Aix*Marseille université
Socialement engagée

NNT : 2023AIXMXXXX

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 15 octobre 2024 par

Tyann DUMERCHAT

From linear to non-linear scales : galaxy and peculiar velocity clustering analysis to constrain Dark Energy

Discipline
Physique et Sciences de la Matière

Spécialité
Astrophysique et Cosmologie

École doctorale
ED352

Laboratoire/Partenaires de recherche
Centre de Physique des Particules de Marseille
Chaire d'excellence AMIDEX

Composition du jury

Étienne BURTIN CEA, Paris-Saclay	Rapporteur
Ariel SANCHEZ MPE, Munich	Rapporteur
Tamara DAVIS UQ, Brisbane	Examinatrice
Pauline ZARROUK LPNHE, Paris	Examinatrice
Cristinel DIACONU CPPM, Marseille	Président du jury
Julian BAUTISTA CPPM, Marseille	Directeur de thèse

A*Midex
Initiative d'excellence Aix-Marseille

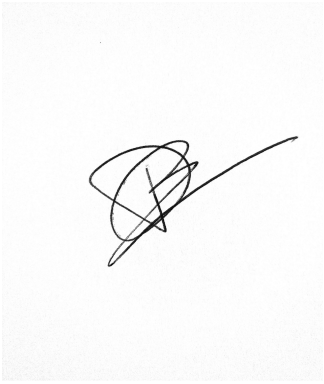
CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE
CPPM

Affidavit

Je soussigné, Tyann Dumerchat, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Julián Ernesto Bautista, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 08/07/2024



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Liste de publications et participation aux conférences

Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse:

1. 2024 **The ACCEL^2 project: simulating Lyman-α forest in large-volume hydrodynamical simulations**, Chabanier and Ravoux et al., [arXiv2407.04473]
2. 2024 **Emulating the non-linear peculiar velocity clustering**, Proceedings of the 56th Rencontres de Moriond : Cosmology
3. 2023 **Galaxy clustering multi-scale emulation**, Dumerchat et al., [A&A]
4. 2022 **Baryon acoustic oscillations from a joint analysis of the large-scale clustering in Fourier and configuration space**, Dumerchat et al., [A&A]

Participation aux conférences et écoles d'été au cours de la période de thèse:

1. 2024 DESI collaboration meeting, Marseille, France
2. 2024 Moriond cosmology conference, La Thuile, Italy
3. 2023 DESI collaboration meeting, Waikoloa, Hawaï
4. 2023 Future cosmology summer school, Cargèse, France
5. 2022 DESI collaboration meeting Cancún, Mexico
6. 2022 Action Dark Energy, Marseille, France
7. 2022 Euclid France Clustering, Marseille, France
8. 2022 DESI collaboration meeting, Berkeley, USA
9. 2022 In2p3 school of statistics, Carry-le-Rouet, France
10. 2022 DESI France, Marseille, France
11. 2021 Action Dark Energy, Marseille, France
12. 2021 Gif summer school, Marseille, France

Remerciements

Paraît-il qu'après trois années acharnées à prouver sa valeur et écrire un manuscrit bien propre tout beau, il faut dire merci. Et puis il faut faire ça bien. On veut de l'émotion, de la gratitude et des anecdotes croustifondantes. Le tout saupoudré d'un sérieux académique, parce que les gens éduqués savent doser l'humour.

Alors je commence par les institutions, c'est franchement le moins marrant. Je remercie l'université d'Aix-Marseille, la chaire d'excellence Amidex et le centre de physique des particules de Marseille pour m'avoir permis d'effectuer cette thèse. Pratique le service publique. Passons aux êtres humains parce que c'est aussi vachement important.

Merci à toute l'équipe du CPPM pour l'accueil et l'ambiance encore inégalée dans un labo de recherche. Merci à tout le groupe de cosmologie pour les échanges scientifiques enrichissants, et aussi pour ceux qui ne l'étaient pas. Ça fait toujours plaisir le contact humain quand on travaille seul derrière un écran. Merci plus particulièrement à l'équipe darkuni, c'était un régale de travailler avec vous et de faire semblant d'écouter vos problèmes de code pendant les réunions du lundi. Merci toujours plus particulièrement à Vincenzo (dit le V) pour ta présence rassurante, mais surtout la qualité de tes blagues. Avec toi l'éducation nationale est sauvée y a pas de doute (il est prof maintenant, faut suivre). Merci bien évidemment à Julian le plus beau des superviseurs. Merci de m'avoir permis de voyager, merci pour ta bienveillance, tes conseils et ta fêtardise. C'est quand même dommage que tu m'aies pas enseigné ton déhancher, mais on peut pas tout avoir.

Merci à toustes mes gâtés de Marseille, même les non-scientifiques (la condescendance de classe c'est moche). Merci pour les lundis Sham, les mardis Papamousse et les autres jours de la semaine où on a fait la fête. N'oublions pas les aprems jeux de sociétés, les sessions sport, les bonnes bouffes et les sorties en plein air. Y a pas que la débauche dans la vie. Merci surtout à mes amis thésards, ça change vraiment la donne de savoir que d'autres personnes partagent votre détresse psychologique. Je vais pas faire de dédicaces personnalisées parce qu'ils sont trop nombreux et que ces remerciements sont déjà trop longs. Aller j'écris quand même leurs prénoms ça leur fera surement plaisir (ils vont trouver le moyen de me reprocher l'ordre d'écriture, ces ingrats) : Marie, Tutur, Maxime, Grode, Luc, Jeannot, Gaya. C'est fou j'ai toujours l'impression d'oublier quelqu'un. En plus d'être longs ces remerciements sont de moins en moins clairs. Je reconnais que j'écris pas de la manière la plus concise possible, mais j'ai besoin de me défouler. Et puis vous avez qu'à en faire une de thèse, comme ça vous pourrez bien écrire comme ça vous chante. "Je remercie mon chien, ma mère, la votre..." Bon je suis aller me calmer ça va mieux maintenant alors je continue.

Je remercie aussi tous mes amis d'enfance. Ils m'ont pas particulièrement aidé à écrire ma thèse, mais les autres non plus remarque. De toute façon on a pas souvent l'occasion de remercier les gens qu'on aime de manière formelle, donc je remercie bien qui je veux si ça me fait plaisir. Petite dédicace quand même à Lulu parce qu'on fait le même métier et parce qu'il est venu travailler avec moi à Marseille. On a même un papier ensemble, et ça, c'est vraiment stylé. Je vais remercier mon amoureuse maintenant parce que c'est trop clicher de la garder pour la fin et puis j'en ai marre d'attendre. Merci madame doudou pour ces heureuses années de vie commune, ça

aurait vraiment été plus rude sans toi. Merci de m'avoir supporté quand j'étais aigri et fatigué. Merci pour ta fantaisie, merci pour nos instants de discussion intense, merci pour ton velouté de brocoli c'est quand même sympa l'hiver. Oui je sais, j'ai tué l'émotion, mais ça dédramatise et c'est important de bien manger.

Aller c'est presque fini, reste plus que la famille, c'était pas si long finalement. Merci à ma famille toute entière ah oui mais là j'utilise deux fois le mot famille ça le fait pas. J'ai vérifié les synonymes et ils me plaisent pas trop alors je laisse comme ça. Merci à ma famille toute entière pour m'avoir soutenu et m'avoir regardé comme si c'était évident que j'allais réussir. Pas mal pour la confiance en soi. Reconnaissons tout de même les mérites du déterminisme social, si vous aviez été moins géniaux j'aurais surement pas pu faire chercheur. J'aurais même pu finir patron.

> "*C'est drôle comme les gens qui se croient instruits éprouvent le besoin de faire chier le monde.*"
>
> Boris Vian, *Les Fourmis*

(J'ai oublié de prévenir mais foutre des citations c'est aussi la coutume)

Abstract and keywords

Galaxies trace the large-scale structures, and contains richfull information about the physical laws governing the evolution of our Universe. The Baryon acoustic oscillations (BAO) and Redshift space distortions (RSD) are powerful probes, imprinted in the spatial distribution of observed galaxies, allowing to measure the expansion and growth history of the Universe and ultimately constrain the cosmological model. In this thesis, I built different analyses beyond the standard methods to leverage the available information and improve the cosmological constraints. I developed a new BAO analysis, jointly fitting the galaxy power spectrum and correlation function while accounting for their strong correlations, and compared the results with the standard method, which directly combines the parameters' correlated posteriors. I found that my new method yields larger estimated uncertainties but smaller biases on BAO parameters while relying on fewer assumptions. I performed a joint analysis of the luminous red galaxy sample of the extended baryon oscillation spectroscopic survey (eBOSS) data release 16, obtaining $D_H/r_d = 19.27 \pm 0.48$ and $D_M/r_d = 17.77 \pm 0.37$, in agreement with the official eBOSS consensus BAO-only results $D_H/r_d = 19.33 \pm 0.53$ and $D_M/r_d = 17.86 \pm 0.33$. While galaxy clustering on non-linear scales is challenging to model, it contains a lot of cosmological information, especially sensible to the growth rate of structures. Simulation based inference (SBI) has seen increasing interest in the past few years as a promising approach to model those scales of galaxy clustering. I present a new Gaussian process model that extends the input parameter space dimensions, and apply it to simultaneously emulate every scale of the non-linear clustering of galaxies in redshift space measured from the AbacusSummit N-body simulations along with the halo occupation distribution formalism. I trained the emulator on scales from 0.3 h^{-1}Mpc to 60 h^{-1}Mpc, and found improved performance in constraining cosmological parameters compared to the standard approach of building one independent emulator for each scale. Conducting a multi-tracer analysis is crucial for obtaining the most accurate measurements of cosmological parameters. I extended the SBI framework to jointly analyse the non-linear galaxy and peculiar velocity auto- and cross-correlations. For a realistic density of tracers and measurement uncertainties, combining peculiar velocities and galaxy information increases the constraining power by 40% for $f\sigma_8$. Finally, I used the MilleniumTNG hydrodynamical simulation to study the errors in the peculiar velocity measured from the Tully-Fisher and fundamental plane relations. I detected that the intrinsic physical scatter of those relations correlates with the local environment on scales typically below $\sim 10h^{-1}$Mpc, biasing the velocity correlation function on small scales if not properly accounted for.

Keywords : Cosmology – Large scale structures – Galaxy clustering – Peculiar velocity – Multi-tracer – Simulation – Machine learning – Emulation

Résumé et mots clés

Le regroupement des galaxies trace les structures à grande échelle et nous informe sur les lois physiques régissant l'Univers. Les oscillations acoustiques des baryons (BAO) et les distorsions spatiales provoquées par le décalage vers le rouge (RSD) sont de puissantes sondes, marquées dans la distribution spatiale des galaxies que nous observons. Elles permettent de tester différents modèles cosmologiques en mesurant l'histoire de l'expansion de l'Univers et la croissance de ses structures. Au cours de cette thèse, j'ai conduit différentes analyses au-delà des méthodes standards, dans le but d'exploiter le maximum d'informations contenues dans les données, et améliorer les contraintes cosmologiques. J'ai développé une nouvelle analyse des BAO, ajustant simultanément le spectre de puissance et la fonction de corrélation des galaxies. Cette nouvelle méthode donne des incertitudes estimées plus importantes pour les paramètres BAO qu'avec la méthode standard, qui combine directement les postérieurs des paramètres corrélés. Cependant, cette nouvelle méthode repose sur moins d'hypothèses et produit des biais systématiques plus faibles. J'ai effectué une analyse jointe de l'échantillon de galaxies rouges lumineuses de la seizième version des données du relevé spectroscopique eBOSS, obtenant $D_H/r_d = 19.27 \pm 0.48$ et $D_M/r_d = 17.77 \pm 0.37$, en accord avec les résultats BAO officiels de eBOSS $D_H/r_d = 19.33 \pm 0.53$ et $D_M/r_d = 17.86 \pm 0.33$. Bien que la distribution des galaxies aux échelles non linéaires soit difficile à modéliser, elle contient une grande quantité d'informations cosmologiques. Elle est particulièrement sensible au taux de croissance des structures. L'inférence basée sur des simulations (SBI), permettant de modéliser ces échelles, a suscité un intérêt croissant ces dernières années. J'ai développé un nouveau modèle de processus gaussien permettant d'étendre la dimension de l'espace des paramètres à émuler. J'ai testé ce formalisme en émulant simultanément chaque échelle de la fonction de corrélation des galaxies (dans l'espace des décalages vers le rouge), mesurée dans les simulations à N-corps AbacusSummit. En entraînant l'émulateur sur des échelles de 0.3 h^{-1}Mpc à 60 h^{-1}Mpc, j'ai obtenu une performance améliorée pour contraindre les paramètres cosmologiques, par rapport à l'approche standard consistant à construire un émulateur indépendant pour chaque échelle. Analyser plusieurs traceurs des grandes structures est crucial pour obtenir des mesures des paramètres cosmologiques les plus précises possibles. J'ai étendu le cadre SBI pour analyser conjointement les échelles non-linéaires des fonctions de corrélation des galaxies et des vitesses particulières. Pour une densité de traceurs et des incertitudes de mesure réalistes, La combinaison de la distribution spatiale des galaxies et de leurs vitesses particulières diminue les barres d'erreurs de 40% pour $f\sigma_8$. Enfin, j'ai utilisé la simulation hydrodynamique MilleniumTNG afin d'étudier les erreurs de mesure dans les vitesses particulières, basées sur les relations de Tully-Fisher et du plan fondamental. J'ai détecté que la dispersion physique intrinsèque de ces relations était corrélée avec l'environnement local, pour des échelles typiquement inférieures à $\sim 10 h^{-1}$Mpc. Si elles ne sont pas correctement prises en compte, ces corrélations pourraient biaiser les petites échelles de la fonction de corrélation des vitesses.

Mots clés : Cosmologie – Structures à grande échelle – Regroupement de galaxies – Vitesse particulière – Multi-traceurs – Simulation – Apprentissage automatique – Émulation

Résumé substantiel en français

"*Le métaphysicien explique l'univers d'après lui-même, le savant s'explique lui-même d'après l'univers.*"

Jack London, *Le Talon de fer*

Notre perception de l'Univers

Selon le modèle standard de la cosmologie, voici l'état actuel de nos connaissances sur l'histoire de l'Univers. L'espace-temps est créé au cours d'une singularité sans aucune description physique à ce jour : le Big Bang. À ce moment initial, l'Univers est dans son état le plus dense et le plus chaud, et toutes les particules élémentaires existantes sont enchevêtrées dans un plasma primordial, parcouru de fluctuations quantiques. Avant même que la première seconde soit écoulée, l'Univers connaît une phase d'inflation brutale, étirant l'espace et ses fluctuations quantiques à des échelles cosmologiques avant de ralentir brusquement son expansion. Tandis que l'Univers continue de s'étendre et de se refroidir, les protons et les neutrons commencent à se former à partir des particules élémentaires. La nucléosynthèse primordiale du Big Bang a lieu dans les premières minutes : les protons et les neutrons se lient ensemble dans un état stable, formant les noyaux atomiques d'éléments légers. L'Univers reste cependant extrêmement dense et chaud, et la matière baryonique et le rayonnement demeurent couplés par des collisions continues. Cet état perdure pendant environ 380 000 ans, jusqu'à ce que l'Univers se dilate et refroidisse suffisamment pour débuter une phase de recombinaison où que les électrons et les protons chargés se lient pour former des atomes neutres. Les photons n'interagissent alors plus avec les atomes, et commencent à voyager librement à travers l'Univers, constituant un fond diffus cosmologique. Après la recombinaison, l'Univers entre dans une ère sombre où la matière continue de se refroidir et s'agglutine sous l'effet de la gravité, formant progressivement des régions plus denses autour des fluctuations primordiales qui continuaient d'évoluer. Après quelques centaines de millions d'années, l'accumulation de gaz et de poussières forme les premières étoiles, initiant ainsi la création d'éléments plus lourds (nucléosynthèse stellaire). Au fil du temps, les étoiles et le gaz s'agglomèrent pour former des galaxies, réionisant progressivement l'Univers avec leurs rayonnements énergétiques. L'Univers continue de s'étendre, et les galaxies suffisamment proches s'attirent et se regroupent pour former des structures à grande échelle constituées de filaments et d'amas : les plus grandes structures liées de l'Univers. Environ 9 milliards d'années après le Big Bang, quelque part dans le superamas de galaxies Laniakea, à l'intérieur de l'amas de la Vierge, au sein de la galaxie de la Voie lactée, notre système solaire et notre planète Terre sont formés. À la même époque, l'expansion de l'Univers commence à s'accélérer et sépare de plus en plus rapidement les structures non liées gravitationnellement. Presque 14 milliards d'années après le commencement, nous observons le ciel, des étoiles les plus proches aux plus anciennes lumières, pour penser cette histoire.

Introduction

La cosmologie est l'étude de l'origine, de l'évolution et du destin de notre Univers dans son ensemble. Bien sûr, notre conception de l'Univers en tant que 'tout ce qui existe' a varié au travers des époques. Notre compréhension grandissante de l'immensité de l'Univers fait finalement émerger la cosmologie comme une discipline scientifique distincte au début du XXe siècle. Le développement de la théorie de la relativité générale d'Albert Einstein fournit alors un cadre mathématique pour décrire les grandes échelles de notre Univers. Dans les années 1920, Hubble et Lemaître découvrent que notre galaxie n'est pas la plus grande structure de l'Univers. En effet, de nombreuses autres galaxies existent au-delà de la nôtre, et elles s'éloignent de nous à des vitesses proportionnelles à leurs distances, indiquant un Univers en expansion. Une dizaine d'années plus tard, l'astronome Zwicky remarque que la masse visible des galaxies est insufisante pour expliquer leur effets gravitationnel. Il conclut alors à l'existence d'une masse "invisible" qu'il désigne comme matière sombre. La découverte la plus importante en cosmologie est peut-être la mesure du rayonnement cosmique dans les années 1960, fournissant une preuve solide en faveur de la théorie du Big Bang. À la fin des années 1990, l'observation de l'expansion accélérée de l'Univers apporte l'ingrédient final au modèle cosmologique : l'énergie sombre, symbolisée Λ (constante cosmologique). Le modèle standard est alors baptisé ΛCDM (cold dark matter).

Ce dernier s'est avéré extrêmement robuste et cohérent avec une large gamme d'observations, permettant de contraindre ses paramètres avec une grande précision. Alors que le modèle standard décrit précisément le contenu, l'évolution et la structure de notre Univers, depuis plusieurs décennies des anomalies significatives et de grandes questions ouvertes persistent. Celles-ci incluent la compréhension de la période inflationnaire initiale de l'Univers, la détermination de la masse des neutrinos, l'élucidation de l'origine des écarts entre les valeurs observées et prédites pour le taux d'expansion et l'amplitude des fluctuations de matière, ainsi que la nature de la matière noire et de l'énergie sombre. Grâce aux relevés passés, en cours et à venir, qui couvrent toujours plus de volume et sondent plus profondément les décalages vers le rouge, la cosmologie est entrée dans son *âge d'or*, foisonnant de données. Le potentiel de rejeter définitivement le modèle ΛCDM en faveur d'un modèle supérieur pourrait se trouver dans des données inexploitées. Le défi actuel réside dans la gestion efficace, l'analyse et l'extraction d'informations significatives à partir de ces ensembles de données, en affinant les stratégies d'analyse, en développant des outils de modélisation théorique et statistique, et en abordant correctement les erreurs systématiques.

Cette thèse est dédiée à l'étude des structures à grande échelle de l'Univers, afin de contraindre les modèles d'énergie sombre et de gravité alternative. Mon attention a principalement été portée sur la distribution des galaxies, reflétant la façon dont la matière s'agglomère sous l'influence de la gravité et de l'énergie sombre. J'ai exploré différentes sondes et observables, utilisant des outils d'apprentissage automatique pour développer de nouvelles méthodes et améliorer nos contraintes et notre compréhension de l'expansion et de la croissance de l'Univers. La plupart des mystères sur l'histoire et la structure de l'Univers étant déjà élucidés, mon travail ne contient aucune découverte révolutionnaire. J'ai tout de même essayé de présenter dans ce manuscript, aussi précisément et simplement que possible, les diverses connaissances que j'ai acquises et ma modeste contribution au domaine.

Ce chapitre résume les principaux résultats et analyses effectuées au cours de cette Thèse. Dans la section 0.1, je discute de l'analyse des échelles linéaires au travers des mesures BAO. La section 0.2 introduit une nouvelle procédure d'émulation des échelles non-linéaires à par-

tir de simulations à N-corps. L'analyse jointe de fonctions de corrélations des vitesses et des galaxies est décrite dans la section 0.3. Enfin, la section 0.4 présente l'étude des erreurs systématiques de mesures de vitesses particulières à partir des relations de Tully-Fisher(TF) et du plan fondamental(FP) dans la simulation hydrodynamique MilleniumTNG.

0.1 Mesure jointe des oscillations baryoniques acoustiques

Les mesures BAO peuvent être effectuées à partir de la fonction de corrélation ou du spectre de puissance (dans l'espace de Fourier) des galaxies. Les deux types d'analyses effectuent une mesure statistique de la distribution des galaxies à partir du même catalogue de positions et d'erreurs systématiques. En pratique, les estimateurs sont construits de manière différente et les échelles étudiées ne correspondent pas parfaitement. Les deux observables perdent donc leur dégénérescence parfaite, et il y a un léger gain à combiner les mesures dans les deux espaces. Afin de produire une valeur de consensus pour les paramètres mesurés, il est important de prendre en compte les fortes corrélations entre les deux méthodes. La méthode standard (GA), introduite dans SÁNCHEZ et al. 2017, a été étendue dans ALAM et al. 2017 pour les dernières mesures du relevé SDSS-IV eBOSS. Cette méthode suppose des distributions postérieures Gaussiennes pour les paramètres afin de les combiner une fois les mesures effectuées. Nous avons développé une analyse alternative (JS) pour effectuer des mesures BAO sans reposer sur cette hypothèse, en ajustant simultanément la fonction de corrélation et le spectre de puissance, tout en prenant en compte les corrélations entre les différentes échelles des deux observables.

Pour ajuster la position du pic BAO et mesurer les paramètres $\alpha_\parallel$ and $\alpha_\perp$ nous utilisons le modèle linéaire pour le spectre de puissance anisotrope décrit dans J. E. BAUTISTA et al. 2021 ; GIL-MARÍN et al. 2021, basé sur de précédents travaux ALAM et al. 2017 ; GIL-MARÍN et al. 2016 ; GIL-MARÍN et al. 2018 ; ROSS et al. 2017 ; J. E. BAUTISTA et al. 2018. Ce modèle s'écrit

$$P_{\rm gg}(k,\mu_k) = b_1^2\left[1+\beta\,(1-S(k))\,\mu_k^2\right]^2 D_g^2(k,\mu_k)\times \\ \times\left[P_{\rm no\ peak}(k) + P_{\rm peak}(k)\,e^{-k^2\Sigma^2_{\rm nl}(\mu_k)/2}\right], \tag{0.1}$$

avec b_1 le biais linéaire de densité, et $\beta = f/b$ le paramètre de distorsion du décalage vers le rouge (RSD). Les mouvements non-linéaires aux petites échelles sont modélisés par une distribution de Lorentz, paramétrée par l'échelle caractéristique Σ_s

$$D_g^2(k,\mu_k) = \frac{1}{1+k^2\mu_k^2\Sigma_s^2/2}. \tag{0.2}$$

Une fonction Gaussienne anisotrope est appliquée au pic BAO pour modéliser l'élargissement causé par les mouvements non-linéaires. Tout d'abord, le spectre de puissance linéaire $P_{\rm L}$ est obtenu à partir de CAMB (LEWIS et al. 2000). Il est ensuite décomposé - selon la procédure décrite dans KIRKBY et al. 2013 - en deux termes avec et sans le pic BAO $P_{\rm L} = P_{\rm peak} + P_{\rm no\ peak}$. La composante $P_{\rm peak}$ est ensuite multiplié par une fonction Gaussienne avec une dispersion donnée par $\Sigma^2_{\rm nl}(\mu_k) = \Sigma^2_\parallel\mu_k^2 + \Sigma^2_\perp(1-\mu_k^2)$. La méthode standard consiste à réduire l'impact de l'évolution non-linéaire des galaxies en appliquant la technique de reconstruction décrite dans BURDEN et al. 2015, et modélisée suivant SEO et al. 2016, avec $S(k) = e^{-k^2\Sigma_r^2/2}$ et $\Sigma_r = 15h^{-1}$Mpc le

paramètre de lissage. La position du pic BAO est ensuite décrite par les paramètres de dilatation (ALCOCK et PACZYNSKI 1979) parallèle et transverse à la ligne de visée

$$\alpha_\perp(z) = \frac{D_M\, r_{\rm d}^{\rm fid}}{D_M^{\rm fid}\, r_{\rm d}}, \quad \alpha_\parallel(z) = \frac{D_H\, r_{\rm d}^{\rm fid}}{D_H^{\rm fid}\, r_{\rm d}}, \tag{0.3}$$

modélisant la dilatation de la séparation (ALCOCK et PACZYNSKI 1979) parallèle et transverse à la ligne de visée. Enfin, le spectre de puissance anisotrope $P_{\rm gg}(k,\mu_k)$ est décomposé en multipôles de Legendre $P_\ell(k)$ pour l'espace de Fourier (FS), ou $\xi_\ell(s)$ pour l'espace des configurations (CS), en ne conservant que les deux premiers multipôles pairs $\ell = [0,2]$ (monopôle, quadrupôle). Une combinaison linéaire de polynômes avec des amplitudes libres est ajoutée à chaque multipôles afin de marginaliser toute erreur de modélisation sans affecter le signal BAO.

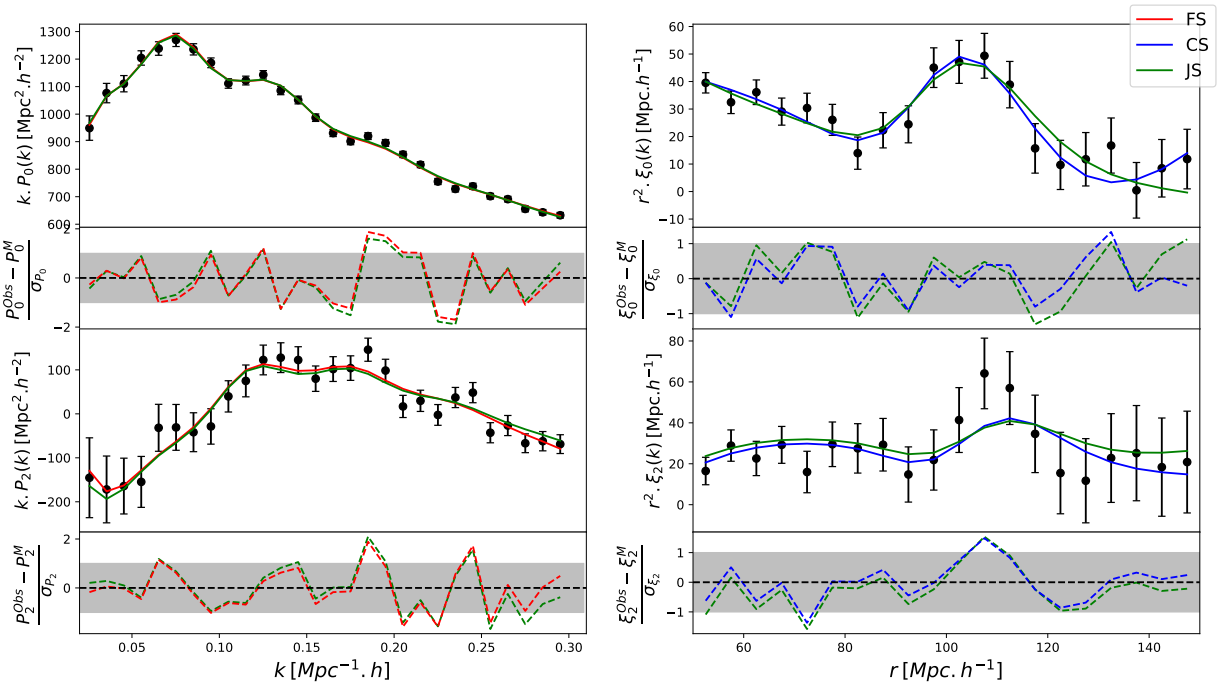


FIGURE 1 : Meilleurs modèles de monopôle et de quadrupôle pour les analyses CS, FS et JS. Les résidus par rapport aux mesures sont standardisés pour chaque point par la barre d'erreur. La zone grisée correspond à une différence de 1σ.

Nous procédons la mesure des paramètres BAO sur 1000 simulations EZmocks, en ajustant séparément puis simultanément le spectre de puissance et la fonction de corrélation. Pour chaque analyse séparée, nous calculons les résultats obtenus par la méthode GA. Les contraintes sur les paramètres sont obtenue en maximisant la fonction de vraisemblance (supposée Gaussienne), avec une matrice de covariance estimée à partir des 1000 EZmocks. La statistique récapitulative des ajustements est présentée dans la table 3.4. Nous observons des résultats cohérents entre chaque méthode, retournant des paramètres non biaisés. Cependant, du fait de la taille supérieure du vecteur de données, la matrice de covariance est affectée par un bruit plus important pour l'analyse JS, et requiert un plus grand nombre de simulations pour être correctement estimée. Par conséquent, la méthode GA donne les contraintes les plus fortes sur les paramètres BAO.

Notre méthode diminue cependant les erreurs systématiques sur le paramètre $\alpha_\parallel$, et performe mieux que la méthode GA quand les contours obtenus dans les analyses séparées sont non Gaussiens. La figure 1 montre les meilleurs ajustements de l'échantillon de galaxies lumineuses rouges des données eBOSS, obtenu pour les analyses CS, FS and JS. Nous obtenons des valeurs de $D_H/r_d = 19.27 \pm 0.48$ et $D_M/r_d = 17.77 \pm 0.37$, en excellent accord avec le consensus officiel BAO de eBOSS $D_H/r_d = 19.33 \pm 0.53$ et $D_M/r_d = 17.86 \pm 0.33$.

0.2 Inférence par simulation

La distribution des galaxies à petite échelle est riche en information et particulièrement sensible au taux de croissance des structures. Aux plus petites échelles, la matière est fortement agglomérée, générant des perturbations non-linéaires que les modèles perturbatifs échouent à décrire. Les échelles non-linéaires de la fonction de corrélation des galaxies ne sont donc pas utilisées dans les analyses standards, faute de modèle théorique valable. Les simulations cosmologiques à N-corps permettent de modéliser ces échelles en résolvant numériquement les équations de mouvements pour chaque 'particule' de matière noire. Cependant, elles sont extrêmement coûteuses numériquement et ne peuvent être produites que pour un nombre limité de cosmologies. Les algorithmes d'apprentissage automatique permettent d'apprendre une fonction de sortie mesurée dans les simulations (comme une fonction de corrélation), à partir d'un échantillon réduit. Le modèle construit, appelé émulateur, est alors capable d'interpoler sur un espace de paramètres cosmologiques continu. Un algorithme couramment utilisé est le processus Gaussien (GP), ne nécessitant d'ajuster qu'un nombre limité d'hyperparamètres interprétables physiquement. De plus, le GP permet d'estimer l'incertitude sur la prédiction du modèle, ce qui est primordial pour ne pas biaiser les contraintes sur les paramètres cosmologiques. Cependant, ce type d'algorithme est limité par la taille de l'échantillon d'entraînement, et il est nécessaire de construire des émulateurs sur l'espace des paramètres cosmologiques et de biais (connexion halo-galaxie) indépendant pour chaque échelle. Nous avons développé un nouvel algorithme permettant d'étendre la dimensionnalité de la fonction à émuler, afin d'inclure plus de variable à l'émulation. Nous testons notre méthode en émulant simultanément les différentes échelles (en plus des autres paramètres) de $0.3h^{-1}$Mpc à $60h^{-1}$Mpc du monopole de la fonction de corrélation des galaxies, mesuré à partir de la simulation à N-corps AbacusSummit (Maksimova et al. 2021) et du formalisme HOD pour la connexion galaxie-halo (Zheng et al. 2007). Nous entraînons également un modèle standard indépendamment pour chaque échelle (IGP), afin de comparer avec la performance du modèle multi-échelle (MGP).

La figure 2 décrit la différence relative médiane (MAE) entre les prédictions des modèles MGP (bleue) et IGP (vert), et l'échantillon test exclu de la phase d'apprentissage, composé de 120 fonctions de corrélation balayant 6 cosmologies différentes, chacune avec 20 modèles HOD. La ligne rouge en pointillée indique le niveau de précision attendue pour l'échantillon test, estimé par la variance cosmique médiane. Les modèles MGP et IGP ont une performance similaire et décrivent correctement l'échantillon test. Nous utilisons ces deux modèles pour procéder à l'ajustement des paramètres cosmologique et HOD sur 25 réalisations de la cosmologie Planck2018 (Planck Collaboration et al. 2020), utilisant les paramètres HOD au centre de l'échantillon d'apprentissage. Nous effectuons des inférences MCMC avec des a priori uniformes sur l'ensemble des paramètres, et une matrice de covariance estimée à partir de 1400 réalisations d'un volume réduit. Les résultats sont présentés dans la table 4.2. Dans les deux cas, nous obtenons des paramètres cosmologiques et HOD non biaisés. Cependant, ne pas inclure l'incertitude du modèle

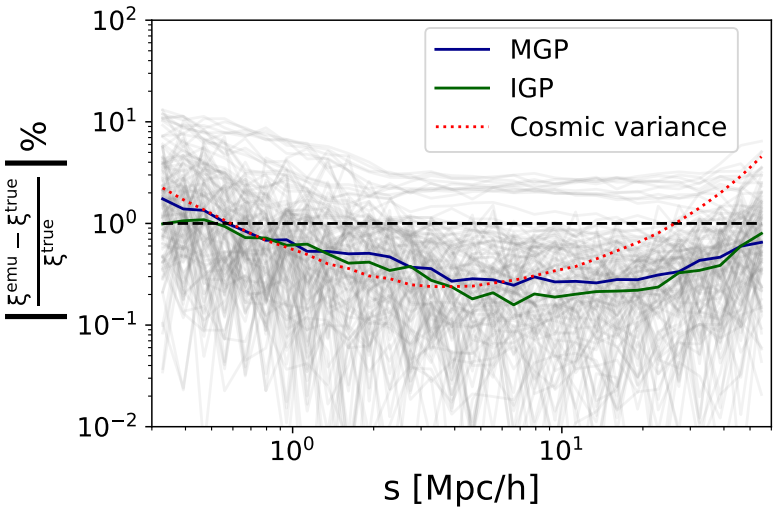


FIGURE 2 : Les lignes grises montrent la différence relative exprimée en pourcentage entre la prédiction de l'émulateur MGP et les valeurs attendues pour l'ensemble de l'échantillon test. Les lignes bleue et verte représentent la médiane de ces déviations pour les modèles MGP et IGP, respectivement. La ligne noire pointillée correspond à une déviation de 1%. La ligne rouge pointillée indique la précision attendue pour l'échantillon test : la médiane de $|\sigma/\xi^{\text{true}}|$ en pourcentage. Ici, σ^2 est la variance cosmique calculée pour la cosmologie Planck2018 et le HOD médian.

lors de l'inférence engendre des erreurs systématiques importantes sur tous les paramètres. Par ailleurs, utiliser le modèle MGP réduit les erreurs systématiques et augmente les contraintes sur les paramètres cosmologiques. Ce nouveau modèle permet de modéliser les distorsions AP, améliorant à nouveau les contraintes cosmologiques.

