Astrophys.* 594 (2016). *http://dx.doi.org/10.1051/0004-6361/201527101* 18

[Planck Collaboration IX et al. 2020] Planck Collaboration IX, Akrami, Y., Arroja, F., et al.: Planck 2018 results: IX. Constraints on primordial non-Gaussianity. *Astron. Astrophys.* 641 (2020), 1–50. *http://dx.doi.org/10.1051/0004-6361/201935891* 17

[Planck Collaboration VI et al. 2020] Planck Collaboration VI, Aghanim, N., Akrami, Y., et al.: Planck 2018 results: VI. Cosmological parameters. *Astron. Astrophys.* 641 (2020), A6. *http://dx.doi.org/10.1051/0004-6361/201833910* 19, 20, 122

[Planck Collaboration VII et al. 2020] Planck Collaboration VII, Akrami, Y., Ashdown, M., et al.: Planck 2018 results: VII. Isotropy and statistics of the CMB. *Astron. Astrophys.* 641 (2020). *http://dx.doi.org/10.1051/0004-6361/201935201* 17

[Planck Collaboration XIII et al. 2016] Planck Collaboration XIII, Ade, P. A., Aghanim, N., et al.: Planck 2015 results: XIII. Cosmological parameters. *Astron. Astrophys.* 594 (2016). *http://dx.doi.org/10.1051/0004-6361/201525830* 18, 19, 108

[Platen et al. 2007] Platen, Erwin, Van De Weygaert, Rien, Jones, Bernard J.: A cosmic watershed: The WVF void detection technique. *Mon. Not. R. Astron. Soc.* 380 (2007), Nr. 2, S. 551–570. *http://dx.doi.org/10.1111/j.1365-2966.2007.12125.x* 33

[Pogosyan et al. 2009] Pogosyan, D., Pichon, C., Gay, C., et al.: The local theory of the cosmic skeleton. *Mon. Not. R. Astron. Soc.* 396 (2009), Nr. 2, S. 635–667. *http://dx.doi.org/10.1111/j.1365-2966.2009.14753.x* 32

[Poudel et al. 2017] Poudel, A, Heinämäki, P, Tempel, E, et al.: The effect of cosmic web filaments on the properties of groups and their central galaxies. *Astron. Astrophys.* 597 (2017), A86. *http://dx.doi.org/10.1051/0004-6361/201629639* 30

[Power et al. 2012] Power, Chris, Knebe, Alexander, Knollmann, Steffen R.: The dynamical state of dark matter haloes in cosmological simulations - I. Correlations with mass assembly history. *Mon. Not. R. Astron. Soc.* 419 (2012), Nr. 2, 1576–1587. *http://dx.doi.org/10.1111/j.1365-2966.2011.19820.x* 114, 115

[Radford et al. 2016] Radford, Alec, Metz, Luke, Chintala, Soumith: Unsupervised representation learning with deep convolutional generative adversarial networks. *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.* (2016), 1–16. *https://arxiv.org/abs/1511.06434* 40

[Ramachandra & Shandarin 2015] Ramachandra, Nesar S., Shandarin, Sergei F.: Multi-stream portrait of the cosmic web. *Mon. Not. R. Astron. Soc.* 452 (2015), Nr. 2, S. 1643–1653. *http://dx.doi.org/10.1093/mnras/stv1389* 33, 34

[Ramakrishnan et al. 2019] Ramakrishnan, Sujatha, Paranjape, Aseem, Hahn, Oliver, Sheth, Ravi K.: Cosmic web anisotropy is the primary indicator of halo assembly bias. *Mon. Not. R. Astron. Soc.* 489 (2019), Nr. 3, 2977–2996. *http://dx.doi.org/10.1093/mnras/stz2344* 30

[Ravanbakhsh et al. 2017] Ravanbakhsh, Siamak, Oliva, Junier, Fromenteau, Sebastien, et al.: Estimating cosmological parameters from the dark matter distribution. *arXiv e-prints* (2017). *https://arxiv.org/abs/1711.02033* 40

[Reid et al. 2016] Reid, Beth, Ho, Shirley, Padmanabhan, Nikhil, et al.: SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: Galaxy target selection and large-scale structure catalogues. *Mon. Not. R. Astron. Soc.* 455 (2016), Nr. 2, S. 1553–1573. *http://dx.doi.org/10.1093/mnras/stv2382* 19

[Ribli et al. 2019] Ribli, Dezső, Pataki, Bálint Ármin, Csabai, István: An improved cosmological parameter inference scheme motivated by deep learning. *Nat. Astron.* 3 (2019), Nr. 1, 93–98. *http://dx.doi.org/10.1038/s41550-018-0596-8* 6, 40

[Robertson 1935] Robertson, H. P.: Kinematics and World-Structure. *Astrophys. J.* 82 (1935), 284. *http://dx.doi.org/10.1086/143681* 12