0.3 Analyse combinée des galaxies et des vitesses particulières

La modélisation des échelles non-linéaires n'est pas le seul enjeu pour la mesure du taux de croissance des structures. Pour les faibles décalages vers le rouge, un autre défi persistant est la variance cosmique, plus importante que les erreurs statistiques lorsque la densité de traceurs est importante. Combiner la distribution des galaxies avec des mesures directes des vitesses particulières aide à réduire l'impact de la variance cosmique en exploitant l'ensemble des données disponibles, et augmente la précision des mesures. Les relevés de vitesses particulières peuvent être construits en combinant des mesures spectroscopiques de galaxies avec des indicateurs de distance tels que la relation de Tully-Fisher (TF), le plan fondamental (FP) ou les supernovae de type Ia (SnIa). Des études récentes ont montré que l'analyse jointe des les fonctions de corrélation des galaxies et des vitesses particulières sur les échelles linéaires augmente le pouvoir de contrainte sur les paramètres cosmologiques (ADAMS et BLAKE 2017 ; TURNER et al. 2023). En utilisant la méthodologie précédement décrite, nous entraînons des émulateurs pour les échelles non linéaires des multipoles $[\xi_0^{gg}, \xi_2^{gg}, \xi_0^{vv}, \xi_2^{vv}, \xi_1^{vg}]$, capables d'interpoler sur les paramètres HOD, cosmologiques, et différentes échelles. Nous testons la précision de nos modèles en mesurant la MAE sur l'échantillon test.

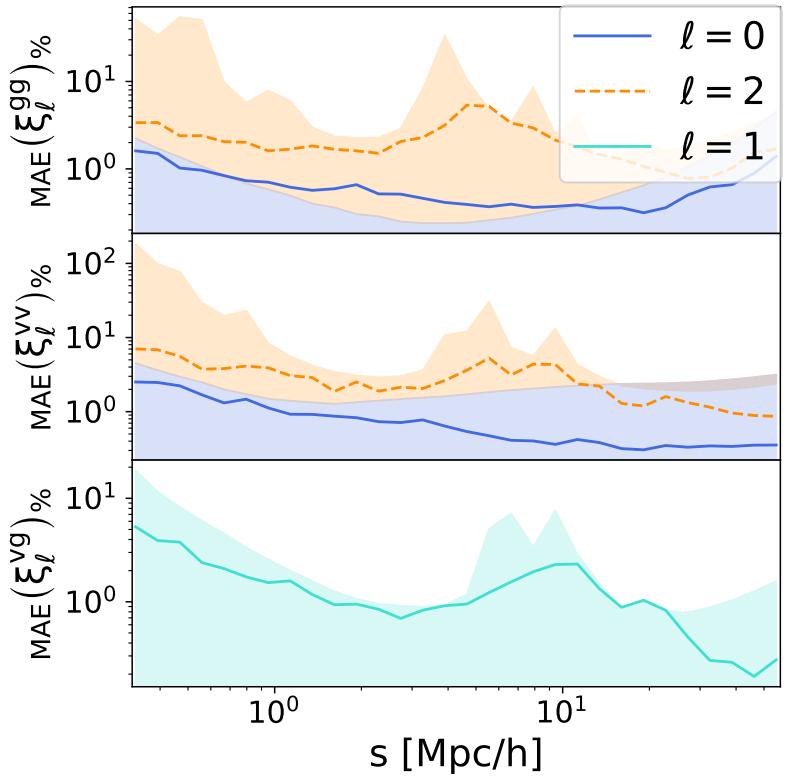


FIGURE 3 : Erreur Médiane Absolue (MAE) évaluée sur l'échantillon test. Les lignes bleue et orange dans le panneau du haut (milieu) montrent les performances des modèles ξ_0^{gg} et ξ_2^{gg} (ξ_0^{vv} et ξ_2^{vv}). La ligne turquoise dans le panneau inférieur montre les performances pour ξ_1^{vg}. Les zones ombrées indiquent la précision attendue pour l'échantillon test : la médiane de $|\sigma/\xi^{\text{test}}|$ avec σ la variance cosmique.

La figure 3 présente la MAE (en %) en fonction de l'échelle pour les cinq multipoles. Bien que nous trouvions de bonnes performances pour chaque modèle, avec des MAE correspondant globalement aux précisions attendues, nous observons quelques divergences. Pour vérifier statistiquement la précision et les biais potentiels de nos modèles, nous construisons des catalogues de galaxies "réalistes" à partir de 25 simulations de la cosmologie Planck2018, reproduisant des effets de grand angles, de distorsions AP, des erreurs sur les mesures de vitesses (BORUAH et al. 2020), et une densité de traceurs réaliste (SAULDER et al. 2023a ; BELLM et al. 2019). Nous ajustons les paramètres cosmologiques et HOD des 25 réalisations en utilisant uniquement les informations des galaxies, les informations des vitesses, les informations croisées, et l'ensemble des informations combinées. Les résultats sont résumés dans la table 5.2.

La figure 4 donne les valeurs moyennes des figures de mérite (FoM) et des figures de biais (FoB) des ajustements. Pour chaque analyse, nous retrouvons des paramètres cosmologiques non biaisés. Par ailleurs, bien que l'inclusion de l'information des vitesses particulières bruitées biase les paramètres HOD (en marginalisant sur le bruit à petite échelle), cela conduit à des contraintes cosmologiques plus fortes. Nous obtenons que, pour la densité de traceurs et les incertitudes de mesure attendues, combiner l'information non-linéaire des vitesses particulières et des galaxies réduit les barres d'ereurs de 40% pour $f\sigma_8$.

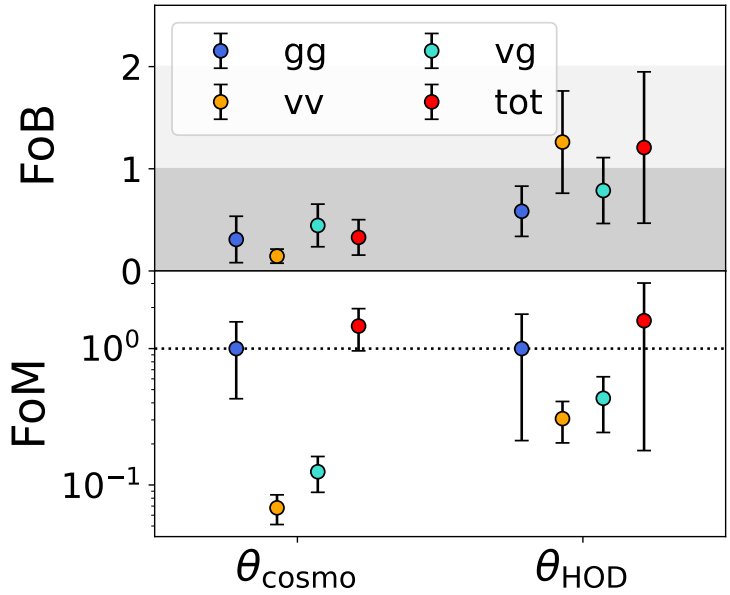


FIGURE 4 : Mesure des paramètres cosmologiques et HOD pour les quatre stratégies d'inférence exécutées sur les mocks réalistes. Panneau supérieur : FoB normalisée par le $\sqrt{\chi^2}$ associé à une p-value de 0.68%, c'est-à-dire 10.4 pour $\theta_{\rm cosmo}$ et 5.9 pour $\theta_{\rm HOD}$. La zone ombrée correspond à 1 et 2 σ CL. Panneau inférieur : FoM normalisée par les valeurs "gg", 4.85×10^{13} pour $\theta_{\rm cosmo}$ et 5.59×10^4 pour $\theta_{\rm HOD}$.

0.4 Relations de Tully-Fisher et du plan fondamentale dans la simulation hydrodynamique MilleniumTNG

Les galaxies spirales montrent une forte corrélation entre leur magnitude absolue M et leur vitesse de rotation maximale, communément connue comme la relation de Tully-Fisher (TF). En calibrant cette relation sur un catalogue de galaxies avec des distances connues, il est possible de déduire la magnitude absolue de n'importe quelle galaxie du même type. En combinant cette estimation à une mesure de magnitude, on peut estimer la distance et donc la vitesse particulière de la galaxie. La relation du plan fondamental (FP) est l'équivalent pour les galaxies elliptiques. Elle relie l'intensité et la dispersion de vitesse au rayon effectif physique. Une mesure de la taille apparente permet alors d'estimer les distances et vitesses de galaxies du même type. Cependant, en plus des incertitudes observationnelles, ces relations souffrent d'une dispersion intrinsèque, entraînant des erreurs considérables dans les mesures de vitesse. Notre analyse précédente a supposé que le bruit dans la mesure des vitesses particulières suivait une distribution aléatoire gaussienne. Nous testons cette hypothèse en construisant les relations TF et FP dans la simulation hydrodynamique haute résolution MilleniumTNG (PAKMOR et al. 2023).

La figure 5 montre la performances des prédictions de nos modèles TF et FP, évaluées pour les mêmes échantillons que ceux utilisés pour la calibration. Bien que les modèles soient non biaisés $\langle M_r - M_r^{\rm TF} \rangle \approx \langle r - r_r^{\rm FP} \rangle \approx 0$, les relations sont caractérisées par une dispersion autour des magnitudes et rayons réels, qui se propage aux vitesses particulières mesurées. Nous mesurons les fonctions d'auto-corrélation des vitesses et les fonctions de corrélation croisée vitesses-galaxies pour les vraies vitesses, et celles estimées. Le clustering est moyenné sur trois lignes de visée en variant la ligne de visée dans la simulation afin de réduire la variance cosmique.

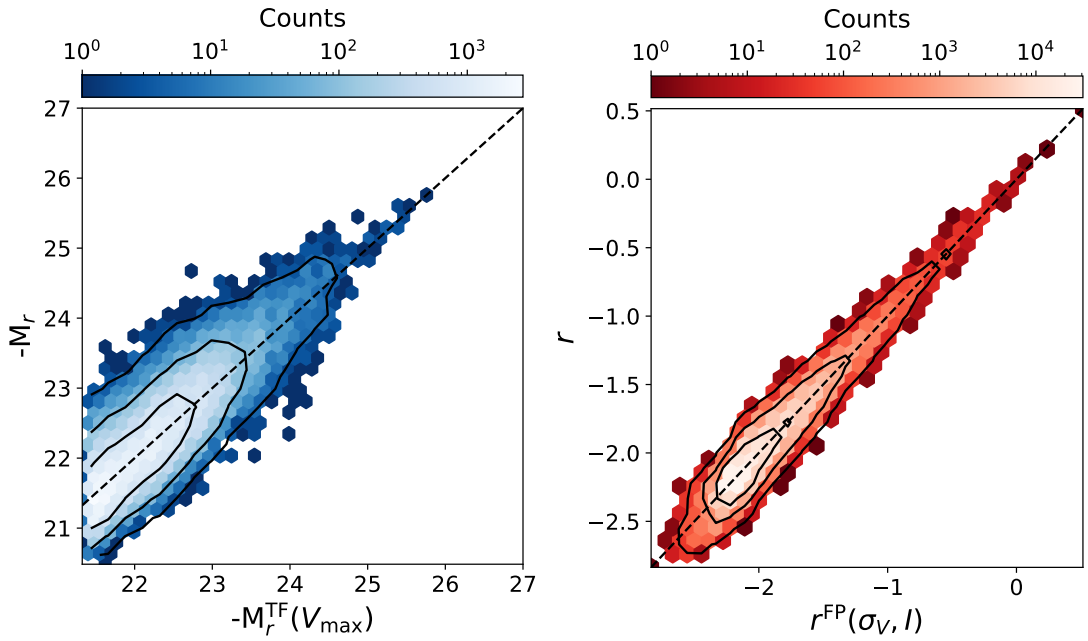


FIGURE 5 : Relations de TF et FP calibrées. Panneau de gauche : magnitudes absolues dans la bande r réelles en fonction de celles prédites à l'aide de la relation TF. Panneau de droite : rayon effectif réel (logarithmique) par rapport à celui prédit à l'aide de la relation FP. Les lignes en pointillés noirs montrent une relation d'équivalence. Les contours noirs correspondent aux régions contenant 68, 95 et 99.7% de la distribution.

La Figure 6 présente la corrélation des vitesses résultant des mesures TF à gauche (en bleu) et FP à droite (en rouge). Pour ξ^0_{vv}, nous constatons que pour les petites échelles, les mesures réalistes s'écartent significativement du clustering attendu, dépassant 1σ du bruit aléatoire à $s \sim 1\ h^{-1}$Mpc pour l'échantillon TF et $s \sim 10\ h^{-1}$Mpc pour l'échantillon FP. Pour le dipôle de la corrélation croisée ξ^1_{vg}, la déviation par rapport au clustering attendu est assez cohérente avec le bruit aléatoire, atteignant à peine 1σ même pour les échelles les plus petites. Le monopole ξ^0_{vg} des vitesses réelles est cohérent avec zéro. Cependant, nous observons un signal fort dans les échantillons TF et FP, dépassant respectivement 3 et 5 σ. En retraçant l'histoire de formation d'étoiles pour différents types de galaxies dans des environnements de forte et faible densité, nous avons découvert que les quantités utilisées pour calibrer les relations TF et FP étaient insuffisantes pour capturer entièrement les propriétés structurelles des galaxies et leur corrélation avec les fluctuations cosmologiques. Utiliser l'information des échelles non linéaires est une tâche difficile puisqu'elles sont affectées par plusieurs effets systématiques observables et théoriques qui doivent être correctement traités. Pour les mesures de vitesses TF et FP, ces échelles devraient être exclues de l'analyse, modélisées avec une expansion de biais supplémentaire, ou leurs incertitudes devraient être augmentées. Il convient également de pousser plus loin notre compréhension des relations TF et FP pour inclure des propriétés (indépendantes de la distance) environnementales supplémentaires, et supprimer ces corrélations résiduelles.

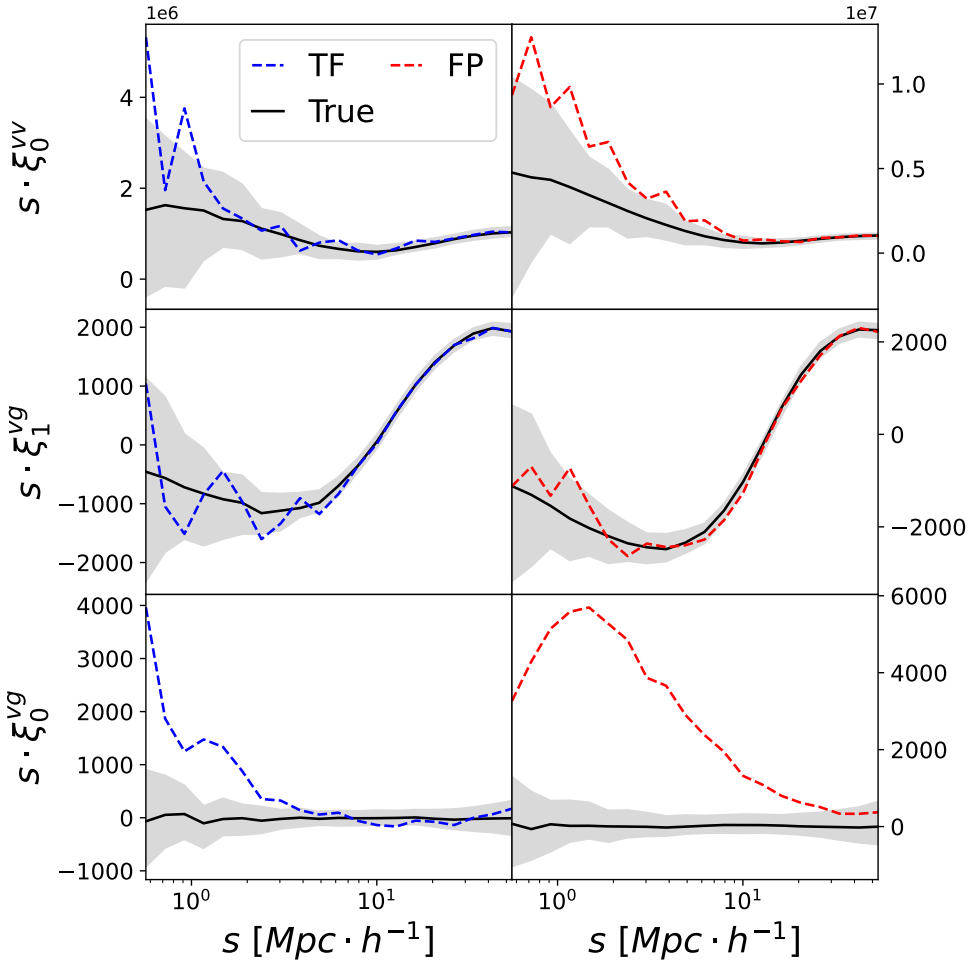


FIGURE 6 : Fonctions de correlations des vitesses mesurées à l'aide des échantillons TF (bleu), FP (rouge) et réelles (noir). Les mesures sont moyennées sur trois lignes de visée en faisant tourner la boîte. La zone ombrée correspond à l'écart type de 1σ par rapport au clustering attendu. Elles sont obtenues en appliquant aux vitesses réelles un bruit gaussien aléatoire (100 réalisations), avec des écarts-types correspondant aux dispersions intrinsèques des relations.

Perspectives

Tout au long de cette thèse, j'ai étudié comment contraindre les paramètres cosmologiques en analysant le regroupement des galaxies et leurs vitesses particulières. Pour relever les défis de la cosmologie moderne, mon objectif a été d'extraire le plus d'informations cosmologiques possible, à partir de la quantité importante de données fournies par les relevés spectroscopiques et photométriques. J'ai comparé différentes méthodes pour combiner les analyses standard et étudié les façons optimales d'utiliser de nouveaux observables. Par ailleurs, j'ai développé des outils d'apprentissage automatique, faisant usage des descriptions théoriques de pointe fournies par les simulations cosmologiques pour modéliser les processus physiques se produisant aux petites échelles. Cet effort pour modéliser précisément le regroupement non linéaire des galaxies pourrait être étendu, afin d'inclure plus de paramètres de biais galactique ou sonder plusieurs

intervalles de décalage vers le rouge. De plus, ce cadre de modélisation par la simulation rend possible la paramétrisation des effets systématiques observationnels tels que les collisions de fibres, principale limitation des petites échelles. La reproduction des effets de sélection soulève cependant le problème de la variance cosmique dans les simulations, notamment à faible décalage vers le rouge où le volume est limité. De futures études pourraient utiliser différentes approches (ANGULO et PONTZEN 2016b ; HADZHIYSKA et al. 2023a) pour réduire la variance associée aux observables émulées. L'émulation de plus grandes échelles nécessiterait de tenir compte correctement des effets de grand angle, en particulier à faible décalage vers le rouge. Notre analyse peut bien sûr être adaptée à toute statistique alternative de la distribution des galaxies. Dans notre analyse multi-traceur, il serait pertinent d'étendre la connexion halo-galaxie au-delà du formalisme standard de la distribution d'occupation des halos, en incluant des biais d'assemblage et de vitesse particulière, ou des approches plus indépendantes telles que la correspondance de l'abondance des sous-halos (SHAM, BEHROOZI et al. 2010), les modèles semi-analytiques (SAM, S. D. M. WHITE et FRENK 1991), ou l'expansion de biais hybride (PELLEJERO IBAÑEZ et al. 2023). Le défi réside dans le fait d'utiliser une connexion halo-galaxie suffisamment flexible pour marginaliser sur la physique non modélisée sans biaiser les résultats, compte tenu de la complexité de la réalité par rapport à nos meilleurs modèles. Il serait également judicieux d'étendre l'analyse 3×2 points pour inclure davantage de multipôles. Pour les travaux futurs, il me semble important d'évaluer et de modéliser l'impact de la cosmologie fiduciaire sur l'estimation des vitesses particulières. Enfin, il sera crucial d'évaluer si les résidus corrélés dans les mesures de vitesse TF et FP seront observés dans les prochaines données (DESI SAULDER et al. 2023a, ZTF BELLM et al. 2019, LSST LSST SCIENCE COLLABORATION et al. 2009), comme cela a été le cas dans les mesures précédentes (JOACHIMI et al. 2015).

Dans les prochaines années, DESI et Euclid observeront le décalage vers le rouge de millions de galaxies pour former la plus grande carte de l'Univers jamais réalisée. Les observations scientifiques de ZTF et celles à venir de Rubin-LSST produiront la plus grande collection de supernovae de type Ia, fournissant une quantité substantielle de mesures de distance et de vitesse particulière. Avec de telles quantités de données, il est crucial de continuer à développer et à affiner nos outils statistiques, nos modèles théoriques et le traitement des effets systématiques. Les progrès continus dans ces domaines sont essentiels pour exploiter pleinement les informations disponibles. L'analyse combinée du regroupement des galaxies, des vitesses particulières, et d'autres sondes (comme le lentillage gravitationnel faible), permettra de mesurer précisément le taux de croissance des structures. Il sera alors possible de tester rigoureusement différents modèles cosmologiques, approfondissant ainsi notre compréhension de l'énergie sombre, de la gravité et des lois de notre Univers.

Alors que j'écris les dernières lignes de ma thèse, il est difficile pour moi de savoir si ces trois années de recherche auront un impact sur notre compréhension de l'Univers (j'en doute un peu). Cependant, si j'ai réussi à vous convaincre de ma capacité à contribuer à la recherche scientifique, alors cela en aura valu la peine.

Table des matières

Affidavit 2

Liste de publications et participation aux conférences 3

Remerciements 4

Abstract and keywords 6

Résumé et mots clés 7

Résumé substantiel en français 8
0.1 Mesure jointe des oscillations baryoniques acoustiques . . . 10
0.2 Inférence par simulation . . . 12
0.3 Analyse combinée des galaxies et des vitesses particulières . . . 13
0.4 Relations de Tully-Fisher et du plan fondamentale dans la simulation hydrodynamique MilleniumTNG . . . 15

Table des matières 19

Introduction 22

1 The standard model of cosmology 24
1.1 The Universe as we see it . . . 25
1.2 The homogeneous Universe . . . 26
1.2.1 General Relativity . . . 26
1.2.2 The homogeneous isotropic expanding Universe . . . 27
1.2.3 Cosmological distances . . . 28
1.2.4 Evolution of the Universe . . . 29
1.3 Non-homogeneous Universe : Linear Perturbations . . . 32
1.3.1 Inflation . . . 34
1.3.2 Boltzmann's equations . . . 35
1.3.3 Perturbed Einstein's equations . . . 38
1.3.4 Cosmic Microwave Background . . . 38
1.3.5 The matter power spectrum : evolution of perturbations . . . 40

2 Galaxy surveys : mapping the large-scale structure of the Universe 44
2.1 Modelling the clustering of galaxies . . . 45
2.1.1 Galaxies as tracers . . . 45
2.1.2 Cosmological probes . . . 46
2.1.3 Linear modelling of the two-points statistics . . . 54
2.1.4 Beyond linear theory . . . 61
2.1.5 Cosmological simulations . . . 62

2.2 Galaxy and peculiar velocity surveys . . . 65
2.2.1 Construction of 3D galaxy maps . . . 65
2.2.2 Measuring Galaxy peculiar velocities . . . 68
2.2.3 Two-point statistics estimators . . . 71
2.2.4 Observational effects . . . 77
2.3 Cosmological Inference . . . 78
2.3.1 Likelihood analysis . . . 78
2.3.2 Fisher forecast . . . 81
2.3.3 Covariance matrix . . . 82
2.3.4 Optimisation algorithms . . . 84

3 Baryon acoustic oscillations from the large-scale clustering of galaxies in Fourier and configuration space 86
3.1 Dataset . . . 88
3.1.1 The eBOSS survey . . . 88
3.1.2 Mock catalogues . . . 88
3.2 Methodology . . . 89
3.2.1 Measuring the clustering . . . 89
3.2.2 BAO modelling . . . 89
3.2.3 Parameter inference . . . 90
3.2.4 Consensus via Gaussian approximation . . . 91
3.2.5 Joint analysis . . . 92
3.3 Results on mock catalogues . . . 94
3.3.1 Fits on average correlations . . . 95
3.3.2 Varying the number of mocks . . . 96
3.3.3 Fits on individual mocks . . . 97
3.3.4 Including the hexadecapole . . . 103
3.4 Application to eBOSS DR16 data . . . 103
3.5 Conclusion . . . 105
Appendix B : pre-reconstruction results . . . 106
Appendix C : modelisation of the cross correlation pattern . . . 107

4 Simulation based inference for galaxy clustering 110
4.1 Emulating clustering from N-body simulations . . . 112
4.1.1 The AbacusSummit Suite of simulations . . . 112
4.1.2 Halo occupation distribution model . . . 113
4.1.3 Observable of interest : The galaxy correlation function . . . 116
4.2 Multiscale Gaussian process model . . . 118
4.2.1 Standard GP regression model . . . 118
4.2.2 Multiscale GP model . . . 124
4.2.3 Application to galaxy clustering . . . 126
4.3 Emulator performance . . . 127
4.3.1 Prediction accuracy . . . 127
4.3.2 Robustness to cosmic variance . . . 128
4.4 Cosmology recovery and constraints . . . 131
4.4.1 Parameter inference using the emulator model . . . 131
4.4.2 Cubic boxes : average fit . . . 133

4.4.3 Cubic boxes : individual fits . . . 134
4.4.4 Observational effects : wide angle and Alcock-Paczynski . . . 137
4.5 Conclusion . . . 142
Appendix A : Emulator covariance choice . . . 144
Appendix B : AP effect on IGP fits . . . 146

5 Combining non-linear peculiar velocity and galaxy clustering 148
5.1 Peculiar velocity and galaxy clustering from N-body simulations . . . 150
5.1.1 Observable . . . 150
5.1.2 Realistic peculiar velocity catalogue . . . 152
5.2 Building the emulators . . . 156
5.2.1 Multiscale Gaussian process . . . 156
5.2.2 Prediction accuracy and precision . . . 157
5.3 Parameters recovery . . . 160
5.3.1 Inference strategy . . . 160
5.3.2 Results . . . 160
5.4 Inference on noisy measurements . . . 164
5.4.1 Constraining power . . . 164
5.4.2 Varying the minimum fitting scale . . . 166
5.5 Conclusion . . . 168
Appendix A : Emulator dynamical range . . . 170
Appendix B : Covariance template scaling . . . 174
Appendix C : Recovery mocks best-fitted parameters . . . 176
Appendix D : Realistic mocks best-fitted parameters . . . 177

6 Tully-Fisher and fundamental plane peculiar velocity measurements in the MilleniumTNG hydrodynamical simulation 179
6.1 Mock galaxy catalogue . . . 181
6.1.1 The Millennium TNG simulation . . . 181
6.1.2 Sample selection . . . 182
6.2 Peculiar velocity measurement . . . 184
6.2.1 Tully-Fisher and fundamental plane relations . . . 184
6.2.2 Peculiar velocity clustering . . . 186
6.3 Dependency of residuals with environment . . . 188
6.3.1 Properties dependency . . . 189
6.3.2 Extended TF and FP relations . . . 192
6.4 Star formation history . . . 193
6.5 Conclusion . . . 195

Conclusion and prospects 197

Bibliographie 200

Introduction

Cosmology is the study of the origin, evolution, and fate of our Universe as a whole. Of course, our conception of the Universe as 'everything that exists' has varied through the epochs. As our understanding of the vastness of the Universe grew, cosmology emerged as a distinct scientific discipline in the early 20th century. The development of Albert Einstein's theory of general relativity notably provided a framework to describe the large scales of our Universe. In the 1920s, Hubble and Lemaître discovered that our galaxy is not the largest structure within the Universe. Many more galaxies exist beyond our own, and they are receding from us with speeds proportional to their distances, indicating an expanding Universe. Perhaps the most important discovery in cosmology is the measurement of the cosmic microwave background radiation in the 1960s, which provided strong evidence for the Big Bang Theory. In the late 1990s, observational evidence of the accelerated expansion of the Universe, theoretically described by Dark Energy, provided the final ingredient for the prevailing standard model of cosmology. The commonly called ΛCDM (for Lambda and cold dark matter) model has proven to be robust and consistent with a wide range of observations, constraining its governing parameters with high accuracy. While the standard model precisely describes the content and the evolution of structures in our Universe, for the past few decades, significant anomalies and big open questions have persisted. These include understanding the dynamics of the initial inflationary period of the Universe, determining the masses of neutrinos, elucidating the origins of discrepancies between observed and predicted values for the expansion rate and the amplitude of matter fluctuations, and unravelling the nature of Dark Matter and Dark Energy.

With past, ongoing, and future surveys covering more volume, spanning higher redshift and multiple wavelengths, cosmology has already entered its *golden age* for quite some time now. The potential to definitively rule out the ΛCDM model in favour of a superior one may lie within untapped data. The challenge lies in effectively managing, analysing, and extracting meaningful insights from these datasets, by refining our analysis strategy, developing theoretical modelling, statistical tools, and addressing systematic errors. This thesis is dedicated to the study of the large scale structure of the Universe to constrain Dark Energy and alternative gravity models. My primary focus is on the distribution of galaxies, reflecting how matter clusters under the influence of gravity and Dark Energy, as well as various baryonic processes and galaxy formation physics. I have explored different probes and observables, employing machine learning tools to develop new methods and improve our constraints and understanding on the expansion and growth of the Universe.

This manuscript is organised as follows. In chapter 1, I introduce the standard model of cosmology and our current state of knowledge about the Universe content and the formation and evolution of large scale structures. In chapter 2, I describe the formalism of galaxy clustering, the construction of galaxy maps, and the different observables and theoretical predictions compared using statistical tools to constrain cosmological parameters. In chapter 3, I focus on baryon acoustic oscillations imprinted in the large scales of the galaxy distribution, which act as a standard ruler tracing the Universe' evolution history. I compare different methods to balance

the data compression and information loss. I describe the simulation based inference framework in chapter 4, introducing a new emulation algorithm to model the small scale clustering and improve constraints on the growth rate of structures. In chapter 5, I study the statistical gain obtained when combining the non-linear clustering of galaxies and peculiar velocities. Finally, in chapter 6, using state-of-the-art hydrodynamical simulation, I expose systematic errors affecting the measurements of galaxies' peculiar velocities and biasing their clustering on the non-linear scales previously considered.

1 The standard model of cosmology

Summary

1.1 The Universe as we see it . . . 25
1.2 The homogeneous Universe . . . 26
1.2.1 General Relativity . . . 26
1.2.2 The homogeneous isotropic expanding Universe . . . 27
1.2.3 Cosmological distances . . . 28
1.2.4 Evolution of the Universe . . . 29
1.2.4.1 Friedmann's equations . . . 29
1.2.4.2 Energy content . . . 29
1.3 Non-homogeneous Universe : Linear Perturbations . . . 32
1.3.1 Inflation . . . 34
1.3.2 Boltzmann's equations . . . 35
1.3.3 Perturbed Einstein's equations . . . 38
1.3.4 Cosmic Microwave Background . . . 38
1.3.5 The matter power spectrum : evolution of perturbations . . . 40

This chapter introduces the standard model of cosmology and the important quantities that we will use throughout this work. The main calculations are based on DODELSON et SCHNEIDER 2013, LESGOURGUES et PASTOR 2006, and BERNARDEAU et al. 2002. We first provide an overview of our current knowledge.

1.1 The Universe as we see it

According to the standard model of cosmology, here is our current state of knowledge about the Universe history. Space and time began with a singularity without any physical description today : the Big Bang. At this initial moment, the Universe was in its densest and hottest state, where every existing elementary particles were entangled in a primordial plasma. Before the first second had passed, the Universe underwent a brutal inflationary phase, stretching space and the quantum fluctuations within it to cosmological scales before abruptly slowing down its expansion. As the universe expands and cools down, protons and neutrons began to form from elementary particles. Within the first few minutes, Big Bang nucleosynthesis occurred, protons and neutrons bound together in an equilibrium state, leading to the formation of atomic nuclei of light elements. As the Universe remained extremely hot and dense, baryonic matter and radiation were coupled in a plasma state, continuously scattering off each other, This state persisted for about 380,000 years until the Universe expanded and cooled sufficiently for charged electrons and protons to combine into neutral atoms, a period known as recombination. At this point, photons decoupled and began to travel freely through the Universe, forming what is known as the Cosmic Microwave Background. Following recombination, the Universe entered a "dark age" where matter continued to cool and clump together under gravity, gradually forming over-dense regions around the primordial fluctuations that kept evolving. After a few hundred million years, the accretion of gas and dust formed the first stars, initiating the creation of heavier elements (stellar nucleosynthesis). Over time, stars and gas clump to form galaxies, and progressively reionized the Universe with energetic radiation. As time passed, the Universe continued to expand, and galaxies cluster under gravity to form large-scale structures made of filament and clusters : the largest bound structures in the Universe. Approximately 9 billion years after the Big Bang, somewhere among the Laniakea supercluster (collection of clusters), inside the Virgo cluster, within the Milky Way galaxy, our solar system and planet Earth are formed. Concurrently, the expansion of the Universe starts accelerating, speeding up the separation of non gravitationally bounded structures. Almost 14 billion years after the beginning, we point our telescopes at the sky, collecting photons that reach us from the closest stars to the most ancient light, and draw this story for our Universe.

1.2 The homogeneous Universe

1.2.1 General Relativity

Albert Einstein's theory of gravitation, known as General Relativity(GR) is a key part of the Standard Model of cosmology. It relies on the following principles :

1. Equivalence Principle : the result of any non-gravitational experiment is independent of its position in space and time.
2. Weak equivalence principle : the trajectory of any particle is independent of its internal structure.
3. Local Lorentz invariance : the outcome of any non-gravitational experiment remains the same regardless of the laboratory's motion when it's in free fall.