[Rodríguez et al. 2018] Rodríguez, Andres C., Kacprzak, Tomasz, Lucchi, Aurelien, et al.: Fast cosmic web simulations with generative adversarial networks. *Comput. Astrophys. Cosmol.* 5 (2018), Nr. 1. *http://dx.doi.org/10.1186/s40668-018-0026-4* 40

[Roeder & Wasserman 1997] Roeder, Kathryn, Wasserman, Larry: Practical Bayesian Density Estimation Using Mixtures of Normals. *J. Am. Stat. Assoc.* 439 (1997), 894—-902. *http://www.jstor.org/stable/2965553* 55

[Rose et al. 1990] Rose, Kenneth, Gurewitz, Eitan, Fox, Geoffrey C.: Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* 65 (1990), Nr. 8, 945–948. *http://dx.doi.org/10.1103/PhysRevLett.65.945* 49, 51, 52

[Rosenblatt 1959] Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65 (1959), S. 386–408. *http://dx.doi.org/10.1037/h0042519* 40

[Rost et al. 2020] Rost, Agustin, Stasyszyn, Federico, Pereyra, Luis, Martinez, Hector J.: A comparison of cosmological filaments catalogues. *Mon. Not. R. Astron. Soc.* 493 (2020), Nr. 2, S. 1936–1947. *http://dx.doi.org/10.1093/mnras/staa320* 94, 108, 117

[Roweis & Saul 2000] Roweis, Sam T., Saul, Lawrence K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science (80-.).* 290 (2000), Nr. December, 2323–2326. *https://science.sciencemag.org/content/290/5500/2323* 66

[Rumelhart et al. 1986] Rumelhart, D. E., Hinton, G. E., Williams, R. J.: Learning Internal Representations by Error Propagation. Version: 1986. *http://dx.doi.org/10.5555/104279.104293*. In: *Parallel Distrib. Process. Explor. Microstruct. Cogn. Vol. 1 Found.* 1986, S. 318–362 40

[Rykoff et al. 2014] Rykoff, E. S., Rozo, E., Busha, M. T., et al.: RedMaPPer. I. Algorithm and SDSS DR8 catalog. *Astrophys. J.* 785 (2014), Nr. 2. *http://dx.doi.org/10.1088/0004-637X/785/2/104* 94

[Saade et al. 2014] Saade, Alaa, Krzakala, Florent, Zdeborová, Lenka: Spectral clustering of graphs with the Bethe Hessian. *Adv. Neural Inf. Process. Syst.* 27 (2014), Nr. January, 406–414. *https://proceedings.neurips.cc/paper/2014/file/63923f49e5241343aa7acb6a06a751e7-Paper.pdf* 69

[Salvati et al. 2018] Salvati, Laura, Douspis, Marian, Aghanim, Nabila: Constraints from thermal Sunyaev-Zel'dovich cluster counts and power spectrum combined with CMB. *Astron. Astrophys.* 614 (2018), S. 1–11. *http://dx.doi.org/10.1051/0004-6361/201731990* 19, 120

[Sarron et al. 2019] Sarron, F, Adami, C, Durret, F, Laigle, C: Pre-processing of galaxies in cosmic filaments around AMASCFI clusters in the CFHTLS. *Astron. Astrophys.* (2019), S. A49. *http://dx.doi.org/10.1051/0004-6361/201935394* 30, 111, 112

[Saxena et al. 2017] Saxena, Amit, Prasad, Mukesh, Gupta, Akshansh, et al.: A review of clustering techniques and developments. *Neurocomputing* 267 (2017), 664–681. *http://dx.doi.org/10.1016/j.neucom.2017.06.053* 44

[Schaap & Weygaert 2000] Schaap, W.E., Weygaert, R.: Continuous fields and discrete samples: reconstruction through Delaunay tessellations. *Astron. Astrophys.* 363 (2000), L29–L32. *http://adsabs.harvard.edu/full/2000A{%}26A...363L..29S* 36

[Schaye et al. 2015] Schaye, Joop, Crain, Robert A., Bower, Richard G., et al.: The EAGLE project: Simulating the evolution and assembly of galaxies and their environments. *Mon.*

Not. R. Astron. Soc. 446 (2015), Nr. 1, S. 521–554. *http://dx.doi.org/10.1093/mnras/stu2058* 23, 31, 106, 108

[Schmittfull et al. 2017] Schmittfull, Marcel, Baldauf, Tobias, Zaldarriaga, Matias: Iterative initial condition reconstruction. *Phys. Rev. D* 96 (2017), Nr. 2, S. 1–26. *http://dx.doi.org/10.1103/PhysRevD.96.023505* 149

[Schwarz 1978] Schwarz, Gideon: Estimating the Dimension of a Model. *Ann. Stat.* 6 (1978), 461–464. *http://dx.doi.org/10.1214/aos/1176344136* 55

[Sefusatti et al. 2016] Sefusatti, E., Crocce, M., Scoccimarro, R., Couchman, H. M.: Accurate estimators of correlation functions in Fourier space. *Mon. Not. R. Astron. Soc.* 460 (2016), Nr. 4, 3624–3636. *http://dx.doi.org/10.1093/mnras/stw1229* 123

[Sereno et al. 2018] Sereno, Mauro, Umetsu, Keiichi, Ettori, Stefano, et al.: Clump-3d. testing λcdm with galaxy cluster shapes. *Astrophys. J.* 869 (2018), Nr. 1, L4. *http://dx.doi.org/10.3847/2041-8213/aac6d9* 30

[Shamir 2009] Shamir, Lior: Automatic morphological classification of galaxy images. *Mon. Not. R. Astron. Soc.* 23 (2009), Nr. 1, 1367–1372. *http://dx.doi.org/10.1111/j.1365-2966.2009.15366.x* 63

[Shannon 1948] Shannon, C. E.: A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27 (1948), Nr. 4, 623–656. *http://dx.doi.org/10.1002/j.1538-7305.1948.tb00917.x* 49

[Sheth et al. 2005] Sheth, Ravi K., Connolly, Andrew J., Skibba, Ramin: Marked correlations in galaxy formation models. *arXiv e-prints* 13 (2005), Nr. July, 1–13. *http://arxiv.org/abs/astro-ph/0511773* 121

[Shirasaki & Yoshida 2014] Shirasaki, Masato, Yoshida, Naoki: Statistical and systematic errors in the measurement of weak-lensing minkowski functionals: Application to the Canada-france-hawaii lensing survey. *Astrophys. J.* 786 (2014), Nr. 1. *http://dx.doi.org/10.1088/0004-637X/786/1/43* 28

[Silver et al. 2016] Silver, David, Huang, Aja, Maddison, Chris J., et al.: Mastering the game of Go with deep neural networks and tree search. *Nature* 529 (2016), Nr. 7587, 484–489. *http://dx.doi.org/10.1038/nature16961* 40

[Silverman 1986] Silverman, B.W.: *Density estimation for statistics and data analysis.* Chapman and Hall, 1986 *https://cds.cern.ch/record/1070306* 87

[Singer 2006] Singer, A.: From graph to manifold Laplacian: The convergence rate. *Appl. Comput. Harmon. Anal.* 21 (2006), Nr. 1, S. 128–134. *http://dx.doi.org/10.1016/j.acha.2006.03.004* 74

[Slezak et al. 1993] Slezak, E., Lapparent, V. de, Bijaoui, A: Objective Detection of Voids and High-Density Structures in the First CfA Redshift Survey Slice. *Astrophys. J.* 409 (1993), 517. *http://dx.doi.org/10.1086/172683* 28

[Smola et al. 2001] Smola, Alexander J., Mika, Sebastian, Sch, Bernhard, Williamson, Robert C.: Regularized Principal Manifolds. *J. Mach. Learn. Res.* 1 (2001), S. 179–209 67

[Smolensky 1986] Smolensky, P.: Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory. Version: 1986. *https://stanford.edu/{~}jlmcc/papers/PDP/Volume1/Chap6{_}PDP86.pdf*. In: *Parallel Distrib. Process. Explor. Microstruct. Cogn. Vol. 1 Found.* MIT Press, 1986, 194–281 40, 62

[Smoot et al. 1992] Smoot, G.F., Bennett, C.L., Kogut, A., et al.: Structure in the COBE Differential Microwave Radiometer First-Year Maps. *Astrophys. J. Lett.* 396 (1992), L1. *http://dx.doi.org/10.1086/186504* 12, 18

[Sorce et al. 2016] Sorce, Jenny G., Gottlöber, Stefan, Yepes, Gustavo, et al.: Cosmicflows Constrained Local UniversE Simulations. *Mon. Not. R. Astron. Soc.* 455 (2016), Nr. 2, 2078–2090. *http://dx.doi.org/10.1093/mnras/stv2407* 22

[Sousbie 2011] Sousbie, T.: The persistent cosmic web and its filamentary structure - I. Theory and implementation. *Mon. Not. R. Astron. Soc.* 414 (2011), Nr. 1, S. 350–383. *http://dx.doi.org/10.1111/j.1365-2966.2011.18394.x* 33, 34

[Sousbie et al. 2008] Sousbie, T., Pichon, C., Colombi, S., et al.: The 3D skeleton: Tracing the filamentary structure of the Universe. *Mon. Not. R. Astron. Soc.* 383 (2008), Nr. 4, S. 1655–1670. *http://dx.doi.org/10.1111/j.1365-2966.2007.12685.x* 32