Under these hypotheses, gravity can be described not as a force between masses, but as the manifestation of a curved space-time, characterised by a metric $g_{\mu\nu}$ defined as

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \tag{1.1}$$

where ds is the spacetime interval between two events and dx^μ represents the infinitesimal displacement in the μ-th coordinate. Under this notation, we sum over repeated indices which run from 0 (the time dimension) to 3. In curved space geometry, particles free from all influences other than gravitation follow geodesics. Those trajectories are defined such that they maximise the proper time, defined as $c^2 d\tau^2 = -ds^2$. Following the principle of least action on $S = \int g_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} d\tau$, we obtain the geodesic equation

$$\frac{d^2 x^\alpha}{d\tau^2} + \Gamma^\alpha_{\mu\nu} \frac{dx^\mu}{d\tau} \frac{dx^\nu}{d\tau} = 0 \tag{1.2}$$

where $\Gamma^\alpha_{\mu\nu}$ are the Christoffel symbols of the second kind. They are defined as

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2} g^{\alpha\beta} \left(\partial_\mu g_{\beta\nu} + \partial_\nu g_{\mu\beta} - \partial_\beta g_{\mu\nu} \right) \tag{1.3}$$

where ∂_μ represents the partial derivative with respect to the μ-th coordinate, and $g^{\alpha\beta}$ denotes the components of the inverse metric tensor.

The dynamical evolution of the space-time metric we can be described using the action

$$S = \frac{c^4}{16\pi G_N} \int d^4x \sqrt{-g} \, (R - \Lambda) + \int d^4x \sqrt{-g} \mathcal{L}_m, \tag{1.4}$$

with G_N Newton's gravitational constant, c the speed of light in vacuum, g the determinant of the metric, and Λ the cosmological constant also known as Dark Energy. $\mathcal{L}_m$ is the matter Lagrangian density, The Ricci scalar R is a functional of the metric, defined by contracting the indices of the Ricci tensor $R_{\mu\nu}$ with the metric tensor $g^{\mu\nu}$ as $R = g^{\mu\nu} R_{\mu\nu}$. The Ricci tensor is defined in terms of second derivatives of the metric as :

$$R_{\mu\nu} = \partial_\alpha \Gamma^\alpha_{\mu\nu} - \partial_\nu \Gamma^\alpha_{\mu\alpha} + \Gamma^\alpha_{\beta\alpha} \Gamma^\beta_{\mu\nu} - \Gamma^\alpha_{\beta\nu} \Gamma^\beta_{\mu\alpha}. \tag{1.5}$$

The distribution of energy, momentum, and stress within spacetime is encapsulated in the energy momentum tensor as :

$$T_{\mu\nu} = -\frac{2}{\sqrt{-g}} \frac{\delta(\sqrt{-g}\mathscr{L}_m)}{\delta g^{\mu\nu}}, \tag{1.6}$$

Minimising the action 1.4, we obtain the Einstein field equations

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G_N}{c^4} T_{\mu\nu}, \tag{1.7}$$

with $G_{\mu\nu}$ the Einstein tensor defined as

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}. \tag{1.8}$$

The Einstein field equations describe the fundamental interaction of gravitation with the components filling the Universe, as a result of spacetime being curved by mass and energy. Finally, we can write the conservation of energy and momentum in a gravitational field as

$$\nabla^\mu T_{\mu\nu} = 0 \tag{1.9}$$

with ∇^μ the covariant derivative, accounting for the curvature of the space-time manifold. This ensures that the total energy and momentum of the system, including both matter and any contributions from the gravitational field, are conserved as the system evolves in spacetime.

1.2.2 The homogeneous isotropic expanding Universe

Due to the finite speed of light, the observable Universe is delimited by the so called particle horizon. This horizon represents the maximum distance from which light could have travelled to us since the Big Bang. Hence, the standard model of cosmology is built only from the observation of a limited portion of the Universe, and therefore relies on several assumptions. On large scales, the Universe is homogeneous. This hypothesis coupled with the Copernican principle -which state that we do not occupy a privileged position in the Universe - leads to the cosmological principle : the Universe is isotropic and homogeneous.

In the 1920s, Hubble and Lemaître observed that galaxies were receding from us, with speeds proportional to their distances. This strong evidence led to the conclusion that we do not live in a static but expanding Universe. The simpler metric describing an expanding Universe which satisfies the cosmological principle is the Friedmann-Lemaître-Robertson-Walker (FLRW) metric defined (in spherical coordinates) as

$$ds^2 = -dt^2 + a(t)^2 \left[\frac{dx^2}{1 - Kx^2} + x^2(d\theta^2 + \sin^2\theta \, d\phi^2) \right] \tag{1.10}$$

where x, θ and ϕ are the radial and angular coordinates, t is the cosmic time, $a(t)$ is the scale factor, describing how distances between objects change with time. Note that the physical distance is $r = a(t)x$, while x is the comoving distance and holds the distance dimension as $a(t)$ is dimensionless. K is the curvature parameter, determining the global spatial curvature of the Universe. $K = 0$ corresponds to a flat Universe, $K > 0$ a closed Universe, and $K < 0$ an open Universe. We define the Hubble function $H(t) = \frac{1}{a}(t) \frac{da}{dt}(t)$ to describe the expansion rate of the

Universe as a function of cosmic time.

One major consequence of an expanding Universe is the redshift of light. Consider that a galaxy located at a distance x_e, emits a photon at time t_e. For an observer located at $x = 0$, the photon follows the null geodesic $ds^2 = 0$ with a radial path ($d\theta = d\phi = 0$), and reaches the observer at time t_0. If the photon is emitted with a wavelength λ_e, the next crest of the wave is emitted at time $\delta t_e = \frac{\lambda_e}{c}$ (with $\delta t_e \ll \frac{1}{H(t_e)}$), and reaches the observer at time $t_0 + \delta t_0$ such that the observed wavelength is $\lambda_0 = \frac{c}{\delta t0}$.
Following Eq 1.10 we can write

$$x_e = \int_{t_e}^{t_0} \frac{c\,dt}{a(t)} = \int_{t_e+\delta t_e}^{t_0+\delta t_0} \frac{c\,dt}{a(t)}, \tag{1.11}$$

leading to

$$\int_{t_e}^{t_e+\delta t_e} \frac{c\,dt}{a(t)} = \int_{t_0}^{t_0+\delta t_0} \frac{c\,dt}{a(t)}, \tag{1.12}$$

and finally

$$\frac{a(t_0)}{a(t_e)} = \frac{\delta t_0}{\delta t_e} = \frac{\lambda_0}{\lambda_e} = 1 + z(t_e), \tag{1.13}$$

where the redshift $z(t_e)$ represents the stretching (time dilation) of a photon emitted at time t_e. This effect causes the light of distant objects to be shifted to larger wavelength (redder for visible light), and is observable through the shift of emission or absorption features. In the following we set $a(t_0) = 1$.

1.2.3 Cosmological distances

Let us recast the FLRW metric using the conformal time η defined as $d\eta = \frac{dt}{a}$, and rewriting $d\chi^2 = \frac{dx^2}{\sqrt{1-Kx^2}}$ such that

$$ds^2 = a(\eta)^2 \left[-d\eta^2 + d\chi^2 + S_K(\chi)^2(d\theta^2 + \sin^2\theta\, d\phi^2)\right] \tag{1.14}$$

with

$$S_K(\chi) = \begin{cases} K^{-1/2}\sin\left(\sqrt{K}\chi\right) & \text{if} K > 0 \\ \chi & \text{if} K = 0 \\ (-K)^{1/2}\sinh\left(\sqrt{-K}\chi\right) & \text{if} K < 0. \end{cases} \tag{1.15}$$

Using this metric parametrisation, we can define different cosmological distances while making minimum assumption about the curvature. The radial comoving distance of an object at redshift z can be defined as

$$\chi(z) = \int_0^z \frac{c\,dz'}{H(z')}. \tag{1.16}$$

The angular diameter distance D_A relates the physical size -perpendicular to the line of sight- of an object with the corresponding solid angle ($d\Omega^2 = d\theta^2 + \sin^2\theta\, d\phi^2$) it occupies within the observer's field of view. This relation is written $ds^2 = D_A^2 d\Omega^2$. Using Eq 1.14 gives

$$D_A(z) = \frac{S_K\left(\chi(z)\right)}{1+z}. \tag{1.17}$$

For a source located at a radial comoving distance $\chi(z)$ from the observer, the energy of emitted photon decreases as a^{-1}, and $dt \propto a d\eta$, the luminosity $L = \frac{dE}{dt}$ scales like $(1+z)^{-2}$. The flux measured by the observer is then given by

$$F = \frac{L(1+z)^{-2}}{4\pi S_K^2\big(\chi(z)\big)} \tag{1.18}$$

and we can define the luminosity distance as

$$D_L(z) = (1+z) S_K\big(\chi(z)\big). \tag{1.19}$$

1.2.4 Evolution of the Universe

1.2.4.1 Friedmann's equations

Setting $c = 1$, for a homogeneous and isotropic Universe filled with a perfect fluid, the general form of the stress-energy tensor is

$$T_{\mu\nu} = (\rho + P) u_\mu u_\nu + P g_{\mu\nu} \tag{1.20}$$

with ρ the energy density of the fluid, P the pressure of the fluid and u_μ the 4-velocity vector of the fluid normalised as $u_\mu u^\mu = -1$. Combined with the Einstein equation 1.7 we can derive the Friedmann's equations

$$\begin{aligned} H^2 &= \frac{8\pi G_N}{3}\rho - \frac{K}{a^2} + \frac{\Lambda}{3} \\ \frac{\ddot{a}}{a} &= -\frac{4\pi G_N}{3}(\rho + 3P) + \frac{\Lambda}{3} \end{aligned} \tag{1.21}$$

where $\dot{} = \frac{d}{dt}$ stands for the derivative with respect to the cosmic time t. From those equations we can derive the energy conservation equation

$$\dot{\rho} = -3H(\rho + 3P). \tag{1.22}$$

Assuming a general fluid equation of state

$$P = w\rho \tag{1.23}$$

where w is a constant, we find that the homogeneous density of a given fluid evolves as

$$\rho(t) = \rho_0 a(t)^{-3(w+1)}. \tag{1.24}$$

1.2.4.2 Energy content

According to the Standard Model, the Universe is composed of different fluids, all interacting through gravity. In cosmology, protons, neutrons and electrons collectively constitute **baryonic matter**. Because baryons are non-relativistic (in most epochs), the equation of state is pressureless and characterised by $w = 0$. Therefore, the energy density evolves as $\rho_b \propto a^{-3}$. Baryonic matter also interacts electromagnetically via photons, resulting in the emission of light across various wavelengths, rendering it observable. The baryon fluid is also called "Luminous mater".

Photons are relativistic, mass-less particles with an equation of state described by $w = 1/3$. The energy density decreases faster than baryons as $\rho_\gamma \propto a^{-4}$. This can also be understood as the photon fluid energy or temperature cooling down with expansion as $T_\gamma \propto a^{-1}$. In the primordial Universe, photons and baryons were coupled in a hot and dense ionized plasma of charged particles through Thomson scattering. Photons scattered off charged particles, primarily electrons, and could not travel far without being scattered again. However, because of expansion, the Universe became cool enough for atoms to form. This process is known as recombination. As the density of free electron fell, the Thomson scattering rate became smaller than the expansion rate, causing the electron-photon plasma to decouple. This decoupling led to the free streaming of photons. These photons travel from every point in space in all directions (almost) without interacting. Due to cosmic expansion, they get redshifted to lower energy, about 2.7255 ± 0.0006K in average (PLANCK COLLABORATION et al. 2020). This signal continuously reaching us is commonly known as the Cosmic Microwave Background (CMB). Because of the vast number of photons composing it, the CMB represents the majority of the photon energy in the Universe today (HAUSER et DWEK 2001 ; HILL et al. 2018). The decoupling happened around the last scattering redshift measured by the Planck collaboration PLANCK COLLABORATION et al. 2020 to be at $z = 1089$ (around 380000 years after the Big Bang). Although nowadays, the proportion of photons in the total energy content is negligible, early time was a radiation dominated era.

Radiation also included relativistic **neutrinos**, extremely light particles weakly interacting with matter. Due to this weak interaction, neutrinos decoupled from the photons-baryon fluid as soon as the interaction rate went below the expansion rate. This happened slightly before electrons started annihilating with positrons, when the cosmic temperature was of the order of the electron mass. As a consequence, the neutrino fluid inherited almost none of the associated energy, while photons did. Furthermore, while photons follow a Bose-Einstein energy distribution (integer spin), neutrinos follow a Fermi–Dirac distribution (half integer spin). This distinction suppresses the neutrinos-photons energy density ratio by a factor $\frac{7}{8}$. Accounting for both effects, the total relativistic neutrinos' energy density can be expressed as

$$\rho_\nu = N_{\rm eff} \frac{7}{8} \left(\frac{4}{11}\right)^{4/3} \rho_\gamma. \tag{1.25}$$

The effective number of relativistic species $N_{\rm eff}$ factor accounts for the 3 different family of neutrinos in the standard model. However, correction coming from a non-instantaneous decoupling leads to $N_{\rm eff} = 3.046$. The total radiation energy density is usually defined as $\rho = \rho_\gamma + \rho_\nu$. Although at early time neutrinos behave like radiation, because the neutrinos are massive particles, as the Universe expands they eventually become non-relativistic and contribute to the total matter energy density.

Measurement of galaxy rotation curves (ZWICKY 1937) shed light on the presence of unaccounted-for total mass. This was latter confirmed with the measurement of the amplitude of anisotropies in the CMB SMOOT et al. 1992, or the separation of mass components of the Bullet Cluster (CLOWE et al. 2006). Invisible non-baryonic matter, which does not interact electromagnetically but only through gravity, has been introduced to account for this missing mass : **Cold Dark Matter**. Some theories describe Dark Matter as beyond Standard Model particles, interacting only through gravity and weak interactions. This definition is referred as weak interacting massive particles (WIMPs). So far Dark Matter particles have not been detected, and Dark Matter could also be interpreted as an inaccurate description of gravity on cosmic scales. Within the standard cold Dark Matter(non-relativistic when it decoupled from the primordial plasma) formalism, Dark

Matter acts as a pressureless fluid $w = 0$, with an energy density scaling as $\rho_c \propto a^{-3}$. The total matter energy density is usually defined as $\rho_m = \rho_c + \rho_b + \rho_\nu$.

The cosmological constant Λ was originally introduced by Einstein to accommodate his theory with a static Universe. As an additional term in Eq. 1.4, the cosmological constant would counteract the gravitational collapse of a static Universe. However, the cosmological constant fell out of favour after Hubble and Lemaître provided observational evidence of an expanding Universe, naturally explained by the Big Bang Theory. It was only in the late 1990s that observational evidence from Type Ia supernovae led to the conclusion that the expansion of the Universe was accelerating. The cosmological constant was revisited to explain this acceleration. It is commonly referred to as **Dark Energy** and can be interpreted as a perfect fluid filling space homogeneously with the equation of state $w = -1$ and an energy density

$$\rho_\Lambda = \frac{\Lambda}{8\pi G_N}. \tag{1.26}$$

While classical physics does not include such a fluid, quantum physics predicts its existence as vacuum energy. However, the energy density predicted by quantum field theory exceeds by several orders of magnitude what we observe with the cosmological constant. Relaxing the vacuum energy hypothesis as a physical description of Dark Energy, one can explore deviations from a pure cosmological constant. A famous parametrisation for Dark Energy equation of state is the Chevallier-Polarski-Linder(CPL) Evolving Dark Energy (CHEVALLIER et POLARSKI 2001)

$$w(a) = w_0 + (1-a)\,w_a, \tag{1.27}$$

where setting $w_0 = -1$ and $w_a = 0$ we retrieve the cosmological constant. The Dark Energy density then evolves as

$$\rho_\Lambda = \rho_{\Lambda,0}\, a^{-3(1+w_0+w_a)}\, e^{3w_a(a-1)}. \tag{1.28}$$

Although this parametrisation does not provide a physical description, it defines a flexible framework to describe various Dark Energy models beyond the simple cosmological constant, and and their impact on the expansion history of the Universe.

Finally, as we saw in the Friedmann's equations, curvature can be rewritten as the energy density

$$\rho_K = -\frac{3K}{8\pi G_N a^2}. \tag{1.29}$$

It is also useful to define the critical density as

$$\rho_c = \frac{3H^2}{8\pi G_N}, \tag{1.30}$$

corresponding to the energy density required for the Universe to be flat. The energy density of a given fluid X can be defined as fraction of the critical density $\Omega_X = \rho_X/\rho_c$, and we can rewrite the first Friedmann equation as

$$\sum_X \Omega_X + \Omega_\Lambda + \Omega_K = 1. \tag{1.31}$$

With the total neutrino energy density today defined as a function of the total neutrino mass as

$$\Omega_\nu h^2 = \frac{\sum m_\nu}{93.14\text{eV}}. \tag{1.32}$$

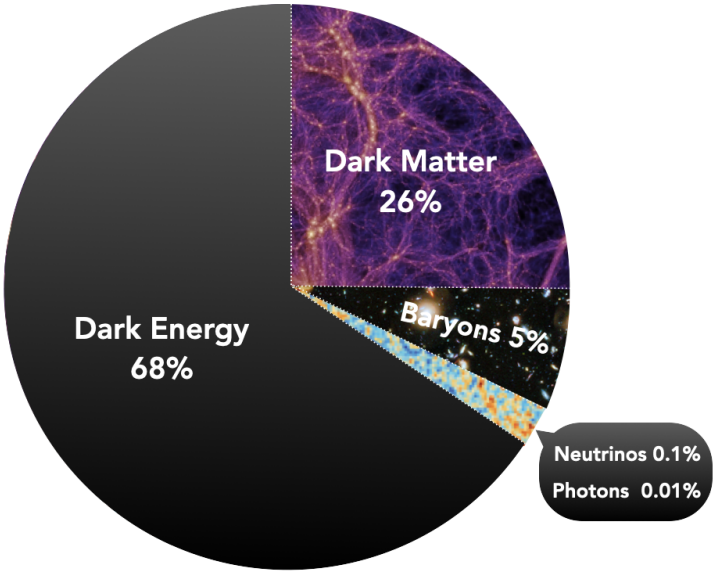


FIGURE 1.1 : Energy content of the Universe according to the Planck results PLANCK COLLABORATION et al. 2020 within the ΛCDM model.

From the Friedmann's equations, we can now derive the Hubble function, describing the expansion rate of the Universe as

$$H(z) = H_0 \sqrt{\Omega_{\mathrm{m},0}(1+z)^3 + \Omega_{\mathrm{r},0}(1+z)^4 + \Omega_{\mathrm{K},0}(1+z)^2 + \Omega_\Lambda} \tag{1.33}$$

with the so-called Hubble constant $H_0 = 100h\mathrm{km}\ s^{-1}\ \mathrm{Mpc}^{-1}$ parametrised with h.
Today, the energy density of each component is pretty well constrained (PLANCK COLLABORATION et al. 2020), with the curvature appearing consistent with a flat Universe. We will maintain this hypothesis throughout this work. The estimated energy content at present time is depicted in figure 1.1. As we also know the redshift evolution of any Ω_X, we can derive the epochs at which each component dominated the energy density. Radiation dominated the Universe until redshift about $z \approx 3400$, when started the Matter domination era. Since $z \approx 0.3$ the Universe is dominated by Dark Energy. Although Dark Matter and Dark Energy does not have a satisfactory physical description yet, the ΛCDM model is able to accurately describe - from few assumptions - the content and evolution of the homogeneous background of our Universe.

1.3 Non-homogeneous Universe : Linear Perturbations

Observed structures such as galaxies, voids, and stars indicates that our Universe is not homogeneous on small scales (typically below 100Mpc). However, considering non-homogeneous energy density for the various components of the Universe makes the dynamical equations analytically unsolvable. To address this, one can introduce small perturbations to the linear order on top of the homogeneous background, enabling analytical treatment of the system's differential equations. The density of a given fluid X at time t and position $\boldsymbol{x}$ can then be written

$$\rho_X(\boldsymbol{x}, t) = \bar{\rho}_X(t)\,(1 + \delta(\boldsymbol{x}, t))\,, \tag{1.34}$$

where $\bar{\rho}_X(t)$ is described by the homogeneous cosmology, and we will consider the perturbation δ to be linear (neglecting orders higher than $O(\delta^2)$). As we will see in the next section, those perturbations are sourced by a quantum field, therefore δ is a stochastic field and cannot be directly predicted. It can only be studied statistically. Following cosmological principle, one can write the expectation value of the perturbation as $\langle\delta\rangle = 0$. In the following we will apply ergodic principle : for a given quantity, the expectation value obtained by averaging over many realisations is equivalent to the one obtained by averaging over a large volume. As we only have one Universe to observe, this is essential. Due to our limited access to a finite volume, particularly at low redshift, averaged quantities may fluctuate around their expected values. This phenomenon is known as cosmic variance, which we will discuss in further detail in the next section. If we assume that the random field is Gaussian, its probability distribution is fully characterised by its two-point function (second moment). According to *Wick's theorem*, the expectation value with any odd number of Gaussian fields vanishes, while the expectation value with any even number of fields is fully characterised by the two-point function. It is also known as the correlation function, and is defined as :

$$\xi(\boldsymbol{s}) = \langle\delta(\boldsymbol{x})\delta(\boldsymbol{x}+\boldsymbol{s})\rangle. \tag{1.35}$$

It can be physically interpreted as the statistical excess of correlation - compared to a random distribution - between the field values at two points. Following the statistical homogeneity assumption, ξ is only a function of the separation $\boldsymbol{s}$ between two locations. Moreover, assuming isotropy, ξ only depends on the norm $s = |\boldsymbol{s}|$. We define the Fourier transform of the perturbation as

$$\delta(\boldsymbol{k}) = \int \delta(\boldsymbol{x})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\, d^3x. \tag{1.36}$$

Because $\delta(\boldsymbol{x})$ is real, its complex conjugate is given by $\delta^*(\boldsymbol{k}) = \delta(-\boldsymbol{k})$. The two-point function in Fourier space is the power spectrum defined as

$$\left\langle\delta(\boldsymbol{k})\delta^*(\boldsymbol{k}')\right\rangle = (2\pi)^3\delta_D\left(\boldsymbol{k}+\boldsymbol{k}'\right)P(\boldsymbol{k}). \tag{1.37}$$

The power spectrum is the Fourier transform of the correlation function

$$P(\boldsymbol{k}) = \int \xi(\boldsymbol{s})\, e^{-i\boldsymbol{k}\cdot\boldsymbol{s}} d^3s. \tag{1.38}$$

For an isotropic field, the power spectrum only depends on the wave vector norm $k \equiv |\boldsymbol{k}|$. Using the spherical volume element $d^3k = k^2 dk d\theta \sin\theta d\phi$ and $\boldsymbol{k}\cdot\boldsymbol{s} = ks\cos\theta$, we can derive the relation

$$\xi(s) = \int \frac{dk}{2\pi^2} k^2 P(k) \frac{\sin(ks)}{ks} \tag{1.39}$$

We will also use the velocity field $\boldsymbol{u}(\boldsymbol{x})$ as a perturbed quantity with zero mean, along its corresponding two-point statistics. The goal of this section is to describe how such perturbations appeared and evolved in the (non)homogeneous Universe. First, we need to derive some initial conditions (IC) for those perturbations.

1.3.1 Inflation

The theory of inflation was first introduced to resolve two inconsistencies within the standard model : the horizon and flatness problem. Observations of the CMB have shown that the temperature of photons travelling since recombination was remarkably uniform across the sky. Thus, suggesting that even regions widely separated were once in thermal equilibrium. However, some regions are so separated that since the Big Bang, light did not have time to travel the distance and equalize their temperatures. The second issue is that today, on large scales, the Universe seems to be almost perfectly flat. From equation 1.33 one can see that the curvature must have been unreasonably small at $z \to \infty$ which is considered as unphysical fine-tuning. Inflation solves both of these problems by providing a mechanism for the extremely rapid expansion of the Universe at primordial times, making the early value of Ω_K more natural, and allowing distant regions to come into causal contact before inflation takes place.

The simplest possibility to generate such a transitory epoch of accelerated expansion is via the potential energy of a scalar field with a negative pressure : the inflaton. The action for a scalar field $\phi(t)$ evolving in a potential $V(\phi)$ can be written

$$S = \frac{1}{16\pi G_N}\int d^4x\sqrt{-g}\left[-\frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi)\right]. \tag{1.40}$$

Minimising this action with respect to the scalar field, we find the equation of motion

$$\ddot{\phi} + 3H\dot{\phi} + V' = 0 \tag{1.41}$$

with $V' = \frac{\partial V}{\partial\phi}$. From Eq 1.6 we can also derive the stress-energy tensor and read off the density and pressure such that we can write the equation of state of the field

$$w = \frac{P}{\rho} = \frac{\dot{\phi}^2 - V(\phi)}{\dot{\phi}^2 + V(\phi)}. \tag{1.42}$$

Note that $\dot{\phi} = 0$ corresponds to the cosmological constant solution, which is not suited as we need the acceleration to stop at some point. Neglecting the curvature, we rewrite Friedmann's equations

$$\begin{aligned} H^2 &= \frac{8\pi G_\mathrm{N}}{3}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) \\ \frac{\ddot{a}}{a} &= \frac{8\pi G_\mathrm{N}}{3}\left(V(\phi) - \dot{\phi}^2\right). \end{aligned} \tag{1.43}$$

In order to produce an accelerated expansion $\ddot{a} > 0$, we need $w < -1/3$, so the field should be slowly varying such that

$$\dot{\phi}^2 \ll V(\phi) \quad \text{and} \quad \ddot{\phi} \ll 3H\dot{\phi}, \tag{1.44}$$

defining the slow-roll regime. From the above equations, we can define the slow roll parameters

$$\epsilon_V \equiv \frac{M_p^2}{2}\left(\frac{V'}{V}\right)^2, \quad \eta_V \equiv M_p^2\frac{V''}{V}, \quad \xi_V^2 \equiv M_p^4\frac{V'V'''}{V^2} \tag{1.45}$$

such that $\epsilon_V \ll 1$, $\eta_V \ll 1$ and $\xi_V \ll 1$, with $M_p = \sqrt{\hbar c/G}$ the Planck mass, with $\hbar$ the reduced Planck's constant. This indicates that the potential must be very flat.

It was soon realized that inflation also provided "for free" a natural mechanism for generating

the seeds of primordial fluctuations. By allowing the inflaton field to have some quantum fluctuations over the homogeneous background, the inhomogeneities would quickly be stretched over cosmological distances by the brutal expansion, and would grow to form the observed structures. Note that Although it is not discussed here, tensor perturbations would be susceptible to produce primordial gravitational waves. Let us decompose the field as

$$\phi(\boldsymbol{x}, t) = \bar{\phi}(t) + \delta\phi(\boldsymbol{x}, t), \tag{1.46}$$

such that $\delta\phi(\boldsymbol{x}, t)$ is a random adiabatic Gaussian field. Expanding the previous equations at linear order in $\delta\phi$, we can derive the equation of motion at linear order in Fourier space

$$\ddot{\delta\phi} + 3H\dot{\delta\phi} + \frac{k^2}{a^2}\delta\phi + V''(\delta\phi) = 0, \tag{1.47}$$

with k the wave number. Because of gravity, the quantum perturbations in the inflaton field will generate perturbations in the gravitational potential, while the inflaton field will decay into Standard Model particles (through unknown chain of reactions). The equation of motion can be directly solved for the curvature perturbation $\mathscr{R}$, which is defined as $\mathscr{R} = \frac{-aH}{\dot{\phi}'}\delta\phi$ during inflation. $\mathscr{R}$ has the property to be conserved during inflation for any mode outside the horizon $\eta k \ll 1$, making it possible to connect the IC in the early Universe to what was generated during inflation. At the end of inflation, the inflaton field rapidly decays and the energy stored in the field is transferred into particles and radiations. When radiation dominates, the curvature perturbations source perturbations in the gravitational field, that we call Ψ such as $\mathscr{R} = -\frac{3}{2}\Psi$. The power spectrum of the curvature perturbation is then given by

$$P_{\mathscr{R}}(k) = 2\pi^2 A_s k^{-3} \left(\frac{k}{k_p}\right)^{n_s - 1 + \frac{1}{2}\alpha_s \ln(k/k_p)} \tag{1.48}$$

where A_s is the variance of curvature perturbations around the pivot scale k_p. The pivot scale is a matter of convention, and is often determined as the scale best constrained by a given set of observations, here $k_p = 0.05$ Mpc^{-1} as in the Planck collaboration PLANCK COLLABORATION et al. 2020. n_s is the spectral index and α_s is the running of the spectral index defined as (PELOSO et al. 2014) :

$$n_s - 1 = 2\eta_V - 6\epsilon_V \quad , \quad \alpha_s = -2\xi_V^2 + 8\epsilon_V\,(n_s - 1) + 24\epsilon_V^2 \tag{1.49}$$

Planck collaboration measured $ns = 0.965 \pm 0.004$ indicating a deviation from scale invariance, and $\alpha_s = 0.0011 \pm 0.0099$ consistent with zero. An important consequence of adiabatic IC is that although different patches of the Universe have different overdensities, the fractional density perturbations are the same for all species (and analogously for velocities). Therefore, for two species i and j we can write $\frac{\delta_i}{1+w_i} = \frac{\delta_j}{1+w_j}$. This sets up initial condition for radiation, Dark Matter, baryons and neutrinos perturbations.

1.3.2 Boltzmann's equations

As mentioned before, the initial quantum perturbations propagate to the gravitational potential because of the coupling with gravity. We can write a corresponding diagonal perturbed metric

with two scalar functions Ψ and Φ, corresponding to the time and space gravitational potentials :

$$ds^2 = a^2(\eta)\left[-(1+2\Psi)d\eta^2 + (1+2\Phi)\delta^D_{ij}dx^i dx^j\right]. \tag{1.50}$$

This is also known as the Newtonian Gauge and is equivalent to fixing a coordinate system. Those gravitational perturbations will couple with the density perturbations for all species in the Universe, and will evolve with time. In order to describe the statistical behaviour of these perturbations, we introduce the distribution function $f(\boldsymbol{x}, \boldsymbol{p}, t)$, describing the number of particles at a given time around a phase-space element $(\boldsymbol{x}, \boldsymbol{p})$, with $\boldsymbol{p}$ the physical momenta of the particles. Taking the three first moments of this function we can write

$$\begin{aligned} n(\boldsymbol{x}, t) &= \int \frac{d^3p}{(2\pi)^3} f(\boldsymbol{x}, \boldsymbol{p}, t), \\ u_i(\boldsymbol{x}, t) &= \frac{1}{n(\boldsymbol{x}, t)} \int \frac{d^3p}{(2\pi)^3} f(\boldsymbol{x}, \boldsymbol{p}, t) \frac{p\hat{p}_i}{E}, \\ u_i(\boldsymbol{x}, t)u_j(\boldsymbol{x}, t) + \sigma_{ij}(\boldsymbol{x}, t) &= \frac{1}{n(\boldsymbol{x}, t)} \int \frac{d^3p}{(2\pi)^3} f(\boldsymbol{x}, \boldsymbol{p}, t) \frac{p^2\hat{p}_i\hat{p}_j}{E^2}, \end{aligned} \tag{1.51}$$

with $n(\boldsymbol{x}, t)$ the number density of particles, $\boldsymbol{u}(\boldsymbol{x}, t)$ the fluid velocity, $\sigma_{ij}(\boldsymbol{x}, t)$ the stress tensor, $E = \sqrt{\boldsymbol{p}^2 + m^2}$ the proper energy and $\hat{p} \equiv \boldsymbol{p}/|\boldsymbol{p}|$ design a unitary vector. The number density can be written $n(\boldsymbol{x}, t) = \bar{n}(t)(1 + \delta(\boldsymbol{x}, t))$. Note that for non-relativistic particles, the energy density is simply obtained by scaling by the particles mass. Hence, the perturbed quantity δ corresponds to inhomogeneities both in the number and energy density. In the linear regime where fluid evolves in a single coherent flow $\sigma_{ij} = 0$. The fluid velocity is by itself a perturbed (small) quantity. We consider the velocities to be irrotational $\nabla \times u = 0$. Thus, in Fourier space it is purely along the wave vector $\boldsymbol{k}$.

The equation governing the distribution function is the Boltzmann transport equation. From the geodesic equation and the four-momentum definition, it can be written as

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial f}{\partial x^i} - \left[H + \dot{\Phi} + \frac{p}{E}\frac{\hat{p}^i}{a}\frac{\partial \Psi}{\partial x^i}\right] p\frac{\partial f}{\partial p} = C\left[f\right]. \tag{1.52}$$

where $C\left[f\right]$ is a collision term.

In the case of photons, we can write the perturbation to their distribution function in terms of a temperature perturbation $\Theta(\boldsymbol{x}, \hat{p}, t)$ instead of energy density. As it is a relativistic species, it follows a Bose–Einstein distribution with zero chemical potential

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{p}, t) &= \left[\exp\left\{\frac{p}{T(t)[1 + \Theta(\boldsymbol{x}, \hat{p}, t)]}\right\} - 1\right]^{-1} \\ &= f^{(0)}(p, t) - p\frac{\partial f^{(0)}(p, t)}{\partial p}\Theta(\boldsymbol{x}, \hat{p}, t), \end{aligned} \tag{1.53}$$

where linearly expanded in Θ, and the zero-th order of the distribution function is

$$f^{(0)} \equiv \left[\exp\left\{\frac{p}{T}\right\} - 1\right]^{-1}. \tag{1.54}$$

Note that we have assumed that Θ does not depend on the amplitude of the momenta. This is

related to the fact that the photon's momenta are unchanged by Compton scattering. The collision term describing the coupling between photons and baryons through Compton scattering can be written

$$C[f] = -p\frac{\partial f^{(0)}}{\partial p} n_e \sigma_{\rm T} \left[\Theta_0 - \Theta + \hat{p}\cdot \boldsymbol{u}_{\rm b}\right], \tag{1.55}$$

where n_e is the number density of electrons, σ_T is the Thomson cross-section, and we have introduced the monopole of the temperature perturbation Θ_0. The Legendre multipoles are defined as

$$\Theta_l(k,\eta) \equiv \frac{1}{(-i)^l}\int_{-1}^{1}\frac{d\mu}{2}\mathscr{P}_l(\mu)\Theta(\mu,k,\eta), \tag{1.56}$$

with $\mathscr{P}_l(\mu)$ the Legendre polynomials and μ the cosine describing the direction $\mu = \frac{\boldsymbol{k}\cdot\hat{p}}{k}$.

We can predict how Compton scattering will influence the distribution of photons. We can see that if there is no overall velocity for the electrons, when Compton scattering is highly effective Θ tends towards Θ_0. In other words, all anisotropies get suppressed. Solving Eq 1.52 for the photons we get the equation describing the evolution of temperature perturbation in Fourier space

$$\Theta' + ik\mu\Theta = -\Phi' - ik\mu\Psi - \tau'\left[\Theta_0 - \Theta + \mu u_b\right] \tag{1.57}$$

were $f^{(0)}$ got factored out, $'$ denote derivatives with respect to conformal time η, and we have defined the optical depth as

$$\tau' = -n_e\sigma_T a. \tag{1.58}$$

Note that we have neglected the angular dependence of Compton scattering and the coupling with the polarisation field.