[Sousbie et al. 2011] Sousbie, T., Pichon, C., Kawahara, H.: The persistent cosmic web and its filamentary structure - II. Illustrations. *Mon. Not. R. Astron. Soc.* 414 (2011), Nr. 1, S. 384–403. *http://dx.doi.org/10.1111/j.1365-2966.2011.18395.x* 34

[Spigler et al. 2019] Spigler, S., Geiger, M., D'Ascoli, S., et al.: A jamming transition from under- To over-parametrization affects generalization in deep learning. *J. Phys. A Math. Theor.* 52 (2019), Nr. 47. *http://dx.doi.org/10.1088/1751-8121/ab4c8b* 41

[Springel et al. 2008] Springel, V., Wang, J., Vogelsberger, M., et al.: The Aquarius Project: The subhaloes of galactic haloes. *Mon. Not. R. Astron. Soc.* 391 (2008), Nr. 4, S. 1685–1711. *http://dx.doi.org/10.1111/j.1365-2966.2008.14066.x* 97

[Springel 2005] Springel, Volker: The cosmological simulation code GADGET-2. *Mon. Not. R. Astron. Soc.* 364 (2005), Nr. 4, S. 1105–1134. *http://dx.doi.org/10.1111/j.1365-2966.2005.09655.x* 22, 122

[Springel et al. 2005] Springel, Volker, White, Simon D. M., Jenkins, Adrian, et al.: Simulating the joint evolution of quasars, galaxies and their large-scale distribution. *Nature* 435 (2005), Nr. 7042, 629–636. *http://dx.doi.org/10.1038/nature03597* 5

[Springel et al. 2001] Springel, Volker, White, Simon D. M., Tormen, Giuseppe, Kauffmann, Guinevere: Populating a cluster of galaxies – I. Results at z = 0. *Mon. Not. R. Astron. Soc.* 328 (2001), Nr. 3, 726–750. *http://dx.doi.org/10.1046/j.1365-8711.2001.04912.x* 63, 108, 111

[Stanford & Raftery 2000] Stanford, Derek C., Raftery, Adrian E.: Finding Curvilinear Features in Spatial Point Patterns : Principal Curve Clustering with Noise. *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2000), Nr. 6, 601–609. *http://dx.doi.org/10.1109/34.862198* 89

[Starck et al. 2010] Starck, J., Murtagh, F., Fadili, J.: *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge University Press, 2010. *http://dx.doi.org/10.1017/CBO9780511730344*. *http://dx.doi.org/10.1017/CBO9780511730344* 42

[Starck et al. 2005] Starck, J. L., Martínez, V. J., Donoho, D. L., et al.: Analysis of the spatial distribution of galaxies by multiscale methods. *EURASIP J. Appl. Signal Processing* 2005 (2005), Nr. 15, S. 2455–2469. *http://dx.doi.org/10.1155/ASP.2005.2455* 28

[Stoica et al. 2005a] Stoica, R. S., Gregori, P., Mateu, J.: Simulated annealing and object point processes: Tools for analysis of spatial patterns. *Stoch. Process. their Appl.* 115 (2005), Nr. 11, 1860–1882. *http://dx.doi.org/10.1016/j.spa.2005.06.007* 32, 51

[Stoica et al. 2005b] Stoica, R. S., Martínez, V. J., Mateu, J., Saar, E.: Detection of cosmic filaments using the Candy model. *Astron. Astrophys.* 434 (2005), Nr. 2, S. 423–432. *http://dx.doi.org/10.1051/0004-6361:20042409* 51

[Stoica et al. 2004] Stoica, Radu, Descombes, Xavier, Zerubia, Josiane: A Gibbs Point Process for Road Extraction from Remotely Sensed Images. *Int. J. Comput. Vis.* 57 (2004), Nr. 2, S. 121–136. *http://dx.doi.org/10.1023/B:VISI.0000013086.45688.5d* 51, 87

[Stoica et al. 2007] Stoica, Radu S., Martínez, Vicent J., Saar, Enn: A three-dimensional object point process for detection of cosmic filaments. *J. R. Stat. Soc. Ser. C Appl. Stat.* 56 (2007), Nr. 4, 459–477. *http://dx.doi.org/10.1111/j.1467-9876.2007.00587.x* 32, 34, 100

[Sutton & Barto 2018] Sutton, Richard S., Barto, Andrew G.: *An introduction to reinforcement learning*. Second. 2018. *http://dx.doi.org/10.4018/978-1-60960-165-2.ch004*. *http://dx.doi.org/10.4018/978-1-60960-165-2.ch004* 40

[Takahashi et al. 2010] Takahashi, Ryuichi, Yoshida, Naoki, Takada, Masahiro, et al.: Non-Gaussian error contribution to likelihood analysis of the matter power spectrum. *Astrophys. J.* 726 (2010), Nr. 1. *http://dx.doi.org/10.1088/0004-637x/726/1/7* 132