Dark Matter is considered as collision-less massive particles. Since it is non-relativistic, characterising its perturbations requires less information. Typically, we do not need an expression for the distribution function, we can integrate Eq 1.52 to get the zero-th and first moments and derive the continuity and Euler equations

$$\begin{aligned} &\delta_c' + iku_c + 3\Phi' = 0 \\ &u_c' + \frac{a'}{a}u_c + ik\Psi = 0. \end{aligned} \tag{1.59}$$

After recombination, electrons and protons remains tightly coupled such that they have a common density $\delta_e = \delta_p = \delta_b$ and velocity $\boldsymbol{u}_e = \boldsymbol{u}_p = \boldsymbol{u}_b$. Because of this tight coupling, we can treat baryons as a non-relativistic fluid. Similarly to Dark Matter we can write

$$\begin{aligned} &\delta_b' + iku_b + 3\Phi' = 0 \\ &u_b' + \frac{a'}{a}u_b + ik\Psi = \frac{\tau'}{R}\left[u_b + 3i\Theta_1\right] \end{aligned} \tag{1.60}$$

with $R = \frac{3\rho_b}{4\rho_\gamma}$. Finally, neutrinos perturbations can be described as a temperature perturbation that we write $\mathscr{N}$. As non-gravitational interactions of neutrino are negligible, one can solve the collision-less Boltzmann's equation for massive particles with a Fermi-Dirac distribution function, and derive

$$\mathscr{N}' + ik\mu\frac{p}{E_\nu(p)}\mathscr{N} - Hp\frac{\partial}{\partial p}\mathscr{N} = -\Phi' - ik\mu\frac{E_\nu(p)}{p}\Psi. \tag{1.61}$$

1.3.3 Perturbed Einstein's equations

To solve this system of differential equations describing the evolution of distributions of the different species, we need to derive two additional equations describing the evolution of the Newtonian potentials. From Einstein field equation and the perturbed metric in the Newtonian gauge, we can derive in Fourier space :

$$\begin{aligned} k^2\Phi + 3\frac{a'}{a}\left(\Phi' - \Psi\frac{a'}{a}\right) &= 4\pi G a^2\left[\rho_{\rm m}\delta_{\rm m} + 4\rho_{\rm r}\Theta_{\rm r,0}\right] \\ k^2(\Phi+\Psi) &= -32\pi G a^2\rho_{\rm r}\Theta_{\rm r,2}. \end{aligned} \tag{1.62}$$

Here, we have included all matter in the subscript m, and all radiation in the subscript r such that

$$\begin{aligned} \rho_{\rm m}\delta_{\rm m} &\equiv \rho_{\rm c}\delta_{\rm c} + \rho_{\rm b}\delta_{\rm b}; \quad \rho_{\rm r}\Theta_{\rm r,0} \equiv \rho_\gamma\Theta_0 + \rho_\nu\mathcal{N}_0 \\ \rho_{\rm m}u_{\rm m} &\equiv \rho_{\rm c}u_{\rm c} + \rho_{\rm b}u_{\rm b}; \quad \rho_{\rm r}\Theta_{\rm r,1} \equiv \rho_\gamma\Theta_1 + \rho_\nu\mathcal{N}_1. \end{aligned} \tag{1.63}$$

The first equation is the Poisson equation written for an expanding Universe. The second equation relates the two Newtonian potentials as a function of the quadrupole of the radiation temperature perturbation, also known as anisotropic stress. Now that the system of equations governing the perturbations is set, with defined IC, it can be solved for density/temperature and velocity perturbations of each species. There are several publicly available Boltzmann codes, such as CAMB[1] and CLASS[2], that we will use to numerically solve those equations.

1.3.4 Cosmic Microwave Background

One of the most important observations in cosmology is the measurement of the CMB and its small temperature anisotropies ($\approx 10^{-5}$) on top of a homogeneous background with a temperature $T = 2.725$K. We can solve the above system of equation for the temperature fluctuation $\Theta(\boldsymbol{x}, \eta, \hat{p})$. Although we can only observe those perturbations at our location and at present time, the anisotropies depend on the direction of the incoming photons. It is useful to decompose into spherical harmonics

$$\Theta(\boldsymbol{x}, \hat{\boldsymbol{p}}, \eta) = \sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{l} a_{\ell m}(\boldsymbol{x}, \eta)\, Y_{\ell m}(\hat{\boldsymbol{p}}), \tag{1.64}$$

so that all of the information is contained in the amplitudes $a_{\ell m}$. By definition the expected value of the anisotropies is zero $\langle a_{\ell m}\rangle = 0$, and as the field is assumed to be Gaussian, its distribution is fully described by its variance

$$\langle a_{\ell m} a^*_{\ell' m'}\rangle = \delta_{\ell\ell'}\delta_{mm'}C(\ell), \tag{1.65}$$

with $C(\ell)$ the angular power spectrum.

[1] https://camb.info/
[2] https://lesgourg.github.io/class_public/class.html

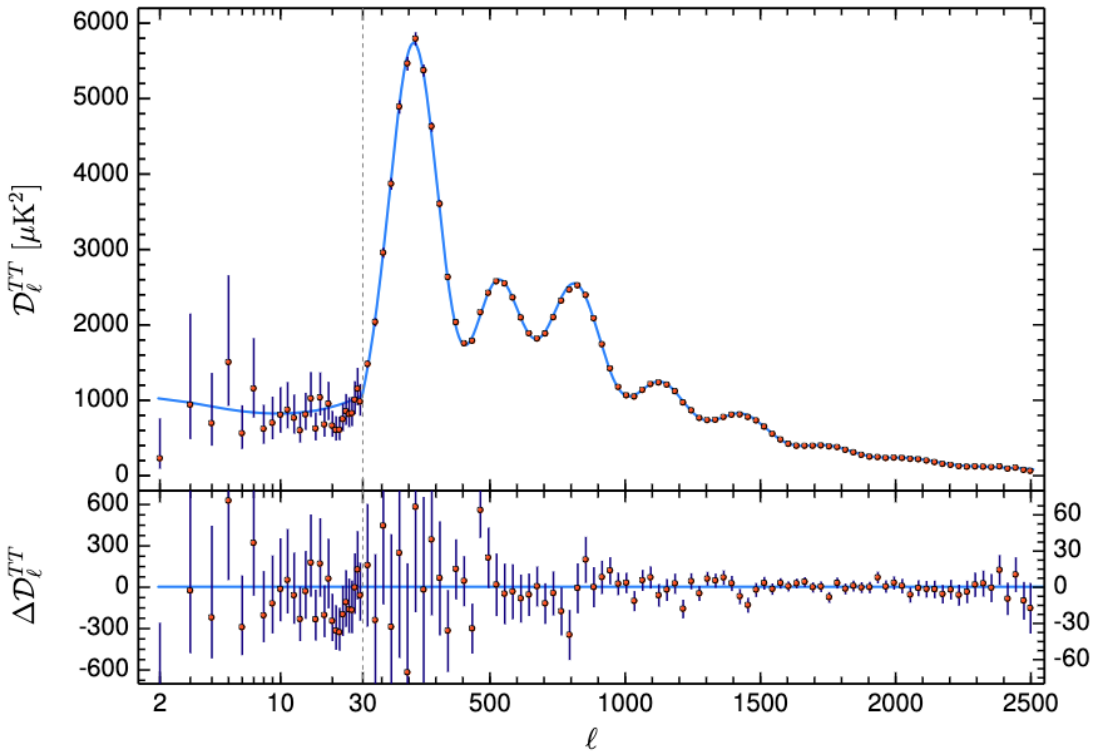


FIGURE 1.2 : The CMB temperature angular power spectrum measured by the Planck collaboration (PLANCK COLLABORATION et al. 2020). The best fit model corresponds to the blue curve plotted best fit model. The horizontal dotted line denotes a switch from logarithmic to linear scale.

The prediction derived for $C(\ell)$ is compared in figure 1.2 to the latest observations performed by the Planck collaboration (PLANCK COLLABORATION et al. 2020). D_ℓ^{TT} is a normalisation defined as $D_\ell^{TT} = \frac{\ell(\ell+1)C(\ell)}{2\pi}$. The red dots correspond to observations while the blue lines shows the best fit. Although we do not show it here, note that the best fit is obtained by using the polarisation fluctuations in addition to the temperature ones. This important result allows to derive precise constraints on the ΛCDM model parameters that we will use as a reference from now on.

There are several interesting physical effects driving the shape of the angular power spectrum, but we will focus on the main oscillation pattern as it will be relevant for this work. As we discussed, before recombination the baryon-photon fluid is in equilibrium through an exchange of momentum caused by Compton scattering. However, because the photon number density is much larger than the baryon number density, the rate of change in the baryon velocity is scaled by the factor $R = \frac{3\rho_b}{4\rho_\gamma}$. Therefore, baryons have a different mean free path with an optical depth $\tau_b' = \tau'/R$. Because decoupling takes place when the optical depth equals one, baryons decouple from photons later than photons decouple from baryons, slightly after recombination. As a consequence, baryons will be dragged for a time by the escaping photon fluid leading to sound wave propagation in the plasma, and leaving an over-dense spherical shell around each initial perturbation. The redshift at which the baryon optical depth equals unity is about $z_\mathrm{d} = 1060$, compared to the last scattering redshift $z^* = 1089$. We can compute the comoving distance

travelled by the sound wave as the horizon at the drag epoch

$$r_{\rm d} = r_s(z_{\rm d}) \equiv \int_0^{\eta_{\rm d}} c_s(\eta) d\eta = \int_\infty^{z_{\rm d}} \frac{c_s(z)}{H(z)} dz, \tag{1.66}$$

with the sound speed in the plasma given by

$$c_s(z) = \sqrt{\frac{1}{3(1+R(z))}}. \tag{1.67}$$

It is evaluated at about $r_{\rm d} = 147$ Mpc by the latest Planck results. This sound wave propagation is described by oscillations in Fourier space, also known as Baryon Acoustic Oscillations (BAO). This feature is imprinted in the matter distribution nowadays as we will study in Chapter 3.

1.3.5 The matter power spectrum : evolution of perturbations

The same initial perturbations generating anisotropies in the CMB during the early Universe also led to the formation of inhomogeneities in the matter field. As the Universe evolves during matter domination era, slightly over-dense regions create gravitational instabilities, aggregating matter over time, and shaping the structures we observe today. Baryonic and Dark Matter perturbations are coupled via gravity, and - once the Universe is dominated by matter - do not depend on radiation perturbations. We neglect the impact of neutrinos and consider baryons and Dark Matter as a single fluid with inhomogeneities δ_m. We are interested in modes growing within the Hubble horizon (sub-horizon modes) such that $\eta k >> 1$. We further assume that in the epochs of interest, there is no anisotropic stress which leads from the second perturbed Einstein equation 1.62 to $\Phi = -\Psi$. In this regime we can write the so called Vlasov-Poisson equations

$$\begin{aligned} &\delta_m' + \nabla \cdot [(1+\delta_m)\boldsymbol{u}_m] = 0, \\ &\boldsymbol{u}_m' + aH\boldsymbol{u}_m + \boldsymbol{u}_m \cdot \nabla \boldsymbol{u}_m = -\nabla\Phi - \frac{1}{\rho_m}\nabla_j \sigma_{ij}, \\ &\Delta\Phi = \frac{3}{2}a^2 H^2 \Omega_m \delta_m. \end{aligned} \tag{1.68}$$

Because at late time perturbations have grown to be non-linear, specially on small scales $\delta_m \sim 1$, we have kept the non-linear terms and the stress tensor for completeness. We recognise the continuity, Euler and Poisson equations describing the non-linear evolution of the matter perturbations. Working at linear order in the perturbations remains useful to get a physical insight of their evolution, and allow to decompose the gravitational perturbations in Fourier space into a scale dependent transfer function $T(k)$ and a time dependent growth factor $D(a)$

$$\Phi(\boldsymbol{k}, a) = \frac{3}{5} T(k) \frac{D(a)}{a} \mathscr{R}(\boldsymbol{k}), \tag{1.69}$$

which propagates to density perturbation trough Poisson equation

$$\delta_m(\boldsymbol{k}, a) = \frac{2}{5} \frac{k^2}{\Omega_{m,0} H_0^2} T(k) D(a) \mathscr{R}(\boldsymbol{k}). \tag{1.70}$$

This convenient decomposition allows to separate early time physics contained in $T(k)$ and $\mathscr{R}(\boldsymbol{k})$ from the scale independent late evolution $D(a)$. Note that both $T(k)$ and $D(a)$ depend on the cosmological parameters. The transfer function describes the evolution of perturbations through the radiation-matter transition and decoupling. It mainly describes two regimes of the gravitational perturbation : the super-Hubble modes ($k\eta \ll 1$) and the sub-Hubble modes ($k\eta >> 1$) at matter-radiation equality. Modes that enter horizon during radiation domination remain constant until matter radiation equality, and are therefore suppressed compared to modes entering horizon during matter domination. Although BAO is more important in radiation - baryons are only a small fraction of total matter - it is also encoded in the matter perturbations through the transfer function. As we mentioned before the publicly available Boltzman codes CAMB[3] and CLASS[4] solves the equations to derive the transfer function.

While we have considered that baryon and Dark matter formed a perfect fluid, we know that baryons and DM have different evolution before recombination. Solving the equations for variables defined as the difference between baryon and Dark Matter we can find a solution corresponding to an initial push for baryons. However, this solution corresponds to decaying modes which are not of our interest as they quickly decay. With this decomposition, still working at linear order, we can derive from the Vlasov-Poisson system a second order differential equation describing the evolution of the growing modes

$$\frac{d^2D}{da^2} + \frac{d\ln\left(a^3H\right)}{da}\frac{dD}{da} - \frac{3\Omega_{m,0}H_0^2}{2a^5H^2}D = 0, \tag{1.71}$$

where we used a instead of η as the time variable. We further define the linear growth rate of structure as the logarithmic derivative of the growth factor

$$f \equiv \frac{d\ln D}{d\ln a}, \tag{1.72}$$

allowing to rewrite the second order differential equation as

$$\frac{df}{d\ln a} + f^2 + \left(2 + \frac{dH}{d\ln a}\frac{1}{H}\right)f - \frac{3}{2}\frac{\Omega_{m,0}}{a^3}\frac{H_0^2}{H^2} = 0, \tag{1.73}$$

With initial and present-day conditions $f(a=0)=1$ and $D(a=1)=1$. While this equation can be quite easily solved numerically, f can be well parametrised as

$$f = \Omega_m(a)^\gamma, \tag{1.74}$$

with the growth index $\gamma = 6/11$ in ΛCDM. As long as density contrast and velocities are small, from continuity equation we can write the velocity divergence as

$$\theta = -faH\delta_m, \tag{1.75}$$

and derive the linear velocity field from the density field

$$\boldsymbol{u}_m = -if\,aH\frac{\boldsymbol{k}}{k^2}\delta_m. \tag{1.76}$$

[3] https://camb.info/
[4] https://lesgourg.github.io/class_public/class.html

At linear order, all the cosmological information is captured by the density field. As we will discuss in Chapter 2 this is no longer true when perturbations become non-linear. Once we have derived solutions for both $T(k)$ and $D(a)$, we can predict the linear matter power spectrum which can be written

$$P_L(k,a) = \frac{8\pi^2}{25}\frac{A_s}{\Omega_{m,0}^2 H_0^4} D(a)^2 T(k)^2 k \left(\frac{k}{k_p}\right)^{n_s - 1 + \frac{1}{2}\alpha_s \ln(k/k_p)} . \tag{1.77}$$

An other important quantity is the dimensionless power spectrum which gives the contribution to the variance of the density field per bin of $d\ln k$

$$\Delta^2(k,a) = \frac{k^3 P_L(k,a)}{2\pi^2}. \tag{1.78}$$

The scenario $\Delta^2(k_{\rm NL}, a) = 1$ indicates that the modes in a unit logarithmic bin around wave number k generate fluctuations of order unity. This defines for a given time a the corresponding scale $k_{\rm NL}$ where linearity breaks (in practice linear models break before $k_{\rm NL}$).

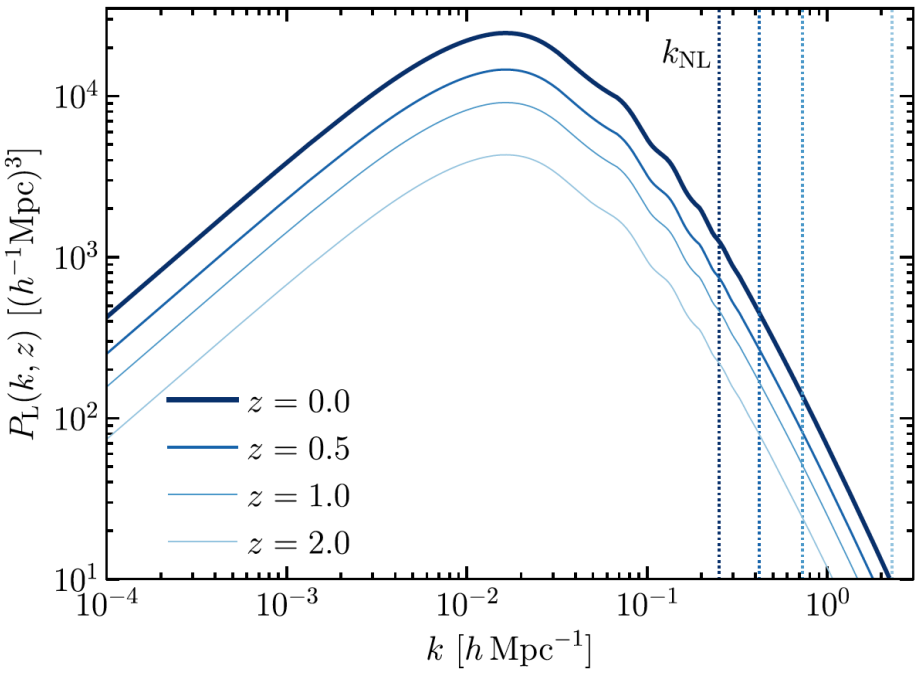


FIGURE 1.3 : The linear matter power spectrum with the Planck2018 ΛCDM cosmology at different redshifts. The vertical lines indicate $k_{\rm NL}$ for each redshift. Taken from DODELSON et SCHNEIDER 2013.

Figure 1.3 shows the linear matter power spectrum at different redshifts. The vertical lines indicate the non-linear scale limit $k_{\rm NL}$ for each redshift. As the Universe evolves, the density perturbations grow, increasing the range of scales for which the perturbations are non-linear. Besides the BAO features around $k \approx \times 10^{-1} h$ Mpc^{-1}, we can visualise the impact of the transfer function damping modes that entered the horizon before matter-radiation equality, where $T(k) \sim k^{-2}$ for $k >> k_{\rm eq}$, with $k_{\rm eq} \approx \times 10^{-2} h$ Mpc^{-1}.

Usually in large scale structure (LSS) analyses, the amplitude of the power spectrum is not

parametrised by A_s but by the standard deviation of density fluctuations in a sphere of $8\text{Mpc}h^{-1}$

$$\sigma_8(z) = \sqrt{\frac{1}{2\pi^2}\int k^2 dk P_m^{\text{lin}}(k,z) W^2(rk)} = \sigma_{8,0} D(z), \tag{1.79}$$

with $r = 8\ h^{-1}\text{Mpc}$, and $W(x) = 3(\sin x - x\cos x)/x^3$ the Fourier transform of the unit top-hat window. Historically, the specific smoothing scale of $8\text{Mpc}h^{-1}$ was chosen because it roughly corresponds to the scale of galaxy clusters, the largest bound structures in the Universe. This quantity is of considerable importance since galaxy surveys allow to measure $f(z)\sigma_8(z)$. We will discuss the methodology in Chapter 2 and apply it in Chapters 4 and 5.

In reality, as neutrinos are massive, they affect the matter perturbations. Because their energy density undergoes a transition between massless($\propto a^{-4}$) to massive particles($\propto a^{-3}$), neutrinos slightly affect the expansion rate and the growth factor amplitude. Moreover, since neutrinos are not cold (they move fast), they escape high density regions on small scales and only cluster on scales larger than the so-called *free streaming scale*, damping the clustering on small scales and making the growth rate scale-dependent : $f(k,z)$. This effect is enhanced as the total mass of neutrinos increases (VILLAESCUSA-NAVARRO et al. 2018).

One last caveat we need to discuss is the baryons-Dark Matter single fluid hypothesis. On small scales, electromagnetic interactions become non-negligible and affect the clustering of matter. Because of baryonic feedback such as supernovae exploding, active galactic nuclei radiating or gas moving towards low pressure, baryon perturbations does not grow as quickly as Dark Matter, affecting the overall clustering of matter on non-linear scales. Finally, one can only directly observe a specific fraction of the matter field via luminous matter (galaxies, quasars, gas, ...). The interconnection between total matter and luminous matter tracing its distribution is an important issue in modern cosmology. It will be discussed in further details in the next section.

2 Galaxy surveys : mapping the large-scale structure of the Universe

Summary

2.1 Modelling the clustering of galaxies . . . 45
2.1.1 Galaxies as tracers . . . 45
2.1.2 Cosmological probes . . . 46
2.1.2.1 BAO and Alcock-Paczynski effect . . . 46
2.1.2.2 Redshift Space Distortions . . . 51
2.1.3 Linear modelling of the two-points statistics . . . 54
2.1.3.1 Modelling for BAO analysis . . . 54
2.1.3.2 Modelling for RSD analysis . . . 58
2.1.4 Beyond linear theory . . . 61
2.1.5 Cosmological simulations . . . 62
2.1.5.1 Dark Matter only . . . 62
2.1.5.2 Simulating a galaxy sample . . . 63
2.2 Galaxy and peculiar velocity surveys . . . 65
2.2.1 Construction of 3D galaxy maps . . . 65
2.2.2 Measuring Galaxy peculiar velocities . . . 68
2.2.2.1 Tully-Fisher relation . . . 69
2.2.2.2 Fundamental plane relation . . . 69
2.2.2.3 From log-distance ratio to peculiar velocity . . . 70
2.2.3 Two-point statistics estimators . . . 71
2.2.3.1 Optimal weighting . . . 71
2.2.3.2 Random catalogues . . . 72
2.2.3.3 Correlation function estimator . . . 73
2.2.3.4 Power Spectrum estimator . . . 76
2.2.4 Observational effects . . . 77
2.3 Cosmological Inference . . . 78
2.3.1 Likelihood analysis . . . 78
2.3.2 Fisher forecast . . . 81
2.3.3 Covariance matrix . . . 82
2.3.4 Optimisation algorithms . . . 84

Galaxies are one of the best tracers of the underlying dark matter distribution in our Universe as a function of time. Spectroscopic galaxy surveys have become a powerful probe of the physical laws governing the formation of the large-scale structures because they provide precise spatial information for millions of galaxies over large volumes. Surveys such as the 2dFGRS (COLLESS et al. 2003), 6dFGS (JONES et al. 2004), WiggleZ (DRINKWATER et al. 2010), VIPERS (GUZZO et al. 2014), Sloan Digital Sky Survey (SDSS) BOSS, and eBOSS (EISENSTEIN et al. 2011; BLANTON et al. 2017; DAWSON et al. 2013; DAWSON et al. 2016), Euclid (A. LAUREIJS et ARDUINI 2011), DESI (D. COLLABORATION et al. 2016), have proven (and continue) to be essential tools for understanding the accelerated expansion of our Universe and for testing models of Dark Energy or modified theories of gravity (WEINBERG et al. 2013; ZHAI et al. 2017; HUTERER 2023; HOU et al. 2023). So far we have derived the linear power spectrum describing the spatial distribution of matter at various epochs. However, several challenges persist to connect theoretical predictions to observational data : linear theory is limited to describing large scales, and the matter field itself is not directly observable. In this chapter we will discuss how we can model and measure the two-point statistics (power spectrum and correlation function) of the galaxy and velocity fields as observables to constraint cosmological models.

2.1 Modelling the clustering of galaxies

2.1.1 Galaxies as tracers

The clustering of galaxies is one of the main probes of the large-scale structure in the Universe. Indeed, galaxies tend to form in dense environments, where enough gas can aggregate and cool down to start forming stars from which we observe the light. It is said that galaxies trace the matter field across the Universe. However, galaxies do not form randomly within the matter field. They rather undergo complex physical processes resulting in a biased distribution relative to the matter field. Perturbations of the galaxy density field can be related to the ones of the matter field as

$$\delta_g(\boldsymbol{x}, z) = \mathscr{F}\left[\delta_m(\boldsymbol{x}, z)\right], \tag{2.1}$$

where $\mathscr{F}$ is a general functional. Note that although we focus on galaxies, this approach is valid for a variety of tracers. Different approaches can be explored to model this relation. The perturbative bias expansion approach assume that the perturbation in the galaxy field can be written as

$$\delta_g(\boldsymbol{x}, z) = \sum_{\mathscr{O}} b_{\mathscr{O}}(z)\mathscr{O}\left(\delta_m(\boldsymbol{x}, z)\right), \tag{2.2}$$

where $b_{\mathscr{O}}$ are the bias parameters associated with a given operator $\mathscr{O}$. The goal of the general approach is to determine which operators of δ_m have to be included in the sum to describe the galaxy field (down to some minimum scale), while respecting the physical symmetries. On large enough scales, linear order can be assumed, and a single coefficient is sufficient to describe the bias

$$\delta_g(\boldsymbol{x}, z) = b_1(z)\delta_m(\boldsymbol{x}, z) \tag{2.3}$$

where b_1 is sensitive to the galaxy sample considered. In practice cutting the Taylor expansion and neglecting higher order terms tend to make the bias parameters scale dependent, as they absorb unmodelled physics. Hence, one need to carefully test the range of scale for which the model is robust. In the following chapters we will work at linear order in the bias expansion, a

more complete description of this formalism is given in DESJACQUES et al. 2018

The galaxy peculiar velocity field $\boldsymbol{u}_g$ provides complementary information to the spatial distribution of galaxies. Although for linear perturbations the velocity field is fully characterised by the density field (Eq 1.76), this is no longer true on non-linear scales. Moreover, the galaxy velocity field is often considered as unbiased with respect to dark matter given that galaxies and matter move along the same trajectories

$$\boldsymbol{u}_g(\boldsymbol{x}, z) = \boldsymbol{u}_m(\boldsymbol{x}, z). \tag{2.4}$$

Of course this approximation also breaks down on very small scales due to baryonic interactions. In the following we will only work with the component parallel to the LoS $\hat{n}$ of the galaxy velocity field $u_r = \boldsymbol{u}_g \cdot \hat{n}$ as this is the only component directly observable.

An alternative approach to the galaxy bias expansion is the so-called halo model. In this model, galaxies are assumed to form only in dark matter halos, gravitationally bound structures, spherical in average. The matter inside halos is assumed to follow the typical distribution introduced in NAVARRO et al. 1996

$$\rho(r) = \frac{\rho_s}{(r/R_s)(1 + r/R_s)^2} \tag{2.5}$$

where ρ_s and R_s are the scale density and scale radius. This distribution can be re-parametrised with the halo total mass M (integral of the distribution) and the concentration $c \equiv R_{\rm vir}/R_s$, with $R_{\rm vir}$ the virial radius. It is defined in spherical collapse model as the radius encompassing a mean over-density of typically 200 so that all the mass contained within this sphere is bound to the halo. The abundance of halos for a given logarithmic mass bin can be described by the halo mass function

$$\frac{dn(M,z)}{d\ln M} = \frac{\bar{\rho}_m}{M} f\left[\frac{\delta_c}{\sigma(M,z)}\right] \left|\frac{d\sigma(M,z)}{d\ln M}\right|, \tag{2.6}$$

where the mass and the comoving radius are related through the mean density enclosed within a sphere $\bar{\rho}_m = \frac{M}{4/3\pi R^3}$. The critical density contrast is $\delta_c \approx 1.686$ in the spherical collapse model, and $\sigma(M,z) \equiv \sigma_{R(M)(z)}$ is the variance of the linear density fluctuations in a sphere of radius as defined in Eq 1.79. f is a function related to the fraction of space contained in collapsed halos with mass larger than M. The free parameters in this model can be related (in some extent) to the bias parameters described previously. This formalism can be used to derive predictive model for the 2pt statistics of the galaxy field (SELJAK 2000). It also provides convenient recipes to populate dark-matter-only simulations with galaxies, as we will discuss in section 2.1.5.

2.1.2 Cosmological probes

Extracting cosmological information from the clustering of galaxies is not straightforward. As we will see in section 2.1.4, the calculation of the model for our observables, the non-linear two-point functions of the galaxy field, is a complex and numerically demanding task, mainly due to the non-linear aspect of it. The historical approach to constrain the cosmological parameters is to identify and isolate physical features (cosmology-dependent) within the observable. We call such features *cosmological probes*.

2.1.2.1 BAO and Alcock-Paczynski effect

Baryon Acoustic Oscillations (BAO) described in section 1.3.4 are imprinted in the matter distribution. In the galaxy field, BAO represent bubbles of overdensities with comoving radius

$r_\mathrm{d} = 147$Mpc where galaxies are more likely to form. This characteristic scale remains constant in comoving coordinates because the physical size of the sound horizon is larger than any collapsed structure, thus mitigating the impact of non-linear evolution. By comparing the measured BAO scale to the expected r_d at a given redshift, it becomes possible to measure distances accurately. This is why BAO in the galaxy field are often referred to as a standard ruler. However, as mentioned previously, BAO have a small effect on the galaxy distribution and are not directly observable if not through the two-point statistics of the whole galaxy field. As we saw in section 1.3.5, the characteristic scale of BAO manifests as oscillations in the matter (and therefore galaxy) power spectrum. In the correlation function, this translates to a peak (excess of probability) located at r_d. Consequently, this feature can be readily measured without needing to account for the general shape of these observables.

Spectroscopic observations of galaxies only yield angular positions and redshifts. The 3D comoving position of a galaxy is given by

$$\boldsymbol{x}_\mathrm{obs} = \chi(z)\hat{\boldsymbol{n}}(\theta,\phi) \tag{2.7}$$

where θ and ϕ are the polar and azimuthal angles, $\hat{\boldsymbol{n}}$ is the unitary LoS vector, and $\chi(z)$ is the comoving radial distance at redshift z. Going from those measured coordinates to physical 3D positions requires to assume a fiducial cosmology to compute $\chi_\mathrm{fid}(z)$, which is not necessarily the true underlying cosmology of our Universe. This effect introduced in Alcock et Paczynski 1979 generates distortions in the observed galaxy field and is commonly referred as the Alcock-Paczynski (AP) effect. While the total number of observed galaxies in a survey is an observable that does not depend on the fiducial cosmology, the comoving volume occupied by these galaxies does. Therefore, the number density of galaxy depends as well on the fiducial cosmology. The conservation of the number of galaxies can be written

$$\begin{aligned} n_\mathrm{g,obs}(\boldsymbol{x}_\mathrm{obs})d^3x_\mathrm{obs} &= n_\mathrm{g}(\boldsymbol{x})d^3x \\ n_\mathrm{g,obs}(\boldsymbol{x}_\mathrm{obs}) &= J n_\mathrm{g}(\boldsymbol{x}), \end{aligned} \tag{2.8}$$

where $J \equiv \left|\frac{d^3x}{d^3x_\mathrm{obs}}\right|$ is the Jacobian of the coordinate transformation between true and observed distances. The average comoving density of galaxies, $\bar{n}$ (in units of h^3 Mpc^{-3}) follows the same relation, therefore, $\delta_g(\boldsymbol{x}) \equiv n_g(\boldsymbol{x})/\bar{n}_g - 1$ can be written

$$\delta_{\mathrm{g,\,obs}}(\boldsymbol{x}_\mathrm{obs}) = \delta_g(\boldsymbol{x}). \tag{2.9}$$

Hence, the observed galaxy correlation is given by

$$\xi_{g,\,\mathrm{obs}}(\boldsymbol{s}_\mathrm{obs}) = \xi_g(\boldsymbol{s}), \quad \text{with} \quad \boldsymbol{s}_\mathrm{obs} = |\boldsymbol{x}_{1,\mathrm{obs}} - \boldsymbol{x}_{2,\mathrm{obs}}|, \tag{2.10}$$

affecting the correlation function scale dependence, specifically shifting the BAO peak location.

Now we need to define the mapping function between true and distorted positions $\boldsymbol{x}(\boldsymbol{x}_\mathrm{obs})$. Consider some galaxies distributed within a thin slice of redshift centred around $\bar{z}$. Let's further use the flat sky approximation, so each pair of galaxy share a fixed line of sight (approximation valid for distant survey and small separation angle). The comoving separation vector between two galaxies can be written $\boldsymbol{s} = (s_\perp, s_\parallel)$ with $s_\parallel$ the component parallel to the line of sight (LoS), and $s_\perp$ the transverse component. If the galaxies are located at redshifts z_1 and z_2 around the

mean redshift $\bar{z}$, we can write

$$s_\parallel = \chi(z_1) - \chi(z_2) \approx \frac{\Delta z}{H(\bar{z})}, \tag{2.11}$$

where $\Delta z = z_1 - z_2$, and we expanded to linear order around $\bar{z}$ under the assumption that the redshift slice is narrow enough. We used $d\chi/dz = 1/H$. For a small angular separation $\Delta\theta$ we have

$$s_\perp = \Delta\theta D_M(\bar{z}), \tag{2.12}$$

where we have defined the comoving angular distance $D_M = D_A(1+z)$. Writing the observed comoving separation vector obtained using the fiducial cosmology analogously $\boldsymbol{s}^{\rm obs} = \left(\Delta\theta D_M^{\rm fid}(\bar{z}), \Delta z/H^{\rm fid}(\bar{z})\right)$, we can express the true separation as

$$\boldsymbol{s}\left(\boldsymbol{s}^{\rm obs}\right) = \left(\alpha_\perp s_\perp^{\rm obs}, \alpha_\parallel s_\parallel^{\rm obs}\right), \tag{2.13}$$

where we have defined the AP parameters as

$$\alpha_\perp(z) = \frac{D_M}{D_M^{\rm fid}}, \quad \alpha_\parallel(z) = \frac{D_H}{D_H^{\rm fid}}, \tag{2.14}$$

with $D_H(z) = c/H(z)$ the Hubble distance. The Jacobian can now be written

$$J = \left|\frac{d^3x}{d^3x_{\rm obs}}\right| = \alpha_\perp^2\alpha_\parallel. \tag{2.15}$$

We can equivalently derive how those distortions behave in Fourier space and affect the power spectrum. We define the true wave vector as a function of the observed wave vector such that

$$\boldsymbol{k}(\boldsymbol{k}_{\rm obs}) = \left(\frac{1}{\alpha_\perp}k_\perp^{\rm obs}, \frac{1}{\alpha_\parallel}k_\parallel^{\rm obs}\right) \tag{2.16}$$

Using Eq 2.9 and the fact that $\boldsymbol{k}_{\rm obs}\cdot\boldsymbol{x}_{\rm obs} = \boldsymbol{k}\cdot\boldsymbol{x}$, we can derive the observed galaxy over-density in Fourier space :

$$\begin{aligned}
\delta_{\rm g,obs}(\boldsymbol{k}_{\rm obs}) &= \int d^3x_{\rm obs}\delta_{\rm g,obs}(\boldsymbol{x}_{\rm obs})e^{-i\boldsymbol{k}_{\rm obs}\cdot\boldsymbol{x}_{\rm obs}} \\
&= \int d^3x_{\rm obs}\delta_g(\boldsymbol{x}[\boldsymbol{x}_{\rm obs}])e^{-i\boldsymbol{k}_{\rm obs}\cdot\boldsymbol{x}_{\rm obs}} \\
&= \left|\frac{d^3x_{\rm obs}}{d^3x}\right|\int d^3x e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\delta_g(\boldsymbol{x}) \\
&= J^{-1}\delta_g(\boldsymbol{k}).
\end{aligned} \tag{2.17}$$

Now the distortion involves the Jacobian J. The reason is that the over-density in Fourier space has units of volume, while in configuration space the over-density is dimensionless. We remind that the galaxy power spectrum is defined as

$$\langle\delta_g(\boldsymbol{k})\delta_g(\boldsymbol{k}')\rangle = (2\pi)^3\delta_D(\boldsymbol{k}-\boldsymbol{k}')P_g(\boldsymbol{k}) \tag{2.18}$$

where δ_D is the Dirac-delta function. Note that both P and δ_D have dimension of volume and

are therefore both affected by the Jacobian J. From the definition of Dirac-delta :

$$(2\pi)^3\delta_D(\boldsymbol{k}) \equiv \int d^3x e^{-i\boldsymbol{k}\cdot\boldsymbol{x}} = J^{-1}\int d^3x_{\,\mathrm{obs}} e^{-i\boldsymbol{k}_{\mathrm{obs}}\cdot\boldsymbol{x}_{\mathrm{obs}}}, \tag{2.19}$$

such that

$$\delta_D(\boldsymbol{k}-\boldsymbol{k}') = J^{-1}\delta_{D,\,\mathrm{obs}}(\boldsymbol{k}_{\mathrm{obs}}-\boldsymbol{k}'_{\mathrm{obs}}). \tag{2.20}$$

Combining with

$$\langle\delta_{\,\mathrm{g,obs}}(\boldsymbol{k}_{\mathrm{obs}})\delta_{\,\mathrm{g,obs}}(\boldsymbol{k}'_{\mathrm{obs}})\rangle = J^{-2}\langle\delta_g(\boldsymbol{k})\delta_g(\boldsymbol{k}')\rangle, \tag{2.21}$$

we obtain that the observed power spectrum is related to the true one by

$$P_{\mathrm{g,\,obs}}(\boldsymbol{k}_{\,\mathrm{obs}}) = J^{-1}P_{\mathrm{g}}(\boldsymbol{k}) = \frac{1}{\alpha_\perp^2\alpha_\parallel}P_{\mathrm{g}}(\boldsymbol{k}). \tag{2.22}$$

In Fourier space, the AP effect shifts the BAO wiggles locations and rescales the total amplitude of the power spectrum. Note that this last effect is small since the α's parameters should be close to one (using Planck prior). An important point is that the AP effect breaks the isotropy of the two-point statistics as the component parallel and transverse to the LoS are affected by different dilation factors.

Historically, and for the reasons we mentioned before, BAO analyses are performed using a fixed template for the galaxy power spectrum (or correlation function). Computing this template requires once again to choose a specific template cosmology, which could differ from the fiducial cosmology, although they are often the same for practical reasons. Therefore, even without AP effect, we also need to rescale the two-point statistics to match the BAO location corresponding to the true (and unknown) cosmology. We can then write

$$\alpha_\perp(z) = \frac{D_M\, r_{\mathrm{d}}^{\mathrm{fid}}}{D_M^{\mathrm{fid}}\, r_{\mathrm{d}}}, \quad \alpha_\parallel(z) = \frac{D_H\, r_{\mathrm{d}}^{\mathrm{fid}}}{D_H^{\mathrm{fid}}\, r_{\mathrm{d}}}, \tag{2.23}$$

with $r_{\mathrm{d}}^{\mathrm{fid}}$ the sound horizon computed at the drag redshift in the fiducial cosmology.

In the past, the limited size of galaxy surveys made it difficult to measure separately the clustering signal along the transverse and parallel to the LoS directions. Instead of $\alpha_\parallel$ and $\alpha_\perp$, one can use the isotropic and anisotropic dilation factors α_V and ϵ respectively, defined as

$$\alpha_V(z) = \frac{D_V\, r_{\mathrm{d}}^{\mathrm{fid}}}{D_V^{\mathrm{fid}}\, r_{\mathrm{d}}}, \quad 1+\epsilon(z) = \left[\frac{D_H}{D_M}\frac{D_M^{\mathrm{fid}}}{D_H^{\mathrm{fid}}}\right]^{1/3}, \tag{2.24}$$

where the spherically averaged distance is given by

$$D_V(z) = \left[zD_H(z)D_M^2(z)\right]^{1/3}. \tag{2.25}$$

From here we can derive the relation with the previous parameters as

$$\alpha_\perp = \alpha_V(1+\epsilon)^{-1} \quad \alpha_\parallel = \alpha_V(1+\epsilon)^2. \tag{2.26}$$

Note that if there is no isotropic nor anisotropic shifts $\alpha_V = 1$ and $\epsilon = 0$. Limited survey with a signal-to-noise ratio (SNR) too weak only measure the spherically averaged distance D_V by

fitting for the parameter α_V. For high enough SNR, two parameters can be left free, allowing to constraint separately the Hubble and comoving angular distances. Doing so for different redshift bins, we can measure the expansion history of the Universe trough $D_H(z)/r_\mathrm{d}$ and $D_M(z)/r_\mathrm{d}$.

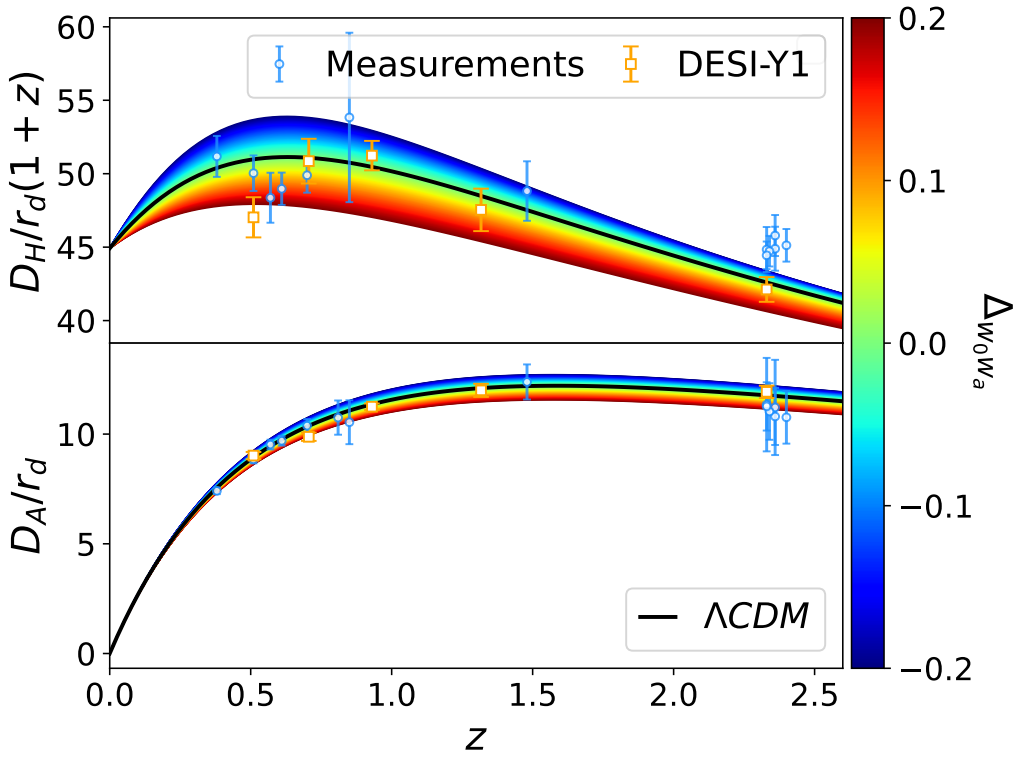


FIGURE 2.1 : Expansion history of the Universe (in h^{-1}Mpc). The black lines show the predicted angular and Hubble distances for the ΛCDM model. The colour gradient illustrates the impact of varying the Dark Energy EoS parameters (w_0, w_a) up to 20% away from ΛCDM. The blue errorbars show measurements from previous surveys while the orange ones show the latest measurement from the DESI collaboration.

Figure 2.1 describe the comoving Hubble and angular distances evolution with respect to redshift. The black lines show the ΛCDM model with the latest Planck measurements (PLANCK COLLABORATION et al. 2020). The coloured lines illustrate the impact of varying simultaneously the Dark Energy EoS parameters (w_0, w_a), with $w_i = w_i^{\Lambda\mathrm{CDM}} + \Delta w_i$. Because those parameters are quite correlated we vary them at the same time, but we should note that w_0 is the parameter affecting the most the signal. The blue points correspond to BAO measurements performed from previous galaxy surveys. The orange points show the latest BAO measurement from the Dark Energy Spectroscopic Instrument (DESI) collaboration D. COLLABORATION et al. 2024, which tend to favour an evolving EoS with $w_0 > -1$ and $w_a < 0$. Those measurements also allow to constraint $\Omega_{m,0}$ and the product $H_0 r_\mathrm{d}$. Big Bang Nucleosynthesis (BBN) theory predict the abundance of light element in the early Universe as function of the baryon to photon ratio. Consequently, observational determination of these abundances allow to constrain the physical baryon density $\omega_b = \Omega_{b,0} h^2$ and to break the degeneracy between H_0 and r_d. This can also be achieved by using the Planck prior on r_d. Assuming the BBN prior, DESI found a H_0 value compatible with the

Planck results with competitive errorbars. In Chapter 3 we report our BAO measurement obtained from the luminous red galaxy sample of the extended baryon oscillation spectroscopic survey (eBOSS) for an effective redshift of 0.7.

2.1.2.2 Redshift Space Distortions

Galaxies have their own peculiar motion within the Hubble flow. The peculiar velocity of a given galaxy with respect to the observer introduces a Doppler shift, and the spectroscopic redshift is given by

$$1 + z_{\rm obs} = (1 + z_{\rm cos})(1 + z_{\rm p}), \tag{2.27}$$

with the relativistic Doppler shift given by

$$1 + z_{\rm p} = \sqrt{\frac{1 + u_r/c}{1 - u_r/c}}, \tag{2.28}$$

where $u_r = \boldsymbol{u} \cdot \hat{\boldsymbol{n}}$ is the radial component of the galaxy velocity. At linear order in u_r/c (galaxies move at a few hundreds of km/s), we can write

$$z_{\rm p} = u_r/c. \tag{2.29}$$

From Eq 2.7, still working at linear order, we can derive the shift introduced in the observed galaxy position

$$\boldsymbol{x}_{\rm obs} = \boldsymbol{x} + \frac{1}{aH} u_r(\boldsymbol{x}) \hat{\boldsymbol{n}}. \tag{2.30}$$

This effect is called Redshift Space Distortion (RSD). Once again following the conservation of the number of Galaxies in Eq 2.8 we can write the Jacobian as

$$J = \left| \frac{dx}{dx_{\rm obs}} \right| \frac{x^2}{x_{\rm obs}^2} \frac{d\Omega}{d\Omega_{\rm obs}}, \tag{2.31}$$

with $x \equiv |\boldsymbol{x}|$. As RSD are purely radial $d\Omega = d\Omega_{\rm obs}$, and we get at linear order

$$J = \left| 1 - \frac{1}{aH} \frac{du_r}{dx} \right| \left(1 - \frac{2}{aH} \frac{u_r}{x} \right). \tag{2.32}$$

Note that now the Jacobian is a function of the position. Therefore, averaging over a large enough survey volume we get $\langle J(x) \rangle = 1$ as the velocity field averages to zero. Hence, the mean number densities are equal in both spaces $\bar{n}_{\rm obs} = \bar{n}$. For a distant observer, the term $\frac{u_r}{x}$ can be neglected and we get

$$\delta_{\rm g,obs}(\boldsymbol{x}_{\rm obs}) = \delta_{\rm g}(\boldsymbol{x}) - \frac{1}{aH} \frac{du_r}{dx}(\boldsymbol{x}). \tag{2.33}$$

Note that we can now evaluate the fields at $\boldsymbol{x}_{\rm obs}$ or $\boldsymbol{x}$ as it will only bring 2nd order corrections. This also implies that the distortions induced in the velocity field itself are second order terms as well.

Figure 2.2 show how the RSD affect the galaxy density and velocity fields in a 20 h^{-1}Mpc slice of the $(500h^{-1}{\rm Mpc})^3$ MileniumTNG simulation (PAKMOR et al. 2023). Top panels correspond to configuration space and bottom panels to the observed redshift space. In this example we make use of the distant observer approximation so the LoS is kept fixed. Negative and positive radial

velocities are respectively shown in blue and red. RSD can be decomposed into two main effects. On large/linear scales, galaxies fall towards over-dense regions. Galaxies located between us and an over-density move away from us and appear farther, while galaxies behind the over-density move towards us and appear closer than they actually are. As a result, over-dense regions appear squashed along the LoS. It is worth noticing that since galaxies move towards each other, the number density increases and the clustering in redshift space is larger than is configuration space. This is known as the Kaiser effect (KAISER 1987). On small scales, non-linear motions generate large velocities. For galaxies falling towards an over-density and located a few megaparsecs away, the induced displacement typically exceeds their distance from the over-density. Galaxies located between us and an over-density move away from us and appear behind the over density. Galaxies located behind the over-density are shifted to appear in front. Accordingly, small over-densities appear elongated along the LoS. This is known as the *Fingers of God* effect.

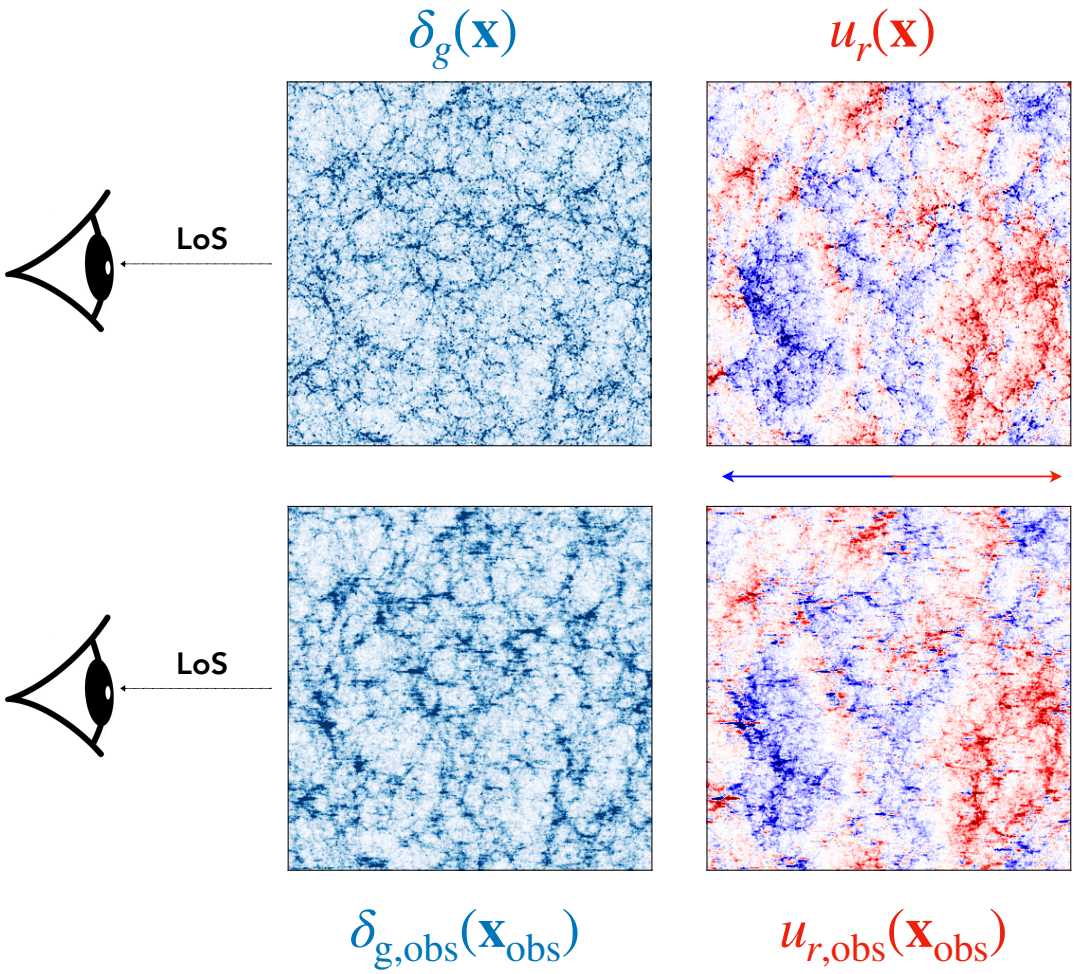


FIGURE 2.2 : Visualisation of the RSD in a slice (20 h^{-1}Mpc thick) of the MileniumTNG simulation (PAKMOR et al. 2023). The total volume of the box is $(500h^{-1}\mathrm{Mpc})^3$. Galaxy density and velocity fields are shown in configuration space in the top panels, and in and redshift space in the bottom panels. Negatives(positives) radial velocities are shown in blue(red).

Let us derive the expression for the galaxy power spectrum on linear scales. From Eq 1.76 we can write the linear velocity field as

$$\frac{u_r(\boldsymbol{x})}{aH} = if\int \frac{dk'^3}{(2\pi)^3}\delta_m(\boldsymbol{k}')\frac{\boldsymbol{k}'\cdot\hat{\boldsymbol{n}}}{k'^2}e^{i\boldsymbol{k}'\cdot\boldsymbol{x}}. \tag{2.34}$$

Writing the galaxy field $\delta_g = b_1\delta_m$, can derive the observed galaxy perturbation in Fourier space :

$$\begin{aligned}\delta_{g,obs}(\boldsymbol{k}) &= \int dx^3 e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\left[b_1\delta_m(\boldsymbol{x}) - if\int\frac{dk'^3}{(2\pi)^3}\delta_m(\boldsymbol{k}')\frac{\boldsymbol{k}'\cdot\hat{\boldsymbol{n}}}{k'^2}\frac{d}{dx}e^{i\boldsymbol{k}'\cdot\boldsymbol{x}}\right]\\ &= b_1\delta_m(\boldsymbol{k}) + f\int\frac{dk'^3}{(2\pi)^3}\frac{(\boldsymbol{k}'\cdot\hat{\boldsymbol{n}})^2}{k'^2}\int dx^3 e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{x}}.\end{aligned} \tag{2.35}$$

Reminding the definition $(2\pi)^3\delta^D(\boldsymbol{k}-\boldsymbol{k}') \equiv \int dx^3 e^{-i(\boldsymbol{k}-\boldsymbol{k}')\cdot\boldsymbol{x}}$, and making use of the flat sky approximation (fixed LoS) $\mu_k \equiv \hat{k}\cdot\hat{\boldsymbol{n}}$, we get

$$\delta_{g,obs}(\boldsymbol{k}) = \left[b_1 + f\mu_k^2\right]\delta_m(\boldsymbol{k}), \tag{2.36}$$

and the redshift space power spectrum is

$$P_{g,obs}(k,\mu_k) = \left[b_1 + f\mu_k^2\right]^2 P_L(k). \tag{2.37}$$

RSD also break the isotropy on the galaxy two-point functions, allowing to measure $f\sigma_8$ at the given redshift (σ_8 is contained in the global amplitude of P_L). $f\sigma_8$ is commonly referred as the growth rate for brevity. Let us note that both Dark Energy and modified gravity are able to produce the accelerated expansion. However, their predictions differ for the growth rate of matter perturbations. Therefore, linear growth rate measurements can be used to discriminate different models of gravity.

Figure 2.3 shows the growth rate evolution with respect to redshift. The black solid line show the ΛCDM (+ General Relativity) model with the latest Planck measurements (Planck Collaboration et al. 2020). The black dotted line show the prediction of the flat DGP gravity model (Dvali et al. 2000). The black dashed line show the prediction of one type of $f(R)$ model (Mirzatuny et Pierpaoli 2019), that typically lead to an alteration of equation 1.73, with an effective gravitation constant $G_{\rm eff}$. Those models are easy to implement as the growth rate can be described from Eq 1.74 with different growth indexes. For DGP (Linder et Cahn 2007) $\gamma = 11/16$, and $\gamma = 0.41$ for $f(R)$ (Gannouji et al. 2009). As previously done for BAO measurements, the coloured lines illustrate the impact of varying the Dark Energy EoS parameters (w_0, w_a), with $w_i = w_i^{\Lambda CDM} + \Delta w_i$. We see that while the effect of varying (w_0, w_a) is prominent at redshift $z \sim 0.5$, different models of gravitation can be easily discriminated at lower redshift. The blue points correspond to previous growth rate measurements from RSD galaxy clustering. Forecast of the DESI RSD constraints are shown is orange. The black points are the forecasted constraints obtained combining the DESI galaxy field in the Northern Hemisphere (+complementary surveys) with the Large Synoptic Survey Telescope (LSST) peculiar velocity forecasts derived from SnIa (Howlett et al. 2017) in the Southern Hemisphere. Ongoing galaxy surveys such as Euclid (A. Laureijs et Arduini 2011; E. Collaboration et al. 2022) and DESI (DESI Collaboration et al. 2022) will detect tens of millions of galaxies, significantly reducing statistical uncertainties in the galaxy two-point functions. At low redshift, because of the limited volume, the density

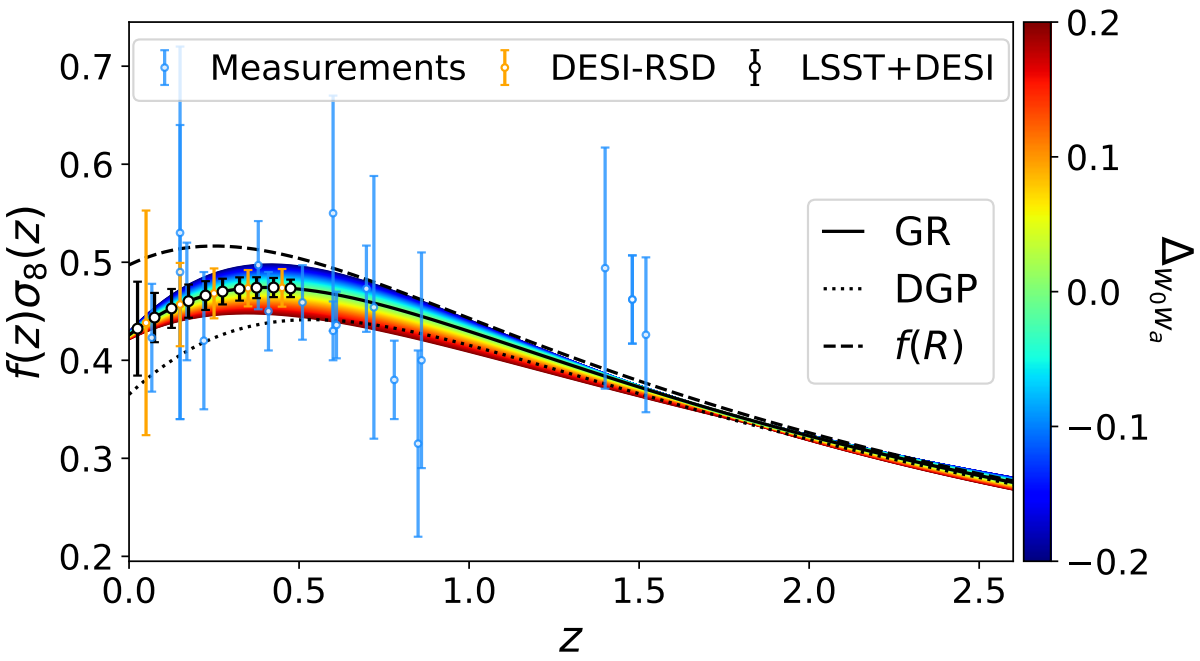


FIGURE 2.3 : Growth history of the Universe. The black solid line shows the Planck2018 ΛCDM prediction. The black dashed and dotted line show the prediction for alternative models of gravity, $f(R)$ and DGP respectively. The colour gradient illustrates the impact of varying the Dark Energy EoS parameters (w_0, w_a) up to 20% away from ΛCDM. Blue errorbars show measurements from previous surveys. The orange ones show the DESI forecast, while combining DESI galaxies with LSST SnIa peculiar velocity measurements are shown in black.

of tracers is less significant compared to cosmic variance which become the primary source of uncertainty. Recent studies have shown that combining galaxy and peculiar velocity clustering allows to leverage additional information from available data, increasing the precision on growth rate measurements (ADAMS et BLAKE 2017; TURNER et al. 2023). We present our work on such measurement in Chapter 5.

2.1.3 Linear modelling of the two-points statistics

Now that we have outlined how the BAO and RSD features in the galaxy field can be used to constraint the cosmological model, let us write down the linear model for our observables : the correlation function in configuration space or the power spectrum in Fourier space. While some models are directly formulated in configuration space, our approach in this work focuses solely on the power spectrum derived in Fourier space. Subsequently, we transform it into configuration space to obtain the correlation function prediction.

2.1.3.1 Modelling for BAO analysis

In BAO analysis, we are only interested in fitting the BAO peak (in configuration space) or wiggles (in Fourier space) locations to extract the parameters $\alpha_\parallel$ and $\alpha_\perp$. However, the non-linear

structure formation tend to smooth the BAO features, reducing the statistical sensitivity of the measurement. The common methodology is to attempt to remove the impact of non-linear evolution of galaxies (and a fraction of RSD) by applying the reconstruction technique described in EISENSTEIN et al. 2007. This technique uses the density field to estimate the displacements using the Zel'dovich approximation (ZEL'DOVICH 1970). In this approximation, the gravitational collapse of matter is approximated by assuming that the particles move along straight-line trajectories, with their initial peculiar velocities determining their motion. This approximation neglects the effects of pressure and other complicated physical processes, focusing solely on the dynamics of gravitational collapse. This is equivalent to using the linear velocity of the density field (Eq 1.76) to move the galaxies "back in time". Doing so sharpens the BAO feature, increasing the precision of the measurement on $\alpha_\perp$ and $\alpha_\parallel$. It as also been shown that the BAO results are robust to the choice of cosmology used to perform the reconstruction (CARTER et al. 2020).

In this work we use the linear-theory-based model for the redshift-space anisotropic galaxy power-spectrum described in J. E. BAUTISTA et al. 2021; GIL-MARÍN et al. 2021, based on previous work ALAM et al. 2017; GIL-MARÍN et al. 2016; GIL-MARÍN et al. 2018; ROSS et al. 2017; J. E. BAUTISTA et al. 2018. This model is written (dropping the redshift dependence for conciseness)

$$P_{\rm gg}(k,\mu_k) = b_1^2\left[1+\beta\left(1-S(k)\right)\mu_k^2\right]^2 D_g^2(k,\mu_k)\times \\ \times\left[P_{\text{no peak}}(k)+P_{\text{peak}}(k)e^{-k^2\Sigma^2_{\rm nl}(\mu_k)/2}\right], \tag{2.38}$$

where b_1 is the linear density bias, $\beta = f/b$ the redshift-space distortions parameter, and the non-linear random motions on small scales (*Fingers of God*) are modelled by a Lorentzian distribution parametrized by a characteristic scale Σ_s

$$D_g^2(k,\mu_k) = \frac{1}{1+k^2\mu_k^2\Sigma_s^2/2}. \tag{2.39}$$

The BAO broadening caused by non-linear clustering is modelled by applying an anisotropic Gaussian smoothing on the BAO features. First, the linear power spectrum $P_{\rm L}$ is obtained from CAMB[1] (LEWIS et al. 2000) (for the given fixed fiducial cosmology). Then, we follow the procedure from KIRKBY et al. 2013 to decompose the BAO peak component $P_{\rm peak}$ from the full linear power-spectrum $P_{\rm L} = P_{\text{no peak}} + P_{\rm peak}$. We start by computing the correlation function by Fourier transforming $P_{\rm L}$, then we replace the correlations over the peak region by a polynomial (or spline) function fitted using information outside the peak region (typically $50 < r < 80$ and $160 < r < 190 h^{-1}$Mpc). The resulting correlation function is then Fourier transformed back to get $P_{\text{no peak}}$. Finally, we multiply the "peak-only" power spectrum $P_{\rm peak} = P_{\rm L} - P_{\text{no peak}}$ by a Gaussian function with dispersion given by $\Sigma^2_{\rm nl}(\mu_k) = \Sigma^2_\parallel \mu_k^2 + \Sigma^2_\perp(1-\mu_k^2)$. The reconstruction procedure is modelled following SEO et al. 2016, with $S(k) = e^{-k^2\Sigma_r^2/2}$ ($S(k) = 0$ for without reconstruction), with $\Sigma_r = 15 h^{-1}$Mpc the smoothing parameter used for the reconstruction.

As described in section 2.1.2.1, the BAO peak position is parametrised via the two dilation parameters scaling separations along and across the LoS $(\alpha_\parallel, \alpha_\perp)$, and the observed power spectrum relates to the true power spectrum as

$$P_{\text{gg, obs}}(k_\parallel, k_\perp) = \frac{1}{{\alpha_\perp}^2\alpha_\parallel} P_{\rm gg}\left(\frac{k_\parallel}{\alpha_\parallel}, \frac{k_\perp}{\alpha_\perp}\right). \tag{2.40}$$

[1] camb.info

The transverse and parallel to the LoS components are related to the coordinates (k, μ_k) through

$$k_\parallel^2 + k_\perp^2 = k^2, \quad k_\parallel = k\mu_k. \tag{2.41}$$

We can derive the relation between true and observed coordinates from Eq 2.16 as

$$\begin{aligned} k &= \frac{k^{\rm obs}}{\alpha_\perp}\left[1 + \left(\mu_k^{\rm obs}\right)^2\left(\frac{\alpha_\perp^2}{\alpha_\parallel^2} - 1\right)\right]^{1/2} \\ \mu_k &= \frac{\mu_k^{\rm obs}\alpha_\perp}{\alpha_\parallel}\left[1 + \left(\mu_k^{\rm obs}\right)^2\left(\frac{\alpha_\perp^2}{\alpha_\parallel^2} - 1\right)\right]^{-1/2}. \end{aligned} \tag{2.42}$$

We apply the scaling factors exclusively to the peak component of the power spectrum $P_{\rm peak}(k)$, effectively removing any dependency of these parameters on the smooth part $P_{\rm no\,peak}(k)$ (KIRKBY et al. 2013).

Furthermore, we compress the two-dimensional power spectrum into multipoles $P_\ell(k)$, by integrating $P_{\rm gg}(k, \mu_k)$ over μ_k, weighted by the Legendre polynomials $L_\ell(\mu_k)$:

$$P_\ell^{\rm gg}(k) = \frac{2\ell+1}{2}\int_{-1}^{1} P_{\rm gg}(k, \mu_k) L_\ell(\mu_k) d\mu, \tag{2.43}$$

such that

$$P_{\rm gg}(k, \mu_k) = \sum_\ell P_\ell^{\rm gg}(k) L_\ell(\mu_k). \tag{2.44}$$

This expression is only valid in the flat sky approximation, where any pair of galaxies located in $\boldsymbol{x}$ and $\boldsymbol{x}'$ share the same LoS $\hat{n} = \hat{x} = \hat{x'}$, such that $\mu \equiv \boldsymbol{s}\cdot\hat{n}$ and $\mu_k \equiv \boldsymbol{k}\cdot\hat{n}$. Because $P_{\rm gg}(k, \mu_k)$ is an even function of μ_k, the odd multipoles vanish. Standard analyses only use the two or three first even multipoles $\ell = [0, 2, 4]$ (monopole, quadrupole and hexadecapole respectively).

The galaxy correlation function multipoles can be express in terms of the power spectrum multipoles. Reminding that the correlation function between galaxies located in $\boldsymbol{x}$ and $\boldsymbol{x}'$ can be written

$$\xi_{\rm gg}(\boldsymbol{s}) = \langle \delta_{\rm g}(\boldsymbol{x})\delta_{\rm g}(\boldsymbol{x}')\rangle = \int \frac{d^3k}{(2\pi)^3} P_{\rm gg}(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot\boldsymbol{s}}, \quad \text{with } \boldsymbol{s} = \boldsymbol{x} - \boldsymbol{x}'. \tag{2.45}$$

Writting the spherical volume element $d^3k = k^2 dk d\Omega_k$, substituting the multipole decomposition of $P_{\rm gg}$ and using the plane-wave expansion

$$e^{i\boldsymbol{k}.\boldsymbol{s}} = \sum_\ell (2\ell+1) i^\ell j_\ell(ks) L_\ell(\hat{k}\cdot\hat{s}), \tag{2.46}$$

where j_ℓ are the spherical Bessel functions, we get

$$\xi_{\rm gg}(\boldsymbol{s}) = \frac{1}{(2\pi)^3}\int dk k^2 \sum_\ell (2\ell+1) i^\ell j_\ell(ks) \sum_{\ell'} P_{gg}^{\ell'}(k) \int d\Omega_k L_\ell(\hat{k}.\hat{s}) L_{\ell'}(\hat{k}\cdot\hat{n}). \tag{2.47}$$

We then express the Legendre polynomials in terms of spherical harmonic

$$L_\ell(\hat{k}.\hat{s}) = \frac{4\pi}{2\ell+1}\sum_m Y_{\ell m}(\hat{k}) Y_{\ell m}^*(\hat{s}), \tag{2.48}$$

so the last integral becomes

$$\begin{aligned}\int d\Omega_k L_\ell(\hat{k}\cdot\hat{s})L_{\ell'}(\hat{k}\cdot\hat{n}) &= \int d\Omega_k \frac{4\pi}{2\ell+1}\sum_m Y_{\ell m}(\hat{k})Y^*_{\ell m}(\hat{s})\frac{4\pi}{2\ell'+1}\sum_{m'} Y^*_{\ell' m'}(\hat{k})Y_{\ell' m'}(\hat{n}) \\ &= \frac{4\pi}{2\ell+1}\frac{4\pi}{2\ell'+1}\sum_m\sum_{m'} Y^*_{\ell m}(\hat{s})Y_{\ell' m'}(\hat{n})\int d\Omega_k Y_{\ell m}(\hat{k})Y^*_{\ell' m'}(\hat{k}).\end{aligned} \tag{2.49}$$

Because spherical harmonics are orthonormal

$$\int d\Omega_k Y_{\ell m}(\hat{k})Y^*_{\ell' m'}(\hat{k}) = \delta_{\ell\ell'}\delta_{mm'}, \tag{2.50}$$

such that Eq 2.49 gives

$$\begin{aligned}\int d\Omega_k L_\ell(\hat{k}\cdot\hat{s})L_{\ell'}(\hat{k}\cdot\hat{n}) &= \delta_{\ell\ell'}\left(\frac{4\pi}{2\ell+1}\right)^2\sum_m Y^*_{\ell m}(\hat{s})Y_{\ell m}(\hat{n}) \\ &= \frac{4\pi}{2\ell+1}L_\ell(\hat{s}\cdot\hat{n})\delta_{\ell\ell'}.\end{aligned} \tag{2.51}$$

Injecting this result in Eq 2.47 we get

$$\xi_{\rm gg}(\boldsymbol{s}) = \frac{1}{2\pi^2}\int dk k^2 \sum_\ell i^\ell j_\ell(ks)P^\ell_{gg}(k)L_\ell(\hat{s}\cdot\hat{n}). \tag{2.52}$$

From the definition of multipole expansion and noting that $\mu = \hat{s}\cdot\hat{n}$, we can finally identify the galaxy correlation function multipoles as the Hankel transform of the $P_\ell(k)$:

$$\xi^{\rm gg}_\ell(s) = \frac{i^\ell}{2\pi^2}\int_0^\infty k^2 j_\ell(ks)P^{\rm gg}_\ell(k)dk. \tag{2.53}$$

These transforms are computed using HANKL[2] that implements the FFTLog algorithm by HAMILTON 2000.