[Tanimura et al. 2020a] Tanimura, H, Aghanim, N, Bonjean, V, et al.: Density and temperature of cosmic-web filaments on scales of tens of megaparsecs. *Astron. Astrophys.* 637 (2020), A41. *http://dx.doi.org/10.1051/0004-6361/201937158* 30, 117

[Tanimura et al. 2020b] Tanimura, H., Aghanim, N., Kolodzig, A., et al.: First detection of stacked X-ray emission from cosmic web filaments. *Astron. Astrophys.* 643 (2020), 1–7. *http://dx.doi.org/10.1051/0004-6361/202038521* 36, 117

[Tanimura et al. 2019] Tanimura, Hideki, Hinshaw, Gary, McCarthy, Ian G., et al.: A Search for Warm/Hot Gas Filaments Between Pairs of SDSS Luminous Red Galaxies. *Mon. Not. R. Astron. Soc.* 483 (2019), 223–234. *http://arxiv.org/abs/1709.05024* 36

[Tanimura et al. 2020c] Tanimura, Hideki, Hinshaw, Gary, McCarthy, Ian G., et al.: Probing hot gas around luminous red galaxies through the Sunyaev-Zel'dovich effect. *Mon. Not. R. Astron. Soc.* 491 (2020), Nr. 2, S. 2318–2329. *http://dx.doi.org/10.1093/mnras/stz3130* 36, 117

[Tempel et al. 2016] Tempel, E., Stoica, R. S., Kipper, R., Saar, E.: Bisous model-Detecting filamentary patterns in point processes. *Astron. Comput.* 16 (2016), S. 17–25. *http://dx.doi.org/10.1016/j.ascom.2016.03.004* 32, 34, 51

[Tempel et al. 2014] Tempel, E., Stoica, R. S., Martínez, V. J., et al.: Detecting filamentary pattern in the cosmic web: A catalogue of filaments for the SDSS. *Mon. Not. R. Astron. Soc.* 438 (2014), Nr. 4, S. 3465–3482. *http://dx.doi.org/10.1093/mnras/stt2454* 32, 36, 94

[Tibshirani 1992] Tibshirani, Robert: Principal curves revisited. *Stat. Comput.* 2 (1992), Nr. 4, S. 183–190. *http://dx.doi.org/10.1007/BF01889678* 67, 74

[Tibshirani 1996] Tibshirani, Robert: Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* 58 (1996), Nr. 1, 267–288. *http://dx.doi.org/10.2307/2346178* 42

[Tikhonov & Arsenin 1977] Tikhonov, Andrey N., Arsenin, Vasiliy Y.: *Solutions of ill-posed problems*. V. H. Winston & Sons, 1977 *https://www.bibsonomy.org/bibtex/2940dd9e8193fd295da5911df36c24851/peter.ralph* 42

[Tinker et al. 2008] Tinker, Jeremy, Kravtsov, Andrey V., Klypin, Anatoly, et al.: Toward a Halo Mass Function for Precision Cosmology: The Limits of Universality. *Astrophys. J.* 688 (2008), Nr. 2, 709–728. *http://dx.doi.org/10.1086/591439* 23

[Tubiana & Monasson 2017] Tubiana, J., Monasson, R.: Emergence of Compositional Representations in Restricted Boltzmann Machines. *Phys. Rev. Lett.* 118 (2017), Nr. 13. *http://dx.doi.org/10.1103/PhysRevLett.118.138301* 41

[Ueda & Nakano 1998] Ueda, Naonori, Nakano, Ryohei: Deterministic Annealing EM Algorithm. *Neural Networks* 11 (1998), 271–282. *http://dx.doi.org/10.1016/s0893-6080(97)00133-0* 48, 51, 52, 87

[Ullmo et al. 2021] Ullmo, M., Decelle, A., Aghanim, N.: Encoding large-scale cosmological structure with generative adversarial networks. *Astron. Astrophys.* (2021). *https://arxiv.org/abs/2011.05244* 40

[Vallés-Pérez et al. 2020] Vallés-Pérez, David, Planelles, Susana, Quilis, Vicent: On the accretion history of galaxy clusters: Temporal and spatial distribution. *Mon. Not. R. Astron. Soc.* 499 (2020), Nr. 2, 2303–2318. *http://dx.doi.org/10.1093/mnras/staa3035* 115

[Valogiannis & Bean 2018] Valogiannis, Georgios, Bean, Rachel: Beyond δ: Tailoring marked statistics to reveal modified gravity. *Phys. Rev. D* 97 (2018), Nr. 2, 023535. *http://dx.doi.org/10.1103/PhysRevD.97.023535* 121

[Van Der Maaten et al. 2009] Van Der Maaten, Laurens, Postma, Eric, Van den Herik, Jaap: Dimensionality reduction: a comparative review. *J. Mach. Learn. Res.* 10 (2009), 66–71. *https://www.bibsonomy.org/bibtex/2ed03568f0e9bca9cdaf6b25304e55940/peter.ralph* 66