In order to properly marginalize over any mismodelling of the smooth part of $P_\ell(k)$ and $\xi_\ell(r)$ or any potential unknown systematic contaminating our measurements, we add to our model a linear combination of polynomial functions of scale, with free amplitudes to be marginalised over. Furthermore, since the $S(k)$ term in Eq. 2.38 is generally not sufficient to accurately model the reconstruction, these smooth functions can also account for this mismodelling. Our final multipoles model can be written as :

$$P^{\rm gg,tot}_\ell(k) = P^{\rm gg}_\ell(\alpha_\perp,\alpha_\parallel,k) + \sum_{i=i_{\rm min}}^{i_{\rm max}} a^P_{\ell,i}k^i, \tag{2.54}$$

$$\xi^{\rm gg,tot}_\ell(s) = \xi^{\rm gg}_\ell(\alpha_\perp,\alpha_\parallel,s) + \sum_{i=i_{\rm min}}^{i_{\rm max}} a^\xi_{\ell,i}s^i, \tag{2.55}$$

where $a^P_{\ell,i}$ and $a^\xi_{\ell,i}$ are the amplitudes for each power i of scale, k or s, and multipole order ℓ. The amount of free parameters of this kind used can vary with the analysis and will be discussed

[2] https://hankl.readthedocs.io/en/latest/

in the next chapter.

So far we have described the modelling for the multipoles of the galaxy density power spectrum and correlation function. However, the observed field depends on the mask of the survey. It can be written as a multiplicative term in configuration space

$$\hat{n}_g = n_g(\boldsymbol{x})W(\boldsymbol{x}), \tag{2.56}$$

with $n_g(\boldsymbol{x}) = \bar{n}(1+\delta_g(\boldsymbol{x}))$ the galaxy density, and $W(\boldsymbol{x})$ the window function. In Fourier space, modelling the observed field $\hat{\delta}_g(\boldsymbol{k})$ require to perform a computationally expensive convolution. We follow the method described in BEUTLER et al. 2017, where we first compute the multipoles in configuration space $\xi_\ell^{\rm gg}(s)$, multiply with the window function multipoles, and transform back into Fourier space to get the convolved $P_\ell^{\rm gg}(k)$. Analyses in configuration space do not require those steps in the modelling process, as the "deconvolved" correlation function is directly estimated. The code producing this overall model is publicly available[3].

2.1.3.2 Modelling for RSD analysis

RSD analyses aim at measuring the growth rate of structures $f\sigma_8$ - in a given redshift bin - by modelling the shape of the two-point statistics. As mentioned previously, the physical relationship between galaxy density and radial peculiar velocity substantially reduces cosmic variance. Consequently, analysing jointly their clustering increases the precision of the parameter measurements. We can use the auto and cross-correlation of the two random fields δ_g and u_r to perform a 3×2 points analysis. Dropping the redshift dependence on $P_L(k,z)$ for conciseness, the galaxy and velocity auto and cross power spectra can be written

$$\begin{aligned}
P_{\rm gg}(k,\mu_k) &= b_1^2\left[1+\beta\mu_k^2\right]^2 D_{\rm g}^2(k,\mu_k)P_L(k),\\
P_{\rm vv}(k,\mu_k) &= \left(\frac{aHf\mu}{k}\right)^2 D_{\rm u}^2(k)P_L(k),\\
P_{\rm vg}(k,\mu_k) &= -i\frac{aHf\mu}{k}b_1\left[1+\beta\mu_k^2\right]D_{\rm g}(k,\mu_k)D_{\rm u}(k)P_L(k)\\
&= -P_{\rm gv}(k,\mu_k).
\end{aligned} \tag{2.57}$$

Unlike the density field, the velocity does not have the squashing Kaiser term. However RSD introduces a damping on the non-linear scales analogously to $D_{\rm g}$. It is empirically modelled (fitted on simulations) with $D_{\rm u}$ (KODA et al. 2014) defined as

$$D_u(k) = \frac{\sin(k\sigma_u)}{k\sigma_u}, \tag{2.58}$$

with σ_u the characteristic damping length in h^{-1}Mpc. In practice, the global amplitude of the clustering is parameterised in those equations by writing $P_L(k,z) = P_L(k,z=0)\,(\sigma_8(z)/\sigma_{8,0})^2$. The multipoles in Fourier or configuration space are computed as described previously, as is the effect of the window function. The velocity power spectrum is also an even function of μ_k and has only even non-vanishing multipoles $[P_0^{\rm vv}, P_2^{\rm vv}, ...]$. While the galaxy-velocity power

[3] https://github.com/julianbautista/baopy

spectrum is neither an odd or even function of μ_k, most of the signal is contained in the dipole $P_1^{\rm gv}$. Furthermore, the modelling of the AP effect must still be considered following Eq 2.42. Consequently, in RSD analyses, the parameters $(\alpha_\parallel, \alpha_\perp)$ are also constrained. Because we assume that galaxy positions are a random Poisson point-process realisation of the underlying matter field, the Fourier modes are affected by shot noise, resulting in an additive term to the power spectrum. Although in practice it is often left as a free parameter, the simplest model predicts

$$P_{\rm noise} = \frac{1}{\bar{n}}, \tag{2.59}$$

with $\bar{n}$ the mean galaxy number density. As we mention before, the standard analysis uses a fixed template for P_L computed for the fiducial cosmology (and the effective redshift and relevant scales). The main advantages of this approach is that it is computational cheap. Furthermore, although the linear template is usually computed within the ΛCDM model, such analysis is only sensible to the overall shape and the BAO dilation. Thus, measuring the quantities $(f\sigma_8, \alpha_\parallel, \alpha_\perp)$ at different redshift allows to compress the information and test different cosmological models. The downside of such template analysis is that all the cosmological information is not properly extracted, compared to performing a full shape analysis letting the cosmological parameters of the template P_L vary during the fit.

A recent approach managed to extract the same amount of information as the full shape analysis in a model-independent way by incorporating just one additional effective parameter to the standard template analysis (BRIEDEN et al. 2021). However, the primary limitation in measuring the growth rate is the number of scales that can be fitted before the linear model ceases to offer a realistic description. The validity range of such models can be tested on high resolution N-body simulations, numerically solving the Vlasov-Poisson system (Eq 1.68) for a given number of dark matter "particles" in a box. While such cosmological simulations are computationally expensive, they accurately describe the non-linear clustering of matter down to small scales, and are extensively used to test analytical models (among many things). We will discuss their use in details in section 2.1.5.

Figure 2.4 compares the linear modelling of $\xi_0^{\rm gg}(s)$, $\xi_0^{\rm vv}(s)$ and $\xi_1^{\rm vg}(s)$ from Eq 2.57 in blue, with the measurements from the redshift $z = 0.2$ snapshot of a high resolution N-body simulation run with Planck2018 cosmology (ABACUSSUMMIT suite) in black. The estimators used for the measurement of our observables (black dots) will be described in section 2.2.3. The errorbars are computed from 25 realisations of the simulations. The bottom panels show the residuals between the linear models and the measurements for different scales. We see that the difference between the model and the simulation exceeds 1σ for scales around 50 to 60 h^{-1}Mpc. Mildly to non-linear scales below the cut-off scale are quite sensitive to $f\sigma_8$ and possess a high SNR (for a given volume, there are fewer available large modes). However, the clustering on those scales is no longer well described by linear theory, and cannot be used to extract cosmological information.

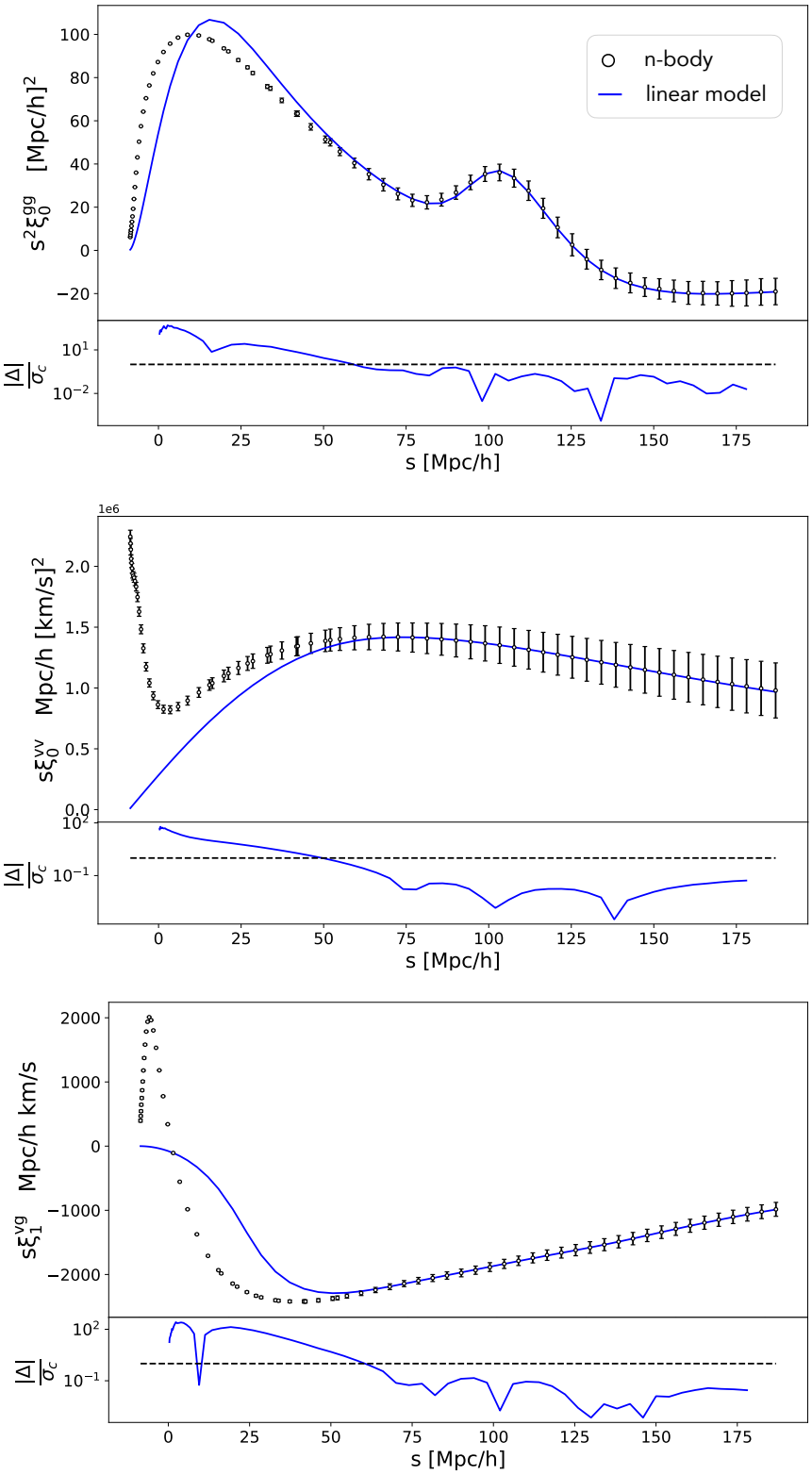


FIGURE 2.4 : Galaxy and velocity auto and cross-correlation functions comparison between linear modelling (blue) and high resolution N-body simulation ABACUSSUMMIT (black). We only show the monopoles and the dipole for the auto and cross-correlations respectively, computed with the Flip implementation. The sub-panels show the residuals with errorbars estimated from 25 realisations of the simulation.

2.1.4 Beyond linear theory

The analytical description of LSS as small density perturbations to the homogeneous Universe is no longer valid below a certain cut-off scale. As redshift decreases, matter has more time to cluster and increasingly larger scales become non-linear. An important point is that while the matter density field becomes fully non-linear on galactic scales, gravity remains weak, and the perturbation in the metric can still be treated linearly. To describe the distribution of these non-linear perturbations, we need to revisit the work done in section 1.3.5 by solving the Vlasov-Poisson system (Eq 1.68). On small scales, due to the discrete nature of matter, it can no longer be considered as an ideal fluid. Instead, we can treat it as an effective fluid, accounting for the stress tensor $\sigma_{ij} \neq 0$ as an expansion in terms of the matter density field itself.

One way to solve Eq 1.68 analytically is to expand the perturbation fields at higher order. In Fourier space we can write

$$\delta_m(k,\eta) = \delta_m(k,\eta)^{(1)} + \delta_m(k,\eta)^{(2)} + \delta_m(k,\eta)^{(3)} + ..., \tag{2.60}$$

where higher indices designate higher order terms. Doing so, one can solve the equations through an iterative approach, replacing in the equations the higher order terms by the solution derived for the previous order. The velocity field gradient can be expanded similarly, with the first order term connected to $\delta_m(k,\eta)^{(1)}$ through linear theory (Eq 1.76). In the end, any high order contribution of the density and velocity fields can be written as a function of the linear perturbation $\delta_m^{(1)}$. This widely used formalism offers an elegant way of pushing down the cut-off scale of the theory, allowing for the description of the clustering on smaller scales, ultimately leading to tighter constraints on cosmological parameters (see BERNARDEAU et al. 2002 for more details). An important consequence of the non-linear evolution of matter perturbations is that the linear description of the matter field evolution in Eq 1.70 breaks down. In the non-linear regime, the scale and time dependence of the matter field can no longer be separated into two independent components, and the growth factor becomes a scale dependent function $D(k,z)$, even without considering massive neutrinos or deviation from GR. Therefore, the measured quantity f is the linear limit at large scales of the growth rate of structures $f(k,z)$. Similarly, the matter-galaxy connection also becomes non-linear on small scales and Eq 2.3 is non longer valid. Consequently, the perturbative approach must be extended to higher orders of the matter density field, introducing additional bias parameters to the model. Small-scale baryonic feedback can also deviate the velocity field of luminous matter with respect to dark matter, introducing a set of velocity biases. Finally, RSD can no longer be treated linearly as the redshift induced by the peculiar velocity of galaxy $z_p = u_r/c$ becomes significantly large for small scales. Standard perturbative theories attempt to model mildly non-linear scales analytically (TARUYA et al. 2010; REID et M. WHITE 2011; CARLSON et al. 2013; L. WANG et al. 2014). Even more modern approaches (including loop corrections), such as the effective field theory of large-scale structure (EFToLSS, PAJER et ZALDARRIAGA 2013; CARRASCO et al. 2014), try to push the model validity to smaller scales, but ultimately such perturbative approaches are limited as the density perturbations on small scales become too large, especially at low redshifts.

While primordial non-Gaussianity in the distribution of matter perturbations - predicted by some inflationary models (KOMATSU et SPERGEL 2001) - is an active field of research, it has not yet been detected. However, even assuming Gaussian initial conditions as we have done so far, the non-linear evolution described above introduces some skewness, i.e. some non-Gaussianity. Consequently, two-point statistics are no longer sufficient to fully characterise the distribution

of the perturbed fields. Higher-order statistics must be considered to extract the complete cosmological information. The first of these higher-order statistics is the three-point function, or the bispectrum in Fourier space defined as

$$\langle \delta_g(\boldsymbol{k_1})\delta_g(\boldsymbol{k_2})\delta_g(\boldsymbol{k_3})\rangle = (2\pi)^3 B(\boldsymbol{k_1},\boldsymbol{k_2},\boldsymbol{k_3})\delta^D(\boldsymbol{k_1}+\boldsymbol{k_2}+\boldsymbol{k_3}). \tag{2.61}$$

Similarly to the two-point function, on small scales those higher order statistics can no longer be treated perturbatively. An alternative to perturbation theory is numerically solving Eqs 1.68 by running cosmological simulations.

2.1.5 Cosmological simulations

2.1.5.1 Dark Matter only

The most commonly used kind of cosmological simulations are dark matter simulations, also known as N-body simulations. Although computationally expensive to produce, brute iterative solving of dynamical equations in N-body simulations provide a high-fidelity description of the non-linear clustering of matter. With this method, the represented dark matter fluid is sub-sampled into point- like fluid element that we call "particles", allowing to describe the phase-space distribution function. The fixed-mass collision-less particles are evolved in a periodic cubic box in comoving space by imposing a gravitational integration under the Newtonian approximation. This way, the particles position and velocity can be predicted for a given redshift, successfully reproducing the non-linear structures and the filamentary pattern that we observe in our Universe. Naturally, with this method, density fluctuations on scales larger than the box dimension cannot be represented. While early cosmological simulations (M. Davis et al. 1985a) used 32^3 particles, state-of-the-art simulation suites such as the AbacusSummit suite (Maksimova et al. 2021) can reach 6912^3 particles. Solving the dynamical equation for a high number of particles allows for either increasing the simulation resolution - by reducing the dark matter particles mass (typically $10^9 - 10^10 M_\odot/h$) - or increasing the simulation volume, thereby accessing larger modes and reducing cosmic variance for smaller modes. Simulations that go beyond the ΛCDM model or involve modified gravity are generally more computationally intensive, as they often require solving a non-linear variant of the Poisson equation. This makes them considerably more demanding than traditional N-body Vlasov-Poisson solvers. Alternative dark matter models may require incorporating new treatments for local interactions, or may not be compatible with an N-body approach at all, such as in the case of fuzzy dark matter models. Although we will not use any beyond ΛCDM-GR simulation in this work, we refer the readers to Angulo et O. Hahn 2022 for further details on the topic and on dark matter simulations in general.

In order to evolve in time the discretised cold collision-less fluid, one needs to set up some primary distribution. As we discussed earlier, the late times non-linear structures formed from the gravitational evolution of small fluctuations around homogeneity in the early Universe. Those initial fluctuations are assumed to be described by a Gaussian homogeneous isotropic random field, that can be written

$$\delta_0(\boldsymbol{k}) = \phi(k) S(\boldsymbol{k}) \tag{2.62}$$

with $\phi(k) = (2\pi)^{3/2}\sqrt{P_L(k)}$ the fluctuations' amplitude, and $S(\boldsymbol{k})$ a complex random "noise" field

describing the spatial distribution such that

$$S^*(\boldsymbol{k}) = S(-\boldsymbol{k}), \quad \langle S(\boldsymbol{k})\rangle = 0, \quad \text{and } \langle S(\boldsymbol{k})S^*(\boldsymbol{k}')\rangle = \delta_D(\boldsymbol{k}-\boldsymbol{k}'). \tag{2.63}$$

This definition ensures that the initial power spectrum is defined as $\langle \delta_0(\boldsymbol{k})\delta_0^*(\boldsymbol{k}')\rangle = (2\pi)^3 P_L(k)\delta_D\left(\boldsymbol{k}-\boldsymbol{k}'\right)$. The field $S(\boldsymbol{k})$ can then be decomposed as

$$S(\boldsymbol{k}) = A(\boldsymbol{k})e^{i\theta(\boldsymbol{k})}, \tag{2.64}$$

where the phase θ follows a uniform distribution on $[0, 2\pi]$, and the modulus A follows a Rayleigh distribution (equivalent to a χ^2 distribution with two degrees of freedom since it is the quadratic sum of two Gaussian fields). Once the initial density fluctuation field $\delta_0(\boldsymbol{k})$ has been specified, it is evolved to a smaller redshift - typically $z \approx 100$ - using perturbation theory (density fluctuations are small at early times). Dark matter particle positions and velocities can then be sampled on the three-dimensional grid, before numerically solving for the dynamical equations. Doing so reduces the number of time steps required to evolve the field to redshift $z = 0$ (or later). The density field is then transformed and evolved in configuration space. However, because the IC are randomly sampled, the random field $S(\boldsymbol{k})$ fluctuates over several realisations. Therefore, the power spectrum associated with the initial field δ_0 is affected by cosmic variance (although it is independent of θ). The expectation value of the two point statistics can be obtained by averaging several realisation of the "same" simulation with different IC. As this method is computationally expensive, different approaches have been developed to reduce cosmic variance without using many realisations (Angulo et Pontzen 2016b; Hadzhiyska et al. 2023a). Yet, many realisations can be produced with fast approximate methods such as EZmocks (Chuang et al. 2015a). Those methods typically use perturbation theory with corrections calibrated on higher precision simulations or observational data. While this approach does not accurately describe clustering on small non-linear scales, it is useful for estimating the variance of observables and testing various observational effects.

2.1.5.2 Simulating a galaxy sample

Because galaxy formation results from complex baryonic processes, which are not included in dark matter only simulations, galaxies have to be added using some post-processing algorithm. A recent approach (Pellejero Ibañez et al. 2022) to generate galaxies from a dark matter-only N-body simulation involves defining the galaxy field through a bias expansion of the numerically evolved dark matter field. Since we can estimate how large the next higher-order contribution is, this hybrid approach provide a rigorous control over mismodelling systematics errors. However, it fails to describe galaxy clustering on scales typically above $k \approx 0.6h/\text{Mpc}$.

Following the halo model formalism, empirical approaches have been developed to populate dark matter halos with galaxies. Halos can be identified using various techniques. The spherical over-density method (Press et Schechter 1974) identifies density peaks and expands spherical shells until the over density falls below a specified threshold, usually determined by the spherical collapse model. Friends-of-Friends algorithms (M. Davis et al. 1985b) links particles that are within a certain distance of each other. The halo occupation distribution (HOD) method (Berlind et Weinberg 2002) populates halos with a given number of galaxies N, following a probabilistic distribution $P(N|M)$ as a function of the halo's total mass M. In a given halo, galaxies are categorized into central and satellite types, each following different distributions. The free

parameters of these distributions can be directly fitted on observations. We give more details on this framework and its application in Chapter 4.

The abundance matching technique (KRAVTSOV et al. 2004) instead assumes a direct relationship between galaxy properties (observed from real data) and dark matter halos in simulations. In this method, both halos and galaxies are ranked by a specific property (such as mass), and associated one-to-one. For sufficiently high-resolution N-body simulations, gravitationally bound structures within halos, known as sub-halos, can be identified. The formalism has subsequently been extended to subhalo abundance matching (BEHROOZI et al. 2010), assuming that the most luminous or massive galaxies are found in the largest bound substructures of halos. These models typically allow for some scatter in the galaxy-halo relationship. Although this method has fewer free parameters than the HOD model, it remains an empirical description of galaxy clustering.

The galaxy formation process relies on the ability of the gas to cool down in order to form stars. Semi-analytical models (SAM) of galaxy formation (S. D. M. WHITE et FRENK 1991) puts into equation the flows between stars, hot gas and cold gas reservoirs, providing a physical treatment of gas cooling, star formation, supernova feedback, metallicity enrichment, and galaxy merging. Applied to N-body simulations, SAM predict the number of galaxies per halo. Because SAM are based on simplifying assumptions, they must be tested against on observed data.

Hydrodynamical simulations provide the most accurate modelling for the complex interplay between baryonic and dark matter. In these simulations, baryons are treated as an additional fluid, and similarly to N-body codes, the dynamical equations for dark matter and baryons can be solved jointly. This can be done with baryons sampled as discrete particles or evolved on a grid. In addition to handling the baryonic gas dynamics, hydrodynamical simulations incorporate various physical processes, such as gas cooling, star formation, stellar feedback (e.g., supernovae), active galactic nuclei (AGN) feedback, radiation, supermassive black holes, magnetic fields, etc. When these processes occur on scales below the simulation resolution, they require effective treatments, often tuned directly to observations or through semi-analytic models (SAM). Figure 2.5 provide a visual representation of the predictions for different N-body and hydrodynamical simulations. Increasing the resolution of a simulation - by reducing the particle mass and increasing the number of particles - allows the modelling of physical processes on smaller scales, eventually enabling the description of complex mechanisms such as galaxy formation. However, to keep computational costs tractable, high-resolution simulations are often run in smaller volumes. This constraint is particularly significant for hydrodynamical simulations, as they involve multiple fluids and shorter time steps. Running simulations with varying sizes and resolutions - down to galactic scales — enable to effectively implement processes occurring on scales smaller than the simulation resolution, by tuning phenomenological descriptions based on higher-resolution simulations. A more comprehensive overview of the methodology and applications of hydrodynamical simulations can be found in VOGELSBERGER et al. 2020.

All methods discussed so far for simulating a galaxy sample depend on simplified assumptions about the connection between galaxies and dark matter. Although each method can be validated against more physically grounded models, the precision of these assumptions is challenging to verify, since the free parameters (including the cosmological parameters) can potentially absorb the mismodelling. Therefore, it is crucial to cross-checks these different methods in a cosmological analysis. Cosmological simulations are extensively used to test the reliability and validity range of analytical modelling, build covariance matrix for the observable, and assert the robustness of pipeline analysis by reproducing systematics effects introduced in the observation process.

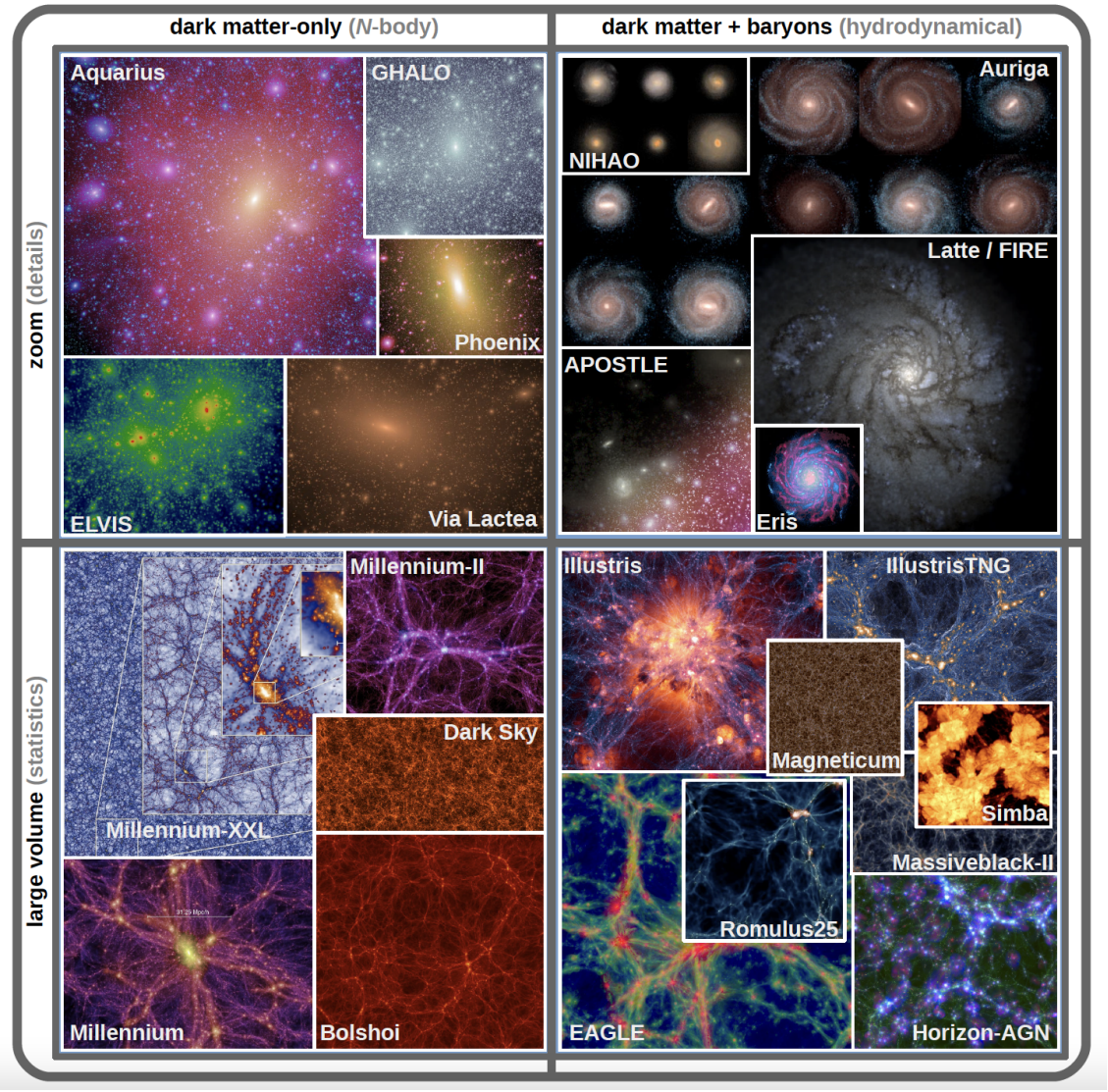


FIGURE 2.5 : Non-exhaustive list of cosmological and galaxy formation simulations. The simulations are divided in N-body (dark matter only) and hydrodynamical (dark matter + baryons). For both types, the simulations are also divided in large volume simulations providing statistical samples of galaxies, and zoom simulations resolving smaller scales in more detail. The figure is taken from VOGELSBERGER et al. 2020

2.2 Galaxy and peculiar velocity surveys

2.2.1 Construction of 3D galaxy maps

Creating 3D maps of galaxies for cosmological analysis involves several key steps. The first stage consists in a photometric survey, which images a large portion of the sky in order to identify galaxies and measure their angular positions. Photometric surveys employ multiple filters, providing rough spectral information though crucial to distinguish between galaxies and stars from our own galaxy. In order to avoid spurious correlations, it is important to select galaxies that form a homogeneous, high-density sample over the observed part of the sky. Different types of galaxies can be distinguished based on their mass, colour, and approximate photometric

redshift (derived from the galaxy's colours in different filters). Figure 2.6 describes the DESI Legacy Imaging Survey in then optical r band. Covering ~ 14000deg^2, it is composed of the BASS and MzLS imaging the north galactic cap (NGC), DECaLS imaging part of the NGC and south galactic cap (SGC), and the external DES survey in the SGC initially designed to measure weak lensing data.

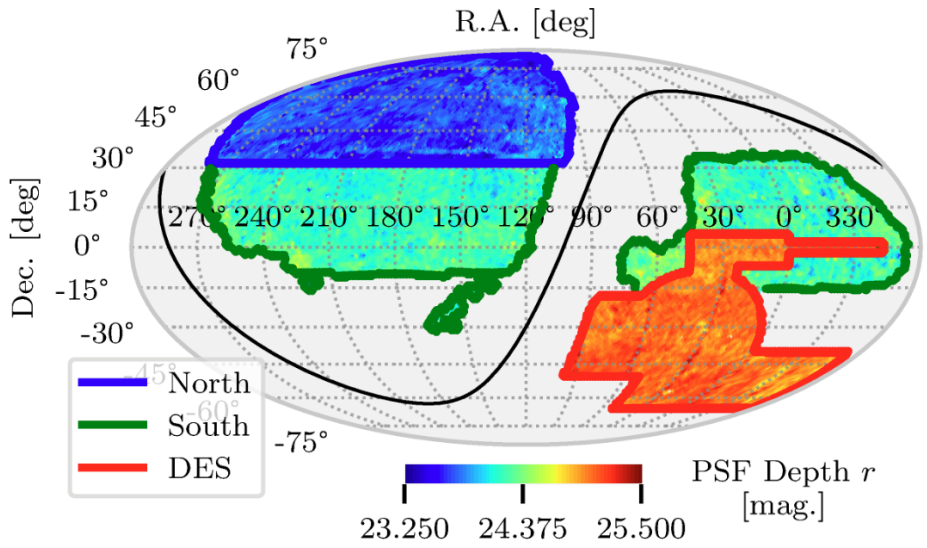


FIGURE 2.6 : Sky coverage of the DESI Legacy Imaging Survey in the r band, composed of the Bass, MzLS, DECaLZ and DES photometric surveys. Figure taken from CHAUSSIDON et al. 2023.

Spectroscopic measurements are necessary for estimating galaxy redshifts accurately. Measuring the spectra of galaxies, is significantly more time-consuming than imaging. Fibre-based telescopes, such as the SDSS and DESI, require precise locations of targets in the sky to place their fibres. Moreover, preliminary photometric observations provide rough estimation of the targets redshifts. Therefore, careful photometric selection of targets for spectroscopy is essential. Instruments like the Sloan Digital Sky Survey (SDSS) or the Dark Energy Spectroscopic Instrument (DESI) measure the redshifts of millions of galaxies by focusing light onto a focal plane where hundreds to thousands of optical fibres simultaneously captures the flux of the corresponding targeted galaxies. A dispersive optical system decomposes the light from the fibres into several wavelengths, which are then redistributed on CCD cameras and analysed by a spectrograph that convert the electron flux into physical spectra. The observing strategy for a spectroscopic survey is usually based on the focal plane dimensions, the field of view of the telescope, the amount of fibres, the total observing time (typically five years) and exposure times (depends on the SNR), the target priority (used for different scientific purpose), and the fraction of targets receiving an optical fibre (completeness). Different types of galaxies present features, such as absorption or emission lines, that make them convenient to probe different epochs of the cosmic history. Figure 2.7 represent the four DESI tracers within a thin slice the Uchuu lightcone simulation (PRADA et al. 2023). Quasars are extremely luminous and distant objects powered by supermassive black holes at their centres. They are used to probe the universe at high redshifts $1 < z < 5$ (including Lyman-alpha forest quasars). Emission line galaxies (ELG) are star-forming galaxies presenting strong emission lines (OII, OIII) in the optical, typically located around redshift $z \sim 1$. Luminous

red galaxies (LRG) are massive, old galaxies with red colours, selected due to their strong 4000 Å break feature, which makes them easily identifiable. LRG are used to map the universe at intermediate redshifts $z \sim 0.6$. The bright galaxy sample (BGS) is a flux-limited sample with r-band magnitude typically below 19.5. Bright galaxies are relatively nearby and help in mapping the local universe.