[Van Nieuwenburg et al. 2017] Van Nieuwenburg, Evert P., Liu, Ye H., Huber, Sebastian D.: Learning phase transitions by confusion. *Nat. Phys.* 13 (2017), Nr. 5, S. 435–439. *http://dx.doi.org/10.1038/nphys4037* 62

[Vapnik 1998] Vapnik, Vladimir: *The Nature of Statistical Learning Theory*. 2. Springer, New York, 1998. – 334 S. *http://dx.doi.org/10.1007/978-1-4757-3264-1*. *http://dx.doi.org/10.1007/978-1-4757-3264-1* 41

[Varma & Zisserman 2009] Varma, Manik, Zisserman, Andrew: A Statistical Approach to Material Classification Using Image Patch Exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2009), Nr. 11, S. 2032–2047. *http://dx.doi.org/10.1109/TPAMI.2008.182* 16

[Villaescusa-Navarro et al. 2013] Villaescusa-Navarro, F., Vogelsberger, M., Viel, M., Loeb, A.: Neutrino signatures on the high-transmission regions of the Lyman α forest. *Mon. Not. R. Astron. Soc.* 431 (2013), Nr. 4, 3670–3677. *http://dx.doi.org/10.1093/mnras/stt452* 121

[Villaescusa-Navarro et al. 2018] Villaescusa-Navarro, Francisco, Banerjee, Arka, Dalal, Neal, et al.: The imprint of neutrinos on clustering in redshift-space. *Astrophys. J.* 861 (2018), Nr. 1, 53. *http://dx.doi.org/10.3847/1538-4357/aac6bf* 136

[Villaescusa-Navarro et al. 2020] Villaescusa-Navarro, Francisco, Hahn, Chang H., Massara, Elena, et al.: The Quijote simulations. *Astrophys. J.* 250 (2020), 2. *http://dx.doi.org/10.3847/1538-4365/ab9d82* 5, 22, 23, 122, 142, 149

[Villaescusa-Navarro et al. 2014] Villaescusa-Navarro, Francisco, Marulli, Federico, Viel, Matteo, et al.: Cosmology with massive neutrinos I: Towards a realistic modeling of the relation between matter, haloes and galaxies. *J. Cosmol. Astropart. Phys.* 2014 (2014), Nr. 3. *http://dx.doi.org/10.1088/1475-7516/2014/03/011* 120

[Vogelsberger et al. 2014] Vogelsberger, Mark, Genel, Shy, Springel, Volker, et al.: Introducing the illustris project: Simulating the coevolution of dark and visible matter in the universe. *Mon. Not. R. Astron. Soc.* 444 (2014), Nr. 2, S. 1518–1547. *http://dx.doi.org/10.1093/mnras/stu1536* 5, 22, 23, 97

[Walker 1937] Walker, A. G.: On Milne's Theory of World-Structure. *Proc. London Math. Soc.* 42 (1937), 90–127. *http://dx.doi.org/10.1112/plms/s2-42.1.90* 12

[Wang et al. 2016] Wang, Weixing, Yang, Nan, Zhang, Yi, et al.: A review of road extraction from remote sensing images. *J. Traffic Transp. Eng.* 3 (2016), Nr. 3, 271–282. *http://dx.doi.org/10.1016/j.jtte.2016.05.005* 87

[Weinmann et al. 2006] Weinmann, Simone M., Van Den Bosch, Frank C., Yang, Xiaohu, Mo, H. J.: Properties of galaxy groups in the Sloan Digital Sky Survey - I. The dependence of colour, star formation and morphology on halo mass. *Mon. Not. R. Astron. Soc.* 366 (2006), Nr. 1, S. 2–28. *http://dx.doi.org/10.1111/j.1365-2966.2005.09865.x* 62

[Wetzel 2017] Wetzel, Sebastian J.: Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E* 96 (2017), Nr. 2, S. 022140. *http://dx.doi.org/10.1103/PhysRevE.96.022140* 62

[van de Weygaert & Platen 2011] Weygaert, Rien van d., Platen, Erwin: Cosmic Voids: Structure, Dynamics and Galaxies. *Int. J. Mod. Phys. Conf. Ser.* 1 (2011), 41–66. *http://dx.doi.org/10.1142/S2010194511000092* 120

[White 2016] White, M: A marked correlation function for constraining modified gravity models. *J. Cosmol. Astropart. Phys.* 11 (2016), 057. *http://dx.doi.org/10.1088/1475-7516/2016/11/057* 121, 132, 145

[White & Frenk 1991] White, Simon D., Frenk, Carlos S.: Galaxy Formation through Hierarchical Clustering. *Astrophys. J.* 379 (1991), Nr. sep, 52–79. *http://dx.doi.org/10.1086/170483* 29, 120