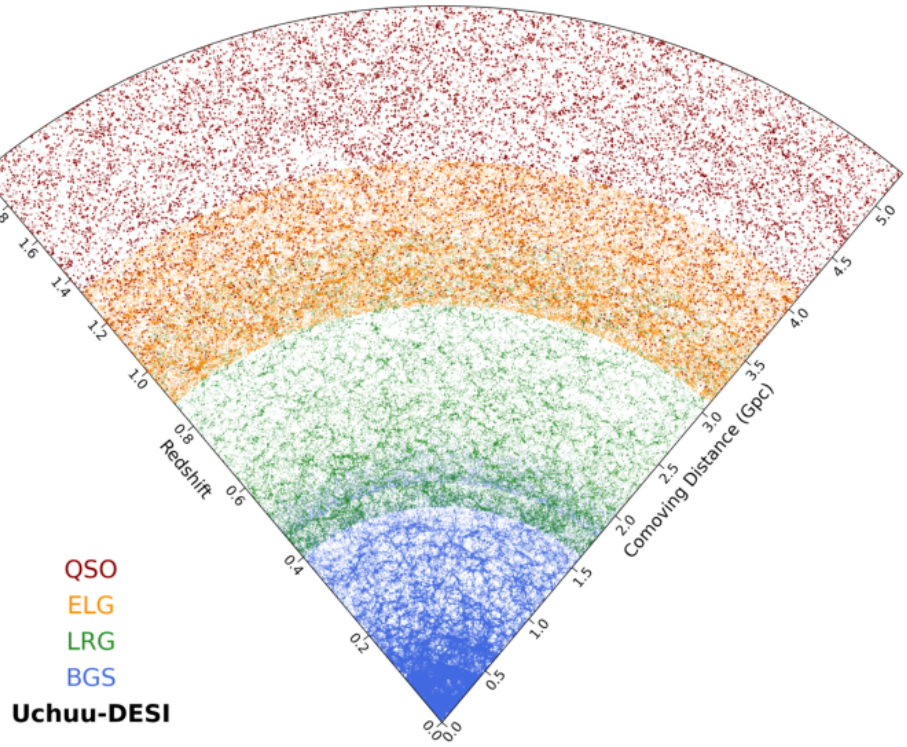


FIGURE 2.7 : Slice through the Uchuu simulation reproducing the DESI different types of targeted galaxies. The different type of tracers are BGS (blue), LRG (green), ELG (orange), QSO (red). Figure taken from PRADA et al. 2023.

Raw 1D spectra are extracted from 2D CCD images, flux-calibrated using spectra of standard stars, and have a sky model removed. Redshift fitting involves identifying the spectral lines and determining how much they have shifted due to the expansion of the universe. Absorption and emission lines in the spectra allow to measure redshifts, along with their uncertainties, by fitting physical templates to the calibrated fluxes. Once the redshifts are determined, the galaxies' three-dimensional positions can be constructed from the angular positions (right ascension and declination) combined with the redshift as described in Section 2.1.2.1. In April 2024 the DESI collaboration released a series of publications describing the results obtained with the first year of data taking. Figure 2.8 show the resulting largest three-dimensional map of our Universe, constructed from more than 12 millions galaxies redshift measurements.

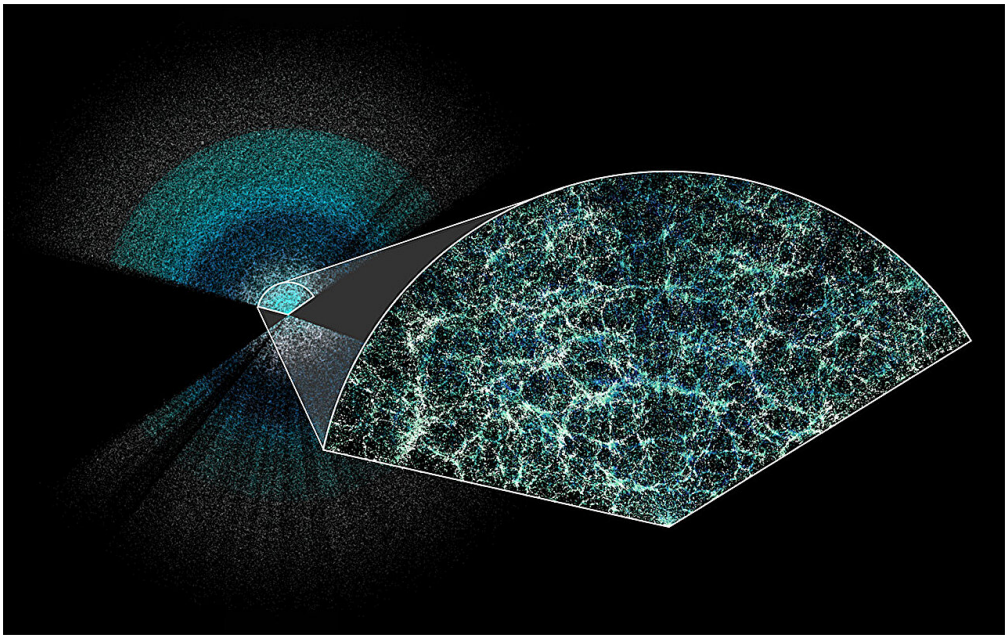


FIGURE 2.8 : Three-dimensional map of our Universe structures from the first year of DESI data taking. The magnified slice correspond to the BGS sample. Figure taken from desi.lbl.gov.

2.2.2 Measuring Galaxy peculiar velocities

As discussed in section 2.1.2.2, the peculiar velocity of a galaxy is imprinted in its observed redshift trough equation 2.27. Measuring the cosmological redshift $z_{\rm cos}$ requires a measurement of the source distance or age which is independent of the observed redshift. Combined with the spectroscopically measured redshift $z_{\rm obs}$, we can derive $z_{\rm p}$ and consequently the radial peculiar velocity. Measuring distance of individual objects on cosmological scales is quite challenging, and is mostly done through three types of standardisable objects : Type-Ia supernovae (SnIa), blue spiral galaxies, and red elliptical galaxies.

SnIa are standardisable candles thanks to the strong correlations between their luminosity and their light-curves colour and duration. By themselves, SnIa are one of the most important cosmological probe of the Universe's expansion, allowing to constraint the Hubble parameter and the Dark Energy EoS from the distance ladder. Through high accuracy distance measurement - about 7% relative error (BORUAH et al. 2020) - they provide the most precise peculiar velocity measurements. Photometric surveys with large sky coverage, such as the Zwicky Transient Facility (ZTF, BELLM et al. 2019) and the Rubin Observatory Legacy Survey of Space and Time (LSST, LSST SCIENCE COLLABORATION et al. 2009) will observe a large and uniform sample of peculiar velocity from SnIa.

Similarly, Tully-Fisher(TF) and fundamental-plane(FP) relations are standardisable empirical relations allowing to estimate distances. While the resulting relative distance error is larger (about 20%), it is compensated by the large amount of TF and FP-based velocity measurements that DESI will obtain (about 186000 SAULDER et al. 2023b). As we will study the velocity clustering derived from the TF and FP relations in Chapter 6, we describe the measurement methodology below.

2.2.2.1 Tully-Fisher relation

Late type galaxies (LTG) are young, blue, star forming (relatively to their masses) spiral galaxies. They formed in lower-density environments where major mergers are less frequent, thus preserving their disk structure. While in practice, their shape is used to identify them, they are physically characterised by their rotational energy being larger than their internal kinetic energy, arising from the velocity dispersion inside the galaxy (CHISARI et al. 2015). The population of LTG show a strong correlation between their absolute magnitude M and their maximum rotation velocity $V_{\rm max}$ measured along the line of sight (in km.s^{-1}). This correlation - which can be physically understood in terms of the virial theorem - is known as the Tully-Fisher (TF) relation R. B. TULLY et FISHER 1977

$$M = a \log V_{\rm max} + c, \tag{2.65}$$

with (a, c) the TF coefficients. In the following we adopt the convention $v_{\rm max} = \log V_{\rm max}$, and we consider the r-band absolute magnitude M_r (SAULDER et al. 2023b). The maximal rotational velocity can be estimated by sampling the rotation curve, placing fibres at different parts of a galaxy's disk and measuring the Doppler shift of emission lines. The TF relation can be calibrated (fitting a and c) on a catalogue of galaxies with known absolute magnitude, derived from known distances (Clusters, Cepheids, etc). This provides a predictive model for the absolute magnitude of any spiral galaxy from its maximum rotation velocity. Furthermore, M_r can also be expressed as

$$M_r(z) = m_r - 5\log\left(\frac{D_L(z)}{1h^{-1}\mathrm{pc}}\right) + 5, \tag{2.66}$$

where m_r is the photometrically observed magnitude, and $D_L(z)$ the luminosity distance in h^{-1}pc. The redshift z can be spectroscopically measured. Because of RSD, the peculiar velocity of a given galaxy will affect its observed magnitude. Following Eqs 2.27 and 2.66, we can define the luminosity log-distance ratio as

$$\eta_L \equiv \log\frac{D_L(z_{\rm obs})}{D_L(z_{\rm cos})} = 0.2\left(M_r(z_{\rm cos}) - M_r(z_{\rm obs})\right). \tag{2.67}$$

Hence once calibrated, the TF relation allows estimating $M_r(z_{\rm cos})$ thus the log-distance ratio η_L by measuring $v_{\rm max}$, m_r and $z_{\rm obs}$.

2.2.2.2 Fundamental plane relation

Early type galaxies (ETG) are old, red, quenched elliptical galaxies where star formation has stopped. They typically formed in high-density environments, undergoing complex interactions and mergers with other galaxies. Since they lost most of their cold gas reservoir, they are no longer forming stars. Contrary to LTG, they are dominated by their internal velocity dispersion rather than their rotational energy. ETG galaxies follow an empirical relation between their physical effective spherical radius R (in h^{-1}kpc) enclosing half the total luminosity, central velocity dispersion σ_V (in km.s^{-1}) and surface brightness I (in $L_\odot$ kpc^{-2}) averaged within R, known as the Fundamental Plane (FP) relation DJORGOVSKI et Marc DAVIS 1987

$$\log R = a \log\sigma_V + b \log I + c, \tag{2.68}$$

with (a,b,c) the FP coefficients. The surface brightness is a distance-independent quantity defined as

$$\begin{aligned} I &= \frac{L}{\pi R^2} \\ &= (1+z_{\rm obs})^4 \frac{1}{\pi\theta_e^2} 10^{10-0.4m} \left(\frac{648}{\pi}\right)^2 \end{aligned} \tag{2.69}$$

with L the luminosity (in $L_\odot$) of the galaxy, m the magnitude and θ_e the angular effective radius (in arcseconds). In the following we adopt the conventions $r = \log R$, $s = \log\sigma_V$ and $i = \log I$. The FP relation can be calibrated on a galaxy catalogue with known physical effective radius, derived from known distances. The physical radius of any elliptical galaxy can then be predicted by measuring s and i. Moreover, r can also be expressed as a function of effective angular radius θ_e and redshift

$$r(z) = \log\theta_e + \log\left(\frac{D_A(z)}{1h^{-1}{\rm Mpc}}\right) + \log\frac{\pi}{648}, \tag{2.70}$$

where $D_A(z)$ is the angular diameter distance in h^{-1}Mpc and $\frac{\pi}{648}$ is a unit conversion factor. As the spectroscopically measured redshift is affected by peculiar velocity, one can derive following Eqs 2.27 and 2.70 the angular log-distance ratio as

$$\eta_A \equiv \log\frac{D_A(z_{\rm obs})}{D_A(z_{\rm cos})} = r(z_{\rm obs}) - r(z_{\rm cos}). \tag{2.71}$$

The calibrated FP relation allows to estimate $r(z_{\rm cos})$, and therefore η_A by measuring s, i,θ and $z_{\rm obs}$.

2.2.2.3 From log-distance ratio to peculiar velocity

Once the log-distance ratio is known, deriving the peculiar velocity is straightforward. From Eq 2.27, the comoving radial distance can be expanded at linear order in $z_p = u_r/c$

$$D(z_{\rm cos}) = D(z_{\rm obs}) + \frac{c}{H(z_{\rm obs})}(z_{\rm cos} - z_{\rm obs}), \tag{2.72}$$

leading to the distance ratio

$$\frac{D(z_{\rm cos})}{D(z_{\rm obs})} = 1 - \frac{(1+z_{\rm obs})}{D(z_{\rm obs})H(z_{\rm obs})} u_r. \tag{2.73}$$

Finally, we can write the comoving radial distance ratio as

$$\begin{aligned} \eta &\equiv -\log\frac{D(z_{\rm cos})}{D(z_{\rm obs})} \\ &= -\log\left[1 - \frac{(1+z_{\rm obs})}{D(z_{\rm obs})H(z_{\rm obs})} u_r\right] \\ &= \alpha(z_{\rm obs}) u_r, \end{aligned} \tag{2.74}$$

where the last equality is only valid at low redshift. The proportionality factor is defined as

$$\alpha(z_{\rm obs}) \equiv \frac{1}{\ln 10}\frac{(1+z_{\rm obs})}{D(z_{\rm obs})H(z_{\rm obs})}. \tag{2.75}$$

Analogous relations can be derived for the luminosity and angular log-distance ratios by writing

$$\begin{aligned}\frac{D_L(z_{\rm cos})}{D_L(z_{\rm obs})} &= \frac{D(z_{\rm cos})}{D(z_{\rm obs})}\frac{1}{1+z_p},\\ \frac{D_A(z_{\rm cos})}{D_A(z_{\rm obs})} &= \frac{D(z_{\rm cos})}{D(z_{\rm obs})}\left(1+z_p\right),\end{aligned} \tag{2.76}$$

leading to

$$\begin{aligned}\eta_L &= \alpha_L(z_{\rm obs})u_r,\\ \eta_A &= \alpha_A(z_{\rm obs})u_r,\end{aligned} \tag{2.77}$$

with

$$\begin{aligned}\alpha_L(z_{\rm obs}) &\equiv \frac{1}{c\ln 10}\left[\frac{(1+z_{\rm obs})c}{D(z_{\rm obs})H(z_{\rm obs})}+1\right],\\ \alpha_A(z_{\rm obs}) &\equiv \frac{1}{c\ln 10}\left[\frac{(1+z_{\rm obs})c}{D(z_{\rm obs})H(z_{\rm obs})}-1\right].\end{aligned} \tag{2.78}$$

It is worth noting that in the regime $\frac{(1+z_{\rm obs})c}{D(z_{\rm obs})H(z_{\rm obs})} >> 1$, we have $\eta \approx \eta_L \approx \eta_A$. Beside observational uncertainties, the TF and FP relations are characterised by an intrinsic scatter (around the true magnitudes and radius respectively), ultimately propagating to the peculiar velocity measurements. Recent TF catalogues Hong et al. 2019; R. Brent Tully et al. 2023 found an average scatter of around 22% in distance ratio. The SDSS FP peculiar velocity catalogue describes a scatter of 23% in distance ratio Saulder et al. 2013; Howlett et al. 2022. Since $\alpha(z)$ is a decreasing function, for a given uncertainty σ_η - corresponding to a constant scatter in the TF/FP relations - the errors on peculiar velocity become increasingly large with redshift. Because of the intrinsic scatter in the standardisable objects, peculiar velocity measurements are limited to low redshifts (also true for SnIa), typically up to $z \sim 0.1$.

2.2.3 Two-point statistics estimators

Once galaxy and peculiar velocity catalogues are defined, any statistics - such as the two-points correlation function or the power spectrum - can be estimated to compare with their corresponding theoretical prediction. In this section, I describe the estimators used in this work as well as different sets of weights and corrections applied to real data to account for observational systematic effects.

2.2.3.1 Optimal weighting

The optimal weighting scheme minimizing the variance of the galaxy two-points statistics estimators for a non-uniform redshift distribution was derived in Feldman et al. 1994. The so-called FKP weights are given by

$$w_g^{\rm FKP} = \frac{1}{P_g\ \bar{n}(z)+1} \tag{2.79}$$

where $\bar{n}(z)$ is the comoving number density in redshift bins, and P_g corresponds to the galaxy power spectrum at some scale of interest, relevant for the studied signal (typically $10^4\ h^{-3}$Mpc^3 for th BAO scale $k \sim 0.1\ h^{-1}$Mpc). Similar weights have been used to reduce the variance of the

velocity clustering measurements (QIN et al. 2019) :

$$w_v^{\mathrm{FKP}} = \frac{1}{P_v\ \bar{n}(z) + \sigma_v^2(z)} \tag{2.80}$$

where $\sigma_v(z) = \sigma_\eta / \alpha(z)$ is the uncertainty of the peculiar velocity measurement, and P_v is the characteristic amplitude of the velocity power spectrum (typically $10^9\ h^{-3}\mathrm{Mpc}^3\ \mathrm{km}^2\mathrm{s}^{-2}$).

Selection effects in the photometric observations can lead to spurious fluctuations in the density field. Imaging weights w_{photo} must be included to correct for potential correlation between galaxy density and non-cosmological photometry-related quantities such as imaging depth, sky flux, galactic extension etc.

Spectra with low SNR might fail to yield confident redshift measurements. If the failure pattern is not homogeneous in the sky, it can contaminate clustering measurements. As it is often correlated with the geometry of the focal plane, redshift failures weights w_{failure} must be included to correct for this effect. Those weights can be defined as proportional to the inverse of the modelled average failure rate as a function of the position within the focal plane.

The redshift of some galaxies, angularly too close to each other, cannot be observed because of the physical size of the optical fibres. This effect typically reduces clustering on scales corresponding to fibre sizes projected on the sky. Fibre-collision weights w_{fc} can be defined by up-waiting a galaxy by the number of near non-observed targets, by removing pairs with small transverse to the LoS separations, or with more sophisticated pair-wise weighting scheme (GUO et al. 2012 ; PERCIVAL et BIANCHI 2017 ; C. HAHN et al. 2017). The total weights assigned to each of the observed galaxy or velocities is then obtained as the combined product

$$w_{\mathrm{tot}} = w^{\mathrm{FKP}} \times w_{\mathrm{photo}} \times w_{\mathrm{failure}} \times w_{\mathrm{fc}}. \tag{2.81}$$

2.2.3.2 Random catalogues

Random catalogues are commonly used to describe the window function of the survey $W_g(\boldsymbol{x})$, equal to 1 if $\boldsymbol{x}$ is within the survey volume and 0 otherwise (in the simplest case). Fake galaxies (called randoms) are randomly drawn from a uniform distribution within the angular footprint, following the photometric survey specific geometry. Afterwards, redshifts sampled from the observed galaxy distribution are randomly attributed to the randoms. The constructed random catalogue of three-dimensional positions $n_s(\boldsymbol{x})$ is then used to describe the product $\bar{n}_g W_g(\boldsymbol{x})$, with $\bar{n}_g$ the mean density of galaxies. In order to properly describe the survey volume and reduce noise in the window function, random points are sampled at a higher number density than the observed galaxies (typically a few to 50 times higher, depending on the galaxy density). The number of random points must then be normalised in any derived statistics to avoid overestimating $\bar{n}_g$. As peculiar velocity and galaxy survey do not always overlap and have different mean densities, a different random catalogue should be built for the subsample of galaxies with peculiar velocity measurements. Due to the observing strategy and time constraints, not all targets have a fibre assigned to them, so their redshifts are not measured. This incompleteness can be corrected by sub-sampling the random catalogue in the concerned regions of the sky, or by attributing fibre completeness weights to the randoms, defined as the ratio between the number of observed redshift to the number of available targets in each region. FKP weights defined in Eq 2.79 are also assigned to the random catalogues.

2.2.3.3 Correlation function estimator

We expose now the derivation of the galaxy and velocity correlation functions. The clustering of peculiar velocity in configuration space is usually described using correlations between components of velocity parallel and perpendicular to the separation vector, as introduced in GORSKI 1988. However, those estimators are built on the hypothesis that the velocity perturbations are linear, and only extract the monopole and quadrupole information. Moreover, it requires computationally expensive pairwise weighting. As we will focus on non-linear clustering in chapters 4, 5, and 6, we derive more agnostic estimators for the velocity-velocity and velocity-galaxy correlation functions. We provide a complete derivation of the state-of-the-art Landy-Szalay estimator (LANDY et SZALAY 1993) for the galaxy two-points correlation function (TPCF), and use the same methodology to build estimators for the velocity auto-correlation and cross-correlation functions.

The correlation function of two random fields $\delta_{\,\mathrm{a}}$, $\delta_{\,\mathrm{b}}$ is defined as

$$\xi_{\,\mathrm{ab}}(\boldsymbol{x}_1 - \boldsymbol{x}_2) \equiv \langle \delta_{\,\mathrm{a}}(\boldsymbol{x}_1)\delta_{\,\mathrm{b}}(\boldsymbol{x}_2)\rangle, \tag{2.82}$$

The observed galaxy spatial distribution can be described as a discrete sum $n_g(\boldsymbol{x}) = \sum_i^{N_g} \delta^{\mathrm{D}}(\boldsymbol{x} - \boldsymbol{x_i})\ w_{g,i}$, with δ^{D} the Dirac delta function, N_g the number of observed galaxies, and $w_{g,i}$ the total weight associated to a given galaxy i. For a complete sample without any observational effect, all weights are set to 1. In order to derive an estimator for the galaxy correlation function, we can also write the observed galaxy density as

$$n_g(\boldsymbol{x}) = \bar{n}_g W_g(\boldsymbol{x}) \left(1 + \delta_{\mathrm{g}}(\boldsymbol{x})\right) \frac{1}{w_g(\boldsymbol{x})}. \tag{2.83}$$

As mentioned before, the mean density and the survey footprint are defined through the random catalogue as $n_s(\boldsymbol{x}) = \bar{n}_g W_g(\boldsymbol{x})$. A number of random samples N_s a few ten times larger than N_g is usually enough to properly sample the footprint. The random catalogue can also have associated weights $w_s(\boldsymbol{x})$. Rewriting the previous relation we define the galaxy over-density field as

$$\delta_{\,\mathrm{g}}(\boldsymbol{x}) = \frac{w_g(\boldsymbol{x})n_g(\boldsymbol{x}) - w_s(\boldsymbol{x})n_s(\boldsymbol{x})}{w_s(\boldsymbol{x})n_s(\boldsymbol{x})}. \tag{2.84}$$

As discussed in section 2.1.3, in configuration space, the underlying density field fluctuation is extracted from the observed galaxy field, so there is no need to consider the convolution with the window function in the modelling. Following Eq 2.82, we square the field and assume ergodicity so the ensemble averaging $\langle . \rangle$ is equivalent to a spatial averaging. Doing so we can derive the well known Landy-Szalay estimator (LANDY et SZALAY 1993) for the galaxy TPCF

$$\xi_{gg}(\boldsymbol{r}) = \frac{DD(\boldsymbol{r}) - DS(\boldsymbol{r}) - SD(\boldsymbol{r})}{SS(\boldsymbol{r})} + 1, \tag{2.85}$$

where $\boldsymbol{r} = (\boldsymbol{x}_1 - \boldsymbol{x}_2)$ and DD, SS, DS and SD are the normalised auto and cross pair counts of the

data and random catalogues given by

$$\begin{aligned}
DD(\boldsymbol{r}) &= \langle n_g(\boldsymbol{x}_1) n_g(\boldsymbol{x}_2)\rangle = \frac{1}{N_{\rm gg}} \sum_i^{N_g} \sum_{j\neq i}^{N_g} \delta^{\rm D}(\boldsymbol{r} - |\boldsymbol{x}_{g,i} - \boldsymbol{x}_{g,j}|)\ w_{g,i} w_{g,j}, \\
SS(\boldsymbol{r}) &= \langle n_s(\boldsymbol{x}_1) n_s(\boldsymbol{x}_2)\rangle = \frac{1}{N_{\rm ss}} \sum_i^{N_s} \sum_{j\neq i}^{N_s} \delta^{\rm D}(\boldsymbol{r} - |\boldsymbol{x}_{s,i} - \boldsymbol{x}_{s,j}|)\ w_{s,i} w_{s,j}, \\
DS(\boldsymbol{r}) &= \langle n_g(\boldsymbol{x}_1) n_s(\boldsymbol{x}_2)\rangle = \frac{1}{N_{\rm gs}} \sum_i^{N_g} \sum_{j}^{N_s} \delta^{\rm D}(\boldsymbol{r} - |\boldsymbol{x}_{g,i} - \boldsymbol{x}_{s,j}|)\ w_{g,i} w_{s,j}, \\
SD(\boldsymbol{r}) &= \langle n_s(\boldsymbol{x}_1) n_g(\boldsymbol{x}_2)\rangle = \frac{1}{N_{\rm sg}} \sum_i^{N_s} \sum_{j}^{N_g} \delta^{\rm D}(\boldsymbol{r} - |\boldsymbol{x}_{s,i} - \boldsymbol{x}_{g,j}|)\ w_{s,i} w_{g,j}.
\end{aligned} \tag{2.86}$$

The auto and cross normalisation factors for two catalogues a and b are defined as

$$\begin{aligned}
N_{\rm aa} &= \left(\sum_i^{N_a} w_{a,i}\right)^2 - \sum_i^{N_a} w_{a,i}^2, \\
N_{\rm ab} &= \left(\sum_i^{N_a} w_{a,i}\right)\left(\sum_j^{N_b} w_{b,j}\right).
\end{aligned} \tag{2.87}$$

Those pair counts are usually separated in (s,μ) bins, allowing to numerically compute the TPCF multipoles from

$$\xi_l(s) = \frac{2l+1}{2} \int_{-1}^{+1} \xi(s,\mu) P_l(\mu) d\mu. \tag{2.88}$$

In the case of BAO analysis where the catalogues have been reconstructed as described in section 2.1.3.1, the estimator is modified to account for the shifted positions of both galaxies and randoms

$$\tilde{\xi}_{gg}(\boldsymbol{r}) = \frac{\tilde{D}\tilde{D}(\boldsymbol{r}) - \tilde{D}\tilde{S}(\boldsymbol{r}) - \tilde{S}\tilde{D}(\boldsymbol{r})}{SS(\boldsymbol{r})} + 1, \tag{2.89}$$

where $\tilde{.}$ indicates the shifted positions after reconstruction.

Following the same methodology we can derive Landy-Szalay-like estimators for the radial velocity auto-correlation and the galaxy-velocity cross-correlation. As in large-scales surveys only the radial component of the galaxy velocity field can be measured, in the following any use of "velocity" implies "radial velocity". Since peculiar velocities can be measured only where galaxies are observed, the field sampled is the galaxy momentum field $p_g(\boldsymbol{x})$. Its spatial distribution can be written as a discrete sum $p_g(\boldsymbol{x}) = \sum_i^{N_v} \delta^{\rm D}(\boldsymbol{x} - \boldsymbol{x}_i)\ u_{g,i}\ w_{v,i}$, with N_v the number of peculiar velocity measurements, $u_{g,i}$ and $w_{v,i}$ the velocity and weight respectively evaluated at position $\boldsymbol{x}_i$. The galaxy momentum can also be expressed as

$$p_g(\boldsymbol{x}) = \bar{n}_v W_v(\boldsymbol{x})\ u_g(\boldsymbol{x})\ \left(1 + \delta_{\rm g}(\boldsymbol{x})\right)\ \frac{1}{w_v(\boldsymbol{x})}, \tag{2.90}$$

with $\bar{n}_v$ the mean density of velocity measurements and W_v the window function describing the peculiar velocity survey footprint. Again, we can use a catalogue $n_r(\boldsymbol{x})$ of N_r random samples to describe $\bar{n}_v W_v(\boldsymbol{x})$. We call the weights associated with the random catalogue $w_r(\boldsymbol{x})$. We rewrite

the previous relation to introduce the galaxy weighted velocity field as

$$v(\boldsymbol{x}) = u_g(\boldsymbol{x})\ \left(1+\delta_{\,\mathrm{g}}(\boldsymbol{x})\right) = \frac{w_v(\boldsymbol{x})\,p_g(\boldsymbol{x})}{w_r(\boldsymbol{x})\,n_r(\boldsymbol{x})}. \tag{2.91}$$

Once again, following Eq 2.82 we can define the auto and cross correlation function by respectively squaring Eq 2.91 and multiplying it by Eq 2.84 :

$$\xi_{vv}(\boldsymbol{r}) = \frac{VV(\boldsymbol{r})}{RR(\boldsymbol{r})}, \tag{2.92}$$

$$\xi_{vg}(\boldsymbol{r}) = \frac{VD(\boldsymbol{r}) - VS(\boldsymbol{r})}{RS(\boldsymbol{r})}, \tag{2.93}$$

Following the same convention for the pair-counts and the normalisation. Note that ξ_{gv} encapsulates as much information as ξ_{vg} and therefore is not used here. In the following we refer to the "galaxy weighted velocity field" as the "velocity field" since they are equivalent in linear theory. However, the reader should keep in mind that according to our definition, the velocity correlation function written as

$$\xi_{vv}(\boldsymbol{r}) = \langle u_g(\boldsymbol{x}_1)\ \left(1+\delta_{\,\mathrm{g}}(\boldsymbol{x}_1)\right) u_g(\boldsymbol{x}_2)\ \left(1+\delta_{\,\mathrm{g}}(\boldsymbol{x}_2)\right)\rangle \tag{2.94}$$

should encapsulate on non-linear scales the information contained in ξ_{gg}, ξ_{uu},ξ_{ug} and higher order statistics.

The most expensive part of computing those estimators is the random-random pair counting. When working with simulated cubic boxes with periodic boundary conditions, under the flat-sky approximation, SS can be analytically computed as

$$SS(\mathbf{r}) = \frac{N_{\mathrm{gal}}\left(N_{\mathrm{gal}}-1\right)}{L_{\mathrm{box}}^3}\left[\frac{4\pi\left(s_2^3-s_1^3\right)}{3}\right]\left[\mu_2-\mu_1\right], \tag{2.95}$$

with N_{gal} and L_{box} the number of galaxies and the size of the simulation box respectively, s_1 and s_2 the lower and upper limits of the separation bins, and μ_1 and μ_2 the lower and upper limits of the angular bins. Furthermore, the galaxy-random pair counts tends to $DS \approx SD \approx SS$, and the velocity-random pair count vanishes $VS \approx 0$. For a cubic box mock where the velocity is known for every galaxy, we can use a unique random catalogue $n_s = n_r$, and make use of the natural estimator PEEBLES P. J. E. 1974 and its analogue for peculiar velocities :

$$\begin{aligned}\xi_{gg}(\mathbf{r}) &= \frac{DD(\mathbf{r})}{SS(\mathbf{r})} - 1, \quad \xi_{vv}(\mathbf{r}) = \frac{VV(\mathbf{r})}{SS(\mathbf{r})},\\ \xi_{vg}(\mathbf{r}) &= \frac{VD(\mathbf{r})}{SS(\mathbf{r})}.\end{aligned} \tag{2.96}$$

We will use those different TPCF estimators [4] in every following Chapters to measure the galaxy clustering, and in Chapter 5 and 6 to measure the peculiar velocity clustering.

[4] Our own implementation is available in the VEGACORR package

2.2.3.4 Power Spectrum estimator

The two-points statistics can also be estimated in Fourier space. We first need to define the over-density function, also known as the FKP field introduced in FELDMAN et al. 1994 as

$$F(\boldsymbol{x}) = w_g(\boldsymbol{x}) \left[n_g(\boldsymbol{x}) - \alpha_r n_s(\boldsymbol{x}) \right] / I^{1/2}, \tag{2.97}$$

where α_r is the ratio between the total weighted number of galaxies and the total weigthed number of randoms, and the factor normalising the amplitude of the measured power spectrum is defined as

$$I \equiv \int dx^3 \left[w_g(\boldsymbol{x}) n_g(\boldsymbol{x}) \right]^2. \tag{2.98}$$

The galaxy power spectrum multipoles can then be written as

$$P_\ell(k) = \frac{(2\ell+1)}{I} \int \frac{d\Omega_k}{4\pi} \int dx_1^3 F(\boldsymbol{x_1}) e^{-i\boldsymbol{k}\cdot\boldsymbol{x_1}} \int dx_2^3 F(\boldsymbol{x_2}) e^{+i\boldsymbol{k}\cdot\boldsymbol{x_2}} \mathscr{L}_\ell(\hat{k}\cdot\hat{n}), \tag{2.99}$$

where the ensemble average has been replaced by averaging over spherical shells of radius k in k-space. Note that for the power spectrum estimator, the fluctuation field estimated is convolved with the window function and must be taken into account in the modelling, as described in section 2.1.3. The full survey volume is embedded into a cubic box with periodic boundary conditions, and subdivided it into grid cells. The mesh used must be large enough to encompass the survey volume, and apply some padding on the edges. Once assigned to the regular grid, the random and observed galaxies can easily be Fourier transformed using Fast Fourier Transform (FFT) algorithms. The scales probed are then defined by the resolution of the grid. For a comoving size L on each side and N^3 voxels, the fundamental mode is $\frac{\pi}{L}$, while the Nyquist mode is $\frac{\pi N}{L}$. For a given volume, the mesh size N has to be chosen to cover scales of interest. Moreover, following YAMAMOTO et al. 2006a, one can choose one of the galaxies of each pair as the line of sight (only valid in the flat sky approximation), allowing to perform the two FT integral in Eq 2.99 separately, using FFT methods such that

$$P_\ell(k) = \frac{(2\ell+1)}{I} \int \frac{d\Omega_k}{4\pi} \left[F_0(\boldsymbol{k}) F_\ell(-\boldsymbol{k}) - P_\ell^{\rm noise} \right], \tag{2.100}$$

with the noise term only contributing to the monopole as

$$P_\ell^{\rm noise} = \delta_{0,\ell} \int dx^3 w_g(\boldsymbol{x}) n_g(\boldsymbol{x})^2. \tag{2.101}$$

$F_0(\boldsymbol{k})$ is a simple Fourier transform which can be computed with FFT. Several techniques exist to compute $F_\ell(\boldsymbol{k})$ as a simple Fourier transform form (BIANCHI et al. 2015a; SCOCCIMARRO 2015a). We will use this estimator in Chapter 3.

The momentum power spectrum can be estimated analogously. Further details can be found in HOWLETT 2019

2.2.4 Observational effects

Integral constraint

So far, we have considered that the expected density of observed galaxies at any position $\boldsymbol{x}$ in the absence of clustering was described by a random catalogue $n_s(\boldsymbol{x}) = \bar{n}W(\boldsymbol{x})$. However, the background number density of galaxies $\bar{n}$ is in practice unknown, and it has to be estimated from a finite survey volume. Therefore, it may differ from the true value due to clustering modes with scales approaching the survey size. By construction, the integral of the inferred density fluctuations over the full survey is forced zero, leading to a so-called integral constraint (IC). The IC results in an additive bias in the two-points statistics, since the fluctuation field is underestimated from the absence of large scales modes $\delta_{\rm large}$

$$n_g(\boldsymbol{x}) = \bar{n}_g W_g(\boldsymbol{x}) \left(1 + \delta_{\rm g}(\boldsymbol{x}) - \delta_{\rm large}\right) \frac{1}{w_g(\boldsymbol{x})}, \tag{2.102}$$

where by definition the mean density(fluctuation) $\delta_{\rm large}$ is constant within the survey volume. Although this effect can be theoretically modelled (Scranton et al. 2002; Wilson et al. 2017; de Mattia et Ruhlmann-Kleider 2019), we do not consider it in this work as it is mostly relevant for large scales.