[White et al. 1993] White, Simon D., Navarro, Julio F., Evrard, August E., Frenk, Carlos S.: The baryon content of galaxy clusters: a challenge to cosmological orthodoxy. *Nature* 366 (1993), Nr. 6454, 429–433. *http://dx.doi.org/10.1038/366429a0* 29

[Wiener 1930] Wiener, Norbert: Generalized Harmonic Analysis. *Acta Math.* 55 (1930), 117–258. *http://dx.doi.org/10.1007/BF02546511* 17

[Wilding et al. 2020] Wilding, Georg, Nevenzeel, Keimpe, Weygaert, Rien Van D., et al.: Persistent homology of the cosmic web . I : Hierarchical topology in Λ CDM cosmologies. *arXiv e-prints* 28 (2020), Nr. November, 1–28. *https://arxiv.org/abs/2011.12851* 29

[Woiselle et al. 2010] Woiselle, A., Starck, J. L., Fadili, J.: 3D curvelet transforms and astronomical data restoration. *Appl. Comput. Harmon. Anal.* 28 (2010), Nr. 2, 171–188. *http://dx.doi.org/10.1016/j.acha.2009.12.003* 28

[Wu 1983] Wu, C. F. J.: On the Convergence Properties of the EM Algorithm. *Ann. Stat.* 11 (1983), 95—-103. *http://dx.doi.org/10.1214/aos/1176346060* 47, 48

[Yadav et al. 2005] Yadav, Jaswant, Bharadwaj, Somnath, Pandey, Biswajit, Seshadri, T. R.: Testing homogeneity on large scales in the Sloan Digital Sky Survey Data Release One. *Mon. Not. R. Astron. Soc.* 364 (2005), Nr. 2, 601–606. *http://dx.doi.org/10.1111/j.1365-2966.2005.09578.x* 13

[Yang et al. 2017] Yang, Xiaohu, Zhang, Youcai, Luo, Wentao, et al.: Revealing the Cosmic Web-dependent Halo Bias. *Astrophys. J.* 848 (2017), 60. *http://dx.doi.org/10.3847/1538-4357/aa8c7a* 27

[York et al. 2000] York, D. G., Adelman, J., Anderson, J.E., et al.: The Sloan Digital Sky Survey: Technical Summary. *Astrophys. J.* 120 (2000), 1579–1587. *http://dx.doi.org/10.1086/301513* 5, 23, 94, 117

[Yoshida et al. 2000] Yoshida, Naoki, Springel, Volker, White, Simon D., Tormen, Giuseppe: Weakly Self-interacting Dark Matter and the Structure of Dark Halos. *Astrophys. J.* 544 (2000), Nr. 2, 87–90. *http://dx.doi.org/10.1086/317306* 30, 94

[Yuille 1990] Yuille, Alan L.: Generalized Deformable Models, Statistical Physics, and Matching Problems. *Neural Comput.* 2 (1990), S. 1–24. *http://dx.doi.org/10.1162/neco.1990.2.1.1* 74

[Zdeborova & Krzakala 2016] Zdeborova, Lenka, Krzakala, Florent: Statistical physics of inference: Thresholds and algorithms. *Adv. Phys.* 65 (2016), S. 453–552. *http://dx.doi.org/10.1080/00018732.2016.1211393* 41

[Zel'dovich 1970] Zel'dovich: Gravitational instability: an approximate theory for large density perturbations. *Astron. Astrophys.* 500 (1970), 13–18. *https://ui.adsabs.harvard.edu/abs/1970A{&}A.....5...84Z* 5, 15, 21, 123

[Zhang et al. 2016] Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, et al.: Understanding deep learning requires rethinking generalization. *Commun. ACM* 64 (2016), Nr. 3, S. 107–115. *http://dx.doi.org/10.1145/3446776* 41

[Zhang et al. 2019] Zhang, Xinyue, Wang, Yanfang, Zhang, Wei, et al.: From Dark Matter to Galaxies with Convolutional Neural Networks. *arXiv e-prints* (2019). *https://arxiv.org/abs/1902.05965* 40

université PARIS-SACLAY

ÉCOLE DOCTORALE
Astronomie
et Astrophysique
d'Île-de-France (AAIF)

Titre : Les environnements de la toile cosmique : identification, caractérisation et quantification de l'information cosmologique

Mots clés : Cosmologie: Structures grandes échelles de l'Univers, Toile cosmique; Méthodes: Méthodes statistiques, Analyse de motifs, Apprentissage automatique non supervisé, Modèles de mélange.

Résumé : La distribution de matière dans l'Univers se présente sous une structure complexe que l'on appelle la toile cosmique. Dans cette disposition spatiale, des régions denses, les nœuds de la toile cosmique, sont reliés par des ponts de matière, les filaments, qui se trouvent à l'intersection de structures planaires moyennement denses appelées murs qui définissent eux-mêmes les bords de vastes régions vides. Cette distribution, façonnée par la gravité depuis des milliards d'années, contient de précieuses informations sur le modèle cosmologique sous-jacent mais également sur les conditions initiales de l'Univers et son évolution. La détection et l'étude des éléments de la toile cosmique, qui jouent également un rôle fondamental dans la formation et l'évolution des galaxies, constituent de véritables défis nécessitant la conception d'outils sophistiqués pour traiter la complexité des structures multi-échelles qui la compose.

Avec pour ambition d'identifier et de caractériser les différents environnements, cette thèse propose plusieurs approches pour analyser des jeux de données spatialement organisés au moyen de méthodes d'apprentissage non supervisé fondées sur les modèles de mélanges. En particulier, des principes dérivés de la physique statistique sont utilisés pour mieux appréhender et comprendre la dynamique d'apprentissage d'un algorithme de classification non supervisé. Nous exposons comment utiliser ce parallèle avec la physique statistique afin d'explorer le jeu de données et obtenir des informations sur sa structure. Afin d'identifier la structure filamentaire de la toile cosmique, nous construisons ensuite une version régularisée de la procédure de classification pour apprendre itérativement une représentation du jeu de données, que l'on suppose généré par une structure uni-dimensionnelle sous-jacente. La méthode modélise cette structure latente par un graphe qui est intégré comme un a priori dans la formulation Bayésienne du problème menant à l'estimation d'un graphe principal passant au centre de la distribution de matière tracée par les galaxies. Nous montrons que cette formulation est particulièrement adaptée à la description des filaments cosmiques puisqu'elle permet la description de leurs propriétés géométriques (longueurs, épaisseurs, etc.) ainsi que l'association, pour les traceurs (galaxies, halos), d'une probabilité d'appartenir à un filament donné. L'algorithme proposé dans la thèse est appliqué avec succès à des simulations numériques. Ces applications ont notamment permis l'étude des relations entre la connectivité des amas de galaxies dans la toile cosmique et leurs propriétés dynamiques et morphologiques. Enfin, nous réalisons, à partir d'un ensemble de simulations à N-corps, une étude approfondie de l'information cosmologique contenue dans les environnements de la toile cosmique (nœuds, filaments, murs et vides). Il est notamment montré que l'analyse des environnements permet de lever les dégénérescences entre certains des paramètres du modèle faisant de la toile cosmique une sonde alternative permettant d'améliorer significativement les contraintes sur les paramètres cosmologiques vis-à-vis des analyses conventionnelles.

Title: Cosmic web environments: identification, characterisation, and quantification of cosmological information

Keywords: Cosmology: Large-scale structure of Universe, Cosmic Web; Methodology: Statistical Methods, Pattern analysis, Unsupervised Machine Learning, Mixture models.

Abstract: The late-time matter distribution depicts a complex pattern commonly called the cosmic web. In this picture, the spatial arrangement of matter is that of dense anchors, the nodes, linked together by elongated bridges of matter, the filaments, found at the intersection of thin mildly-dense walls, themselves surrounding large empty voids. This distribution, shaped by gravitational forces since billions of years, carries crucial information on the underlying cosmological model and on the evolution of the large-scale structures. Detecting and studying elements of cosmic web, playing also a key role in the formation and evolution of galaxies, are challenging tasks requiring the elaboration of optimised methods to handle the intrinsic complexity of the pattern made of multi-scale structures of various shapes and densities.

With the aim of identifying and characterising the cosmic web environments, we propose several approaches to analyse spatially structured point-cloud datasets, not restricted to cosmological ones, by means of unsupervised machine learning methods based on mixture models. In particular, we use principles emanating from statistical physics to get a better understanding of the learning dynamics of a clustering algorithm and expose how statistical physics can be used to explore the data distribution and obtain key insights on its structure. In order to identify the filamentary part of the pattern, its most prominent feature, we propose a regularisation of the clustering procedure to iteratively learn a non-linear representation of structured datasets, assuming it was generated by an underlying one-dimensional manifold. The method models this latent structure as a graph embedded as a prior in the Bayesian formulation of the problem to estimate a principal graph passing in the ridges of the matter distribution as traced by galaxies or halos. We show that this formulation is especially well-suited for the description of the filaments since it allows the description of their geometrical properties (lengths, widths, etc.) and associates to each tracer a probability of residing in a given filament. The resulting algorithm is successfully used to detect filaments in state-of-the-art numerical simulations. It also allows us to study the relation between the connectivity of galaxy clusters to the cosmic web and their dynamical and morphological properties. Finally, based on a large suite of N-body simulations, we perform a comprehensive analysis of the cosmological information content based on the two-point statistics derived in the cosmic web environments (nodes, filaments, walls and voids). We show that they can break some degeneracies among key parameters of the model making them a suitable alternative probe to significantly improve the constraints on cosmological parameters obtained by standard analyses.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France