Bulk flow

The peculiar velocity clustering could suffer from a similar bias if the survey is sampled on top of a large velocity mode, inducing a mean velocity also known as *bulk flow*. In that case the observed velocity field has a non-vanishing expectation value. Because we can only estimate the density-weighted velocity field, the resulting clustering could also suffer from IC. We can write the observed momentum field as

$$p_g(\boldsymbol{x}) = \bar{n}_v W_v(\boldsymbol{x}) \left(u_g(\boldsymbol{x}) - u_{\rm bf}\right) \left(1 + \delta_{\rm g}(\boldsymbol{x}) - \delta_{\rm large}\right) \frac{1}{w_v(\boldsymbol{x})}, \tag{2.103}$$

where by definition, the bulk flow $u_{\rm bf}$ is constant within the survey volume. Note that if the flat sky approximation breaks down, the bulk flow correction is no longer homogeneous as it is function of the LoS orientation. While another bulk flow could in practice be induced by our own motion in the Universe, one should be able to correct it from the measured CMB dipole induced by the motion of the solar system barycentre with respect to the CMB rest frame (Lineweaver et al. 1996; Hinshaw et al. 2009). Previous work already explored the possibility to constraint our bulk motion from SnIa observations (Feindt et al. 2013). We leave this issue to future work and consider that peculiar velocities are sampled over a volume large enough, and are corrected from our own peculiar motion.

Selection effect

Distance - thus velocity - measurements based on standardisable objects suffer from different kind of biases. Because of the intrinsic scatter of empirical relations (see previous section), a selection criterion such as a SNR or magnitude cut could bias the average distance of objects close to this selection boundary. The work of Strauss 1995 further describes the so-called Malmquist bias. The estimated distances might be biased by some non-uniform distributions of objects in

the distance bins. A homogeneous bias arises from the fact that the number of detected objects scales with the distance squared. An inhomogeneous malmquist bias might also be caused by density fluctuation along the LoS. Indeed, around structures such as clusters, the number of object in a given distance bin will not be uniformly distributed. It is therefore essential to properly account for those biases to correct peculiar velocity measurements.

Wide-angle effect

Because galaxy peculiar velocity is a vector field, its two-point statistics is anisotropic. This effect propagates to the galaxy two-points statistics through RSD. The flat sky approximation assumes an infinitely distant observer, so that any pair of object shares a unique LoS $\hat{n}$. As such, any TPCF (in configuration space) depends on the separation vector $\boldsymbol{s} = \boldsymbol{x_1} - \boldsymbol{x_2}$. Assuming azimuthal symmetry around the LoS, it can be parametrised by $s = |\boldsymbol{s}|$ and $\mu = \hat{s} \cdot \hat{n}$. However, when s is no longer negligible with respect to the object distances $|\boldsymbol{x_1}|$, $|\boldsymbol{x_2}|$, the flat sky approximation is no longer valid. Because each galaxy has its own LoS, the TPCF depends on the triangles formed by each pair of galaxies and the observer, described by s, $\mu_1 = \hat{s} \cdot \hat{x_1}$, and $\mu_2 = \hat{s} \cdot \hat{x_2}$. It is convenient to use the parametrisation $\xi(s, \mu, \theta)$, with μ the cosine of the angle between the observer LoS - passing through the midpoint between the galaxies - and the separation vector of the galaxy pair, and θ the angle between the two individual lines of sight. In our estimator described previously, we neglected this angular dependence by projecting the TPCF in the two dimensions (s, μ). Meaning that we are estimating (YOO et SELJAK 2015)

$$\xi(s, \mu) = \frac{\int \xi(s, \mu, \theta) P(s, \mu, \theta) d\theta}{\int P(s, \mu, \theta) d\theta}, \tag{2.104}$$

where $P(s, \mu, \theta)$ is the probability distribution of triangle configuration, described by the normalised random-random pair counts. Since the dependence on θ - containing cosmological information - mainly concern the largest scales, this compression should not lead to a loss of information for the scales of interest. For distant survey, $P(s, \mu, \theta) = P(s, \mu)\delta^D(\theta)$ since for any pair $\theta \to 0$, and we retrieve the flat sky approximation with $\xi(s, \mu) = \xi(s, \mu, \theta = 0)$. For low-redshift survey (and large separation bins), $P(s, \mu, \theta)$ deviates from a Dirac distribution along θ, and $\xi(s, \mu) \neq \xi(s, \mu, \theta = 0)$. If the θ dependence is not modelled and properly integrated out, the clustering can be biased. This is commonly referred to as the wide-angle effects. We will carefully check this effect on the measured clustering in Chapters 4 and 5.

2.3 Cosmological Inference

We have described how some observables - physically motivated by the theoretical framework - can be estimated from observations, and modelled either numerically or analytically. Moreover, systematic effects can be either included in the modeling or corrected in the measurements. In this section inspired from the review TROTTA 2017, we describe how different theories can be quantitatively tested and their parameters statistically inferred from data.

2.3.1 Likelihood analysis

Bayesian statistics have been widely used in cosmology to analyse and interpret observational data, ultimately seeking for the best theoretical description of our Universe. Bayes theorem

states that, given some n_d-dimensional data vector $\boldsymbol{d}$, the probability distribution for the n_p-dimensional parameter vector $\boldsymbol{\theta}$ of a model $\mathcal{M}$ [5] is given by

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{d})}. \tag{2.105}$$

The conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$ is commonly called the *posterior* probability, representing our degree of belief about the value of $\boldsymbol{\theta}$ after observing the data $\boldsymbol{d}$. An essential ingredient of Bayesian statistics is the *prior* probability distribution $p(\boldsymbol{\theta})$, describing our knowedge of $\boldsymbol{\theta}$ before measuring the dataset. The normalisation factor $p(\boldsymbol{d})$ is called the *evidence* (or marginal-likelihood). It is defined as

$$p(\boldsymbol{d}) = \int d\boldsymbol{\theta}\, p(\boldsymbol{d}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{2.106}$$

and is important for Bayesian model comparison. The last quantity in Eq 2.105 is the *likelihood* function, often denoted $\mathscr{L}(\boldsymbol{d}|\boldsymbol{\theta}) \equiv p(\boldsymbol{d}|\boldsymbol{\theta})$, describing the likelihood of the observed data given a certain value of the parameters.

While recent approaches have been developed for non-Gaussian likelihood analyses (C. HAHN et al. 2019), the common approach assumes a multivariate Gaussian distribution. Indeed, if the studied perturbation δ is a random multivariate Gaussian field, the power spectrum of a given mode $\langle|\delta|^2(k)\rangle$ follows a gamma distribution. However, averaging over a sufficient number of modes (valid for high SNR), the power spectrum would tend towards a Gaussian distribution by the Central Limit Theorem. In the realistic case of a non-Gaussian perturbation field, although the power spectrum no longer fully describes its distribution, this limit is still valid. The Gaussian likelihood is then given by

$$\mathscr{L}(\boldsymbol{d}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n_d/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{d}-\boldsymbol{\mu}(\boldsymbol{\theta}))^\top C^{-1}(\boldsymbol{d}-\boldsymbol{\mu}(\boldsymbol{\theta}))\right), \tag{2.107}$$

where the measurements fluctuations are described by the covariance $C \equiv \langle(\boldsymbol{d}-\langle\boldsymbol{d}\rangle)(\boldsymbol{d}-\langle\boldsymbol{d}\rangle)^\top\rangle$, and the expectation value $\langle\boldsymbol{d}\rangle = \boldsymbol{\mu}(\boldsymbol{\theta})$ is given by the parameter-dependent theoretical predictions. While in practice the covariance matrix is also parameter-dependent, it has been shown (CARRON 2013) that cosmological analyses of two-point functions are more robust when fixing the covariance for a set of fiducial parameters $C(\boldsymbol{\theta}_{\rm fid})$. The treatment of the covariance matrix is discussed in details in section 2.3.3.

The principle of the statistical inference is then to find the set of parameters $\boldsymbol{\theta}^*$ maximising the posterior $p(\boldsymbol{\theta}|\boldsymbol{d}) \propto \mathscr{L}(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, ignoring the evidence for now. Since we set the prior to be a known analytical distribution, the goal of the inference is to maximise $\mathscr{L}$ with respect to $\boldsymbol{\theta}$ such that

$$\left.\frac{\partial\mathscr{L}}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*} = 0, \quad \text{and} \quad \left.\frac{\partial^2\mathscr{L}}{\partial\boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}^*} < 0. \tag{2.108}$$

In practice there are only few situations where this can be done analytically, and it is more convenient to numerically minimise the so-called chi-squared (also known as the mahalanobis distance)

$$\chi^2 = -2\ln\mathscr{L} + \text{cste} \propto (\boldsymbol{d}-\boldsymbol{\mu}(\boldsymbol{\theta}))^\top C^{-1}(\boldsymbol{d}-\boldsymbol{\mu}(\boldsymbol{\theta})). \tag{2.109}$$

[5] In our case, the theory $\mathcal{M}$ is the ΛCDM model, the n_p parameters to infer are $\boldsymbol{\theta} = \{H_0, \Omega_m, \sigma_8, w, ...\}$, and the data vector of dimension n_d is typically a concatenation of two-point statistics multipoles over a range of scales $\boldsymbol{d} = [\xi_0(s_1), \xi_0(s_2), ..., \xi_2(s_1), \xi_2(s_2), ..., P_0(k_1), P_0(k_2), ..., P_2(k_1), P_2(k_2), ...]$.

Parameters confidence levels

The confidence levels (uncertainty) on the best fitted parameters can be obtained by exploring the posterior shape in the vicinity of the best fitted parameters. While a wide distribution around $\boldsymbol{\theta}^*$ results in large uncertainties, a peaked distribution describes strong constraints on the parameter values. Note that the posterior and the likelihood are two distinct quantities with different meaning (even with a flat uniform prior making them functionally identical). While the distribution of data points described by the likelihood is assumed to be n_d dimensional multivariate Gaussian, the posterior distribution of the parameters can be anything (again independently of the prior), and is no guaranteed to be a n_p dimensional multivariate Gaussian. For most well-defined problems however, it is fortunately the case. For a one dimensional parameter vector, the uncertainty Δ over the best fitted parameter θ^* is obtained by integrating the posterior from each side of θ^*, until a certain confidence interval is reached

$$\int_{\theta^*-\Delta}^{\theta^*+\Delta} p(\theta|\boldsymbol{d})d\theta = CL. \tag{2.110}$$

Generally the errorbars written after a "±" sign correspond to a 68.3% confidence interval, equivalent to one standard deviation for a Gaussian distribution. This defines a 1σ uncertainty, but it can be generalised to any number of σ, according to a certain confidence level. The multidimensional case is analogous and usually represented as a set of two-dimensional CL called "contours", describing the correlations between different parameters. Individual parameters' confidence level can be obtained by integrating the posterior probability

$$p(\theta_1|\boldsymbol{d}) = \int p(\boldsymbol{\theta}|\boldsymbol{d})d\theta_2...d\theta_n. \tag{2.111}$$

Note that it also allows to marginalise over nuisance parameters. We give further details on how these calculations are performed in practice done in section 2.3.4.

Goodness of a model

Beside constraining the model parameters, it is essential to assess the validity of the model. The χ^2 defined in Eq 2.109 can be written as a sum of $n_d - n_p$ independent Gaussian random variables (the covariance matrix can be diagonalised). It is therefore a random variable following a chi-squared distribution $p(\chi^2)$ with $n_d - n_p$ degrees of freedom (dof), expectation value $\langle\chi^2\rangle = n_d - n_p$ and variance $\mathrm{var}(\chi^2) = 2(n_d - n_p)$. A commonly used quantity in cosmology is the reduced chi-squared defined as ${}_r\chi^2 = \chi^2/(n_d - n_p)$ which has an expectation value of 1. However, since a given deviation from unity does not quantify the same fit goodness for different dof, one can use a more quantitative indicator, the so-called p-value. It is defined as the probability to measure a chi-square value larger than the one estimated χ^2_* :

$$p-\text{value} = \int_{\chi^2_*}^{+\infty} p(\chi^2)d\chi^2. \tag{2.112}$$

It is important to precise that this quantity does not correspond to the probability that the model is valid. Instead, it describes how likely the observations are, assuming the model true and the parameters fixed at $\boldsymbol{\theta}^*$. Therefore, a small p-value allows to discard a model.

In cosmology, since several Dark Energy or alternative gravity models have been developed to

solve the tensions, it is quite convenient to compare them to choose the 'best' theory. We already mentioned that the evidence was a normalisation factor to the posterior overall amplitude. While the evidence is parameter independent, it is model-dependent and should be written as $p(\boldsymbol{d}|\mathscr{M})$ when considering several models. Applying once again Bayes' theorem to invert the conditioning order, we find that the evidence is proportional to the posterior probability of a given model $p(\mathscr{M}|\boldsymbol{d})$. Therefore, the model with the largest evidence is favoured by the data. For two models, the ratio of their evidences is called the Bayes factor.

Bayesian and frequentist

It is worth mentioning that an animated debate exists among statisticians about the use of priors. The frequentist approach focuses on the likelihood function, treated as a proper probability distribution. The choice of prior is considered as subjective, besides, the parameters are not considered as random variables but rather fixed unknown quantities. For Bayesians, the use of "standard" priors is an objective approach to reflect our state of knowledge, updating it with new data to enforce constraints. However, the choice of prior must be physically motivated. Even flat uniform prior - the posterior becomes proportional to the likelihood - can have strong implications. Under a change of variable $\theta \rightarrow \Theta(\theta)$, according to the conservation of probability, the prior becomes

$$p(\Theta) = p(\theta) \left| \frac{\partial \theta}{\partial \Theta} \right|, \tag{2.113}$$

and is no longer necessarily flat. This effect is especially important for high dimensional spaces. Yet, for a sufficiently large data set, the posterior distribution usually converges to a prior-independent regime.

2.3.2 Fisher forecast

Fisher forecast in cosmology is a crucial tool for predicting the expected uncertainties on cosmological parameters for future experiments or surveys, allowing to optimise their design and observing strategies. We can Taylor expand at second order the χ^2 around its minimum evaluate at $\boldsymbol{\theta}^*$ as

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \left. \frac{\partial \chi^2}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}^*} + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \left. \frac{\partial^2 \chi^2}{\partial \boldsymbol{\theta}^2} \right|_{\boldsymbol{\theta}^*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*). \tag{2.114}$$

Because the linear term vanishes by definition (Eq 2.108), the posterior distribution of the parameters $\boldsymbol{\theta}$ is a multivariate Gaussian - in the vicinity of $\boldsymbol{\theta}^*$ - fully characterised by its expectation value $\boldsymbol{\theta}^*$ and its $n_p \times n_p$ covariance matrix

$$C_\theta = F^{-1}, \tag{2.115}$$

where we have defined the Fisher matrix as

$$F_{ij} = - \left. \frac{\partial^2 \ln \mathscr{L}}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta}^*}. \tag{2.116}$$

The Fisher matrix is analogous to the log-likelihood curvature, measuring how fast it varies away from its maximum value. It also defines the Cramér–Rao bound, stating that the parameters' covariance matrix measured through an inference is bounded by $C_\theta \geq F^{-1}$. It can be shown that

for a Gaussian likelihood, the fisher matrix can be written (TEGMARK et al. 1997)

$$F_{ij} = \frac{\partial \boldsymbol{\mu}^\top}{\partial \theta_i} C^{-1} \frac{\partial \boldsymbol{\mu}}{\partial \theta_j} + \mathrm{Tr}\left[C^{-1} \frac{\partial C}{\partial \theta_i} C^{-1} \frac{\partial C}{\partial \theta_j} \right] \tag{2.117}$$

Hence, for a given configuration $\boldsymbol{\theta}^*$ and covariance matrix C, the parameter uncertainties can be estimated.

2.3.3 Covariance matrix

Estimating the uncertainty of an observable is crucial in physics. Errors determine the limits of the model's validity, and propagate to the constraints on cosmological parameters. Besides observational noise, we only have access to one unique realisation of our Universe, so any statistics derived from a cosmological random field suffers from cosmic variance. Averaging over a large volume helps reduce cosmic variance, but ultimately the available volume in our observable Universe is limited. Consequently, our observables — the two-point statistics multipoles — are random variables fluctuating around their expectation value, according to the distribution described in Eq 2.107. Because the distribution is Gaussian, it is fully characterised by the $n_d \times n_d$ covariance matrix with determinant $|C|$ and inverse C^{-1}.

Analytical covariance

The covariance matrix of the power spectrum for a generic random field δ (matter, galaxy or velocity fields) is defined as

$$\mathrm{Cov}\left[\hat{P}(\boldsymbol{k}), \hat{P}(\boldsymbol{k}')\right] \equiv \langle \hat{P}(\boldsymbol{k}) \hat{P}(\boldsymbol{k}') \rangle - \langle \hat{P}(\boldsymbol{k}) \rangle \langle \hat{P}(\boldsymbol{k}') \rangle. \tag{2.118}$$

If the field δ is assumed to be Gaussian, this covariance can be estimated following *Wick's theorem*. We can write the four-points function as

$$\begin{aligned} \langle P(\boldsymbol{k}) P(\boldsymbol{k}') \rangle = & \langle \delta(\boldsymbol{k}) \delta(-\boldsymbol{k}) \rangle \langle \delta(\boldsymbol{k}') \delta(-\boldsymbol{k}') \rangle \\ & + \langle \delta(\boldsymbol{k}) \delta(-\boldsymbol{k}') \rangle \langle \delta(-\boldsymbol{k}) \delta(\boldsymbol{k}') \rangle \\ & + \langle \delta(\boldsymbol{k}) \delta(\boldsymbol{k}') \rangle \langle \delta(-\boldsymbol{k}) \delta(-\boldsymbol{k}') \rangle, \end{aligned} \tag{2.119}$$

with the definition of the discrete power spectrum

$$\langle \delta(\boldsymbol{k}) \delta(\boldsymbol{k}') \rangle = (2\pi)^3 \delta^D(\boldsymbol{k} - \boldsymbol{k}') \left[P(\boldsymbol{k}) + P_N \right]. \tag{2.120}$$

The discrete definition of the galaxy field Fourier transform leads to the estimated power spectrum covariance

$$\mathrm{Cov}\left[\hat{P}(\boldsymbol{k}), \hat{P}(\boldsymbol{k}')\right] = \frac{2(2\pi)^3}{V} \delta^D(\boldsymbol{k} - \boldsymbol{k}') \left[P(\boldsymbol{k}) + P_N \right]^2, \tag{2.121}$$

where each mode is independent (diagonal covariance matrix) and V is the volume of the sample. As we observe a discrete sample of a continuous field, the covariance matrix is also affected by the shot noise term P_N, also known as sample variance [6]. High redshift surveys typically suffer

[6] Note that the difference between cosmic and sample variance is often ambiguous. In this work we stick to the convention where cosmic variance appears because we only observe one realisation of a random field, while sample variance refers to the discrete nature of the tracers observed.

more from sample variance, as it is difficult to detect a high number of galaxies. Low redshift uncertainties are dominated by cosmic variance, as most of the galaxies are observed (for an infinite dense sample $P_N \to 0$) but the volume is restricted. Of course those considerations neglect systematic observational errors. An analytical description of the galaxy power spectrum and correlation function multipole covariance matrices is given in GRIEB et al. 2016. Once again, this modelling is only valid on large scales since non-linearities render the matter/galaxy/velocity fields non-Gaussian, affecting the overall variance of the two point functions and correlating the different scales. Moreover, analytical descriptions fail to include precisely the impact of observational effects.

Simulation based covariance

A common solution - that we will use throughout this work - is to use simulations to estimate the covariance matrix from many realisations. Fast approximate methods fail to describe the covariance matrix on non-linear scales, but the observational effects can be forward modelled on the mock galaxy samples and included in the covariance matrix. Non-linear scales description requires more expensive N-body simulations. The unbiased estimate of the covariance matrix from $n_{\rm mock}$ realisations is given by

$$\hat{C}_{ij} = \frac{1}{n_{\rm mock}-1} \sum_{k}^{n_{\rm mock}} \left(d_i^k - \langle d_i \rangle\right)\left(d_j^k - \langle d_j \rangle\right) \tag{2.122}$$

where d_i^k is the kth realisation of the data vector d_i ($P_\ell(k_i)$ or $\xi_\ell(r_i)$). In practice, the quantity used in cosmological inferences is the inverse of the covariance matrix, known as the precision matrix $\Psi = C^{-1}$. The distribution of matrices computed from multiple realisations of independent simulations follows a Wishart distribution, and its inverse Ψ follows an inverse-Wishart distribution (TAYLOR et al. 2013). Since we only have an estimate of the covariance matrix and its inverse from Eq 2.122, we should marginalise over the possible values of Ψ in Eq 2.107. Because the marginalisation over every element of Ψ is challenging, approximate methods are used where the estimate of the true Ψ is used, leading to two corrections.
First, because of the skewed nature of the inverse Wishart distribution, the inverse covariance matrix estimate is biased. HARTLAP et al. 2007 showed that the unbiased estimator for the precision matrix was a function of the number of independent simulations $n_{\rm mocks}$ and the size of the data vectors n_d, introducing a corrective factor to the matrix inversion

$$\hat{\Psi} = (1-D)\hat{C}^{-1}, \quad \text{with} \quad D = (n_d+1)/(n_{\rm mocks}-1). \tag{2.123}$$

Second, the marginalisation over the noisy elements of Ψ propagates to the uncertainties on the cosmological parameters. It has been shown in DODELSON et SCHNEIDER 2013 that this additional noise is directly proportional to the parameters' covariance obtained after inference C_θ ($n_p \times n_p$). We therefore need to apply correction factors to the obtained C_θ. These factors are given in PERCIVAL et al. 2014 as :

$$m_1 = \frac{1+B(n_d-n_p)}{1+A+B(n_p+1)}, \tag{2.124}$$

with

$$A = \frac{2}{(n_{\text{mocks}} - n_d - 1)(n_{\text{mocks}} - n_d - 4)},$$
$$B = \frac{(n_{\text{mocks}} - n_d - 2)}{(n_{\text{mocks}} - n_d - 1)(n_{\text{mocks}} - n_d - 4)}. \tag{2.125}$$

If an inference is run on several mock data vectors that have been used to estimate the covariance matrix, the following factor must scale the variance of the best-fitted parameters distribution.

$$m_2 = \frac{m_1}{1 - D}, \tag{2.126}$$

with D the Hartlap factor defined above. We see that the number of mocks required to build a covariance matrix depends on the size of the data vector and on the number of parameters to infer. The production of mocks being computationally expensive, compressing the data vector is essential to prevent the covariance noise from propagating to parameter uncertainties. As we entered the big data era, balancing compression and information loss is an important issue in modern cosmology.

Estimating covariance matrices for cosmological inferences is an active field of research. A large panel of alternative methods exist. Covariance matrices can be directly estimated from the observed dataset. The Jackknife method (EFRON 1982) and its application to LSS (NORBERG et al. 2009 ; MOHAMMAD et PERCIVAL 2022) involves dividing the survey data into several subsamples and then systematically leaving out one subsample at a time, in order to create different realisations of the data vector. For each realisation, the statistic of interest is recalculated, and the covariance matrix can be estimated from those measurements. This method is useful, though it may underestimate the true covariance in some cases. Hybrid methods (PHILCOX et al. 2020) combines analytic approaches with numerical simulations to efficiently generate large numbers of mock catalogues. Non-linear shrinkage (LEDOIT et WOLF 2012 ; JOACHIMI 2017) typically correct from the ill-conditioning of the covariance matrix generated when the matrix dimension is close to the sample size ($n_d \approx n_{\text{mock}}$). It is worth mentioning that fluctuations on scales larger than the survey can also affect the covariance. Such modes increase the overall variance and correlate small-scale fluctuations. This effect can be described from the Trispectrum contribution neglected in Eq 2.119. It is known as the super sample variance (TAKADA et HU 2013). In our work, we expect the standard cosmic variance to dominate the error budget, therefore we will neglect this contribution.

2.3.4 Optimisation algorithms

Gradient descent

Minimising numerically a loss function such as our χ^2 defined in Eq 2.109 it a standard procedure in science, owing to the small number of problems simple enough to be amenable to analytic treatment. A widely used procedure for numerical minimization is based on gradient descent algorithms :

- Start from an initial guess for the parameters $\boldsymbol{\theta}_i$.
- Compute the gradient of the function with respect to each parameter (faster if analytically implemented).

- Update the parameters in the direction opposite to the gradient.
- Iterate until convergence towards the minimum.

Some popular algorithms incorporate random processes in order to avoid local minima. Once the loss function is minimised, it is possible to sample it to estimate CL for the parameters. Throughout this work, we will use the minimisation procedure implemented in the `Scipy` [7] and `Minuit` [8] packages. As we will see in Chapter 4, due to their computational efficiency and scalability, gradient descent algorithms are particularly useful in machine learning for training models.

Markov Chain Mote Carlo

Another popular optimisation method is the Markov Chain Mote Carlo (MCMC) method. The most used MCMC algorithm is the Metropolis-Hastings algorithm. For a specific proposal distribution $g(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})$, a symmetric Gaussian centred on $\boldsymbol{\theta}_i$ for example, it obeys the following steps

- Start from an initial guess for the parameters $\boldsymbol{\theta}_i$.
- Draw from the proposal distribution a new paramter configuration $\boldsymbol{\theta}_{i+1}$.
- Compute the ratio $r(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}) = \frac{p(\boldsymbol{\theta}_{i+1}|\boldsymbol{d})}{p(\boldsymbol{\theta}_i|\boldsymbol{d})}$.
- If $r > 1$ accept the step and add it to the chain.
- Otherwise draw a number α from a uniform distribution $U(O,1)$. If $r > \alpha$ accept the step, otherwise reject it and draw another proposal.

The probability of transition from the step i to $i+1$ is then given by $T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}) = r(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1}) g(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})$. One can show that the algorithm satisfies the detailed balance conditional

$$\frac{T(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i+1})}{T(\boldsymbol{\theta}_{i+1}, \boldsymbol{\theta}_i)} = \frac{p(\boldsymbol{\theta}_{i+1}|\boldsymbol{d})}{p(\boldsymbol{\theta}_i|\boldsymbol{d})}, \tag{2.127}$$

guaranteeing that the parameters sampling frequency of the algorithm is proportional to the posterior distribution. Once the chain has converged, it is trivial to compute the expectation values, CL, or any other information since we have sampled the posterior distribution. We will use the `emcee` [9] implementation in Chapters 4 and 5.

[7] https://scipy.org/
[8] https://scikit-hep.org/iminuit/index.html
[9] https://emcee.readthedocs.io/en/stable/

3 Baryon acoustic oscillations from the large-scale clustering of galaxies in Fourier and configuration space

Summary

3.1 Dataset . . . 88
3.1.1 The eBOSS survey . . . 88
3.1.2 Mock catalogues . . . 88
3.2 Methodology . . . 89
3.2.1 Measuring the clustering . . . 89
3.2.2 BAO modelling . . . 89
3.2.3 Parameter inference . . . 90
3.2.4 Consensus via Gaussian approximation . . . 91
3.2.5 Joint analysis . . . 92
3.3 Results on mock catalogues . . . 94
3.3.1 Fits on average correlations . . . 95
3.3.2 Varying the number of mocks . . . 96
3.3.3 Fits on individual mocks . . . 97
3.3.4 Including the hexadecapole . . . 103
3.4 Application to eBOSS DR16 data . . . 103
3.5 Conclusion . . . 105
Appendix B : pre-reconstruction results . . . 106
Appendix C : modelisation of the cross correlation pattern . . . 107

In the past decade, spectroscopic survey have given access to the three-dimensional distribution of galaxies, enabling us to study the expansion history of the Universe through BAO measurements. The latest cosmological constraints from a spectroscopic survey BAO measurements were produced by the DESI collaboration (D. COLLABORATION et al. 2024). The BAO measurements were extracted from the clustering of different types of tracers in the redshift range $0.1 < z < 2.33$ (DESI COLLABORATION et al. 2024a; DESI COLLABORATION et al. 2024b) : the magnitude-limited Bright Galaxy Survey for $0.1 < z < 0.4$, the Luminous Red Galaxies for $0.4 < z < 1.1$, the Emission Line Galaxies for $0.8 < z < 1.6$, the quasars for $0.8 < z < 2.1$, and the quasars with visible Lyman-α forests for $1.96 < z < 2.8$.Previously, BAO measurements have been conducted by the cosmological component of the fourth generation of the Sloan Digital Sky Survey (SDSS-IV, BLANTON et al. 2017), also known as the extended Baryon Oscillation Spectroscopic Survey (eBOSS, DAWSON et al. 2016). The eBOSS BAO measurements were obtained using luminous red galaxies at an effective redshift $z_{\rm eff} = 0.7$ (J. E. BAUTISTA et al. 2021; GIL-MARÍN et al. 2021), emission line galaxies at $z_{\rm eff} = 0.85$ (RAICHOOR et al. 2021; TAMONE et al. 2020; DE MATTIA et al. 2021), quasars as tracers of the matter field at $z_{\rm eff} = 1.48$ (HOU et al. 2021; NEVEUX et al. 2020), Lyman-α forests at $z_{\rm eff} = 2.33$ (DU MAS DES BOURBOUX et al. 2020). Prior to that, two lower redshift measurements were obtained at $z_{\rm eff} = 0.38$ and 0.51 from the third generation of SDSS, BOSS (EISENSTEIN et al. 2011; DAWSON et al. 2013; ALAM et al. 2017). The SDSS-II Main Galaxy Sample (ROSS et al. 2015b; HOWLETT et al. 2015) provided measurements at $z_{\rm eff} = 0.15$. Other surveys performed similar measurements, the 6dFGS (BEUTLER et al. 2011) at $z_{\rm eff} = 0.10$ and the WiggleZ Dark Energy Survey (KAZIN et al. 2014) on three redshift bins in the range $0.2 < z < 1.0$. Angular BAO measurements have also been produced by the photometric Dark Energy Survey (DES, DARK ENERGY SURVEY COLLABORATION et al. 2016), using three years of its data (DES COLLABORATION et al. 2021) in five redshift bins within $0.6 < z < 1.1$.

As we mentioned in Chapter 2, BAO measurements can be performed from the galaxy field two-point function, either in Fourier space with the power spectrum $P(\boldsymbol{k})$, or in configuration space with the two-point correlation function (TPCF) $\xi(\boldsymbol{s})$. Both types of analyses perform a statistical measurement of the galaxy distribution, starting from the same dataset, which is defined as a list of angular positions, redshifts, and weights, as well as some definition of the window function of the survey.In principle, both types of analyses should produce identical cosmological constraints since they share a common starting point. However, in practice, selecting a limited range of scales for fitting models slightly breaks this perfect degeneracy. Furthermore, correlation function is commonly done with pair-counting techniques, while power spectrum analysis assign galaxies into a three-dimensional mesh for Fast Fourier Transform. Noise properties or systematic effects on the estimated (and binned) statistics might differ between $\xi(\boldsymbol{s})$ and $P(\boldsymbol{k})$. These differences slightly reduce the correlations between the measured parameters in both spaces to levels below 100%. As an example, using the best-fit values for the dilation parameters on an ensemble of mock catalogues, J. E. BAUTISTA et al. 2021 show that BAO results in configuration and Fourier spaces are roughly 90% correlated. Therefore, there should be a slight statistical gain in producing a combined result. Systematic errors might as well be reduced by this combination, since they may manifest differently in both spaces. Collaboration groups often produce two separate analyses in a single space, making it difficult to decide which results to quote as final. Therefore, there are several advantages to produce joint results between Fourier and configuration spaces.

When producing a consensus result, it is important to take into account the strong correlations between measurements in configuration and Fourier space. SÁNCHEZ et al. 2017 described a method to compute a consensus result, used for the first time in SDSS-III BOSS (ALAM et al. 2017)

and extended to the latest measurements from SDSS-IV eBOSS survey. This method was used to produce consensus results on BAO and RSD analyses of a given survey, yielding constraints on D_H/r_d, D_M/r_d and $f\sigma_8$. However, the method from SÁNCHEZ et al. 2017 requires the individual likelihoods on BAO parameters to be Gaussian, which is not necessarily true, particularly in a low signal-to-noise ratio regime.

In this chapter, we develop a framework (DUMERCHAT et J. E. BAUTISTA 2022) to perform simultaneous BAO analyses in both Fourier and configuration spaces. The main advantages of our framework over past work are

1. Individual likelihoods on BAO parameters from each space are not assumed to be Gaussian.
2. The posterior distribution is not approximated to be Gaussian, resulting in more reliable uncertainties.
3. It is a straightforward Bayesian analysis.
4. It yields smaller systematic biases than the analyses in each individual space.

However, our method relies on a sufficiently good estimate of the full data-vector covariance matrix, particularly the cross-covariance between two-point functions in Fourier and configuration spaces. We use realistic mock catalogues to validate our method and performed comparison with previous work.

This chapter is organized as follows. In section 3.1, we introduce the dataset used. In section 3.2, we describe the BAO modelling, which is mostly the same used in J. E. BAUTISTA et al. 2021 ; GIL-MARÍN et al. 2021. Section 3.3 presents several statistical and systematic tests of both methods and finally the application on real data is described on section 3.4.

3.1 Dataset

3.1.1 The eBOSS survey

The extended Baryon Oscillation Spectroscopic Survey (eBOSS, DAWSON et al. 2016) was a 5-year observing program using multiobject fiber-fed spectrographs (SMEE et al. 2013) mounted on the focal plane of the 2.5 meter Sloan Foundation Telescope (GUNN et al. 2006) at the Apache Point Observatory. During eBOSS, 174,816 LRG redshifts were obtained over 4,242 deg^2 of both northern and southern skies, in the redshift interval $0.6 < z < 1.0$. These were combined with BOSS galaxies in the same redshift range covering 9,493 deg^2 of sky, resulting in a total of 377,458 LRG redshifts (ROSS et al. 2020). As discussed in 2.2, the survey geometry was defined using a set of randomly distributed points, taking into account masked areas. The final data sample has correction weights accounting for the photometric and spectroscopic incompleteness, as well as some spurious correlations (e.g., with stellar density, Galactic extinction). The final data catalogue serves as reference for mock catalogue production.

3.1.2 Mock catalogues

To validate our methodology, we use a sample of 1000 mock realisations of the eBOSS LRG sample. The simulations were produced using the EZMOCK method (CHUANG et al. 2015b), which employs the Zel'dovich approximation to compute the density field at a given redshift. This method is

faster than n-body simulations and has been calibrated to reproduce the two- and three-point statistics of the given galaxy sample, to a good approximation and up to mildly nonlinear scales. The angular and redshift distributions of the eBOSS LRG sample were reproduced, including both photo and spectro-incompleteness effects as well as systematic effects introducing spurious correlations. These mocks aim to include all known features in real data. A detailed description of the EZMOCKS can be found in C. ZHAO et al. 2021. We use these mocks to estimate covariance matrices and to perform statistical studies of our method. As we will describe in section 3.2.4, mocks are also required by the Gaussian approximation method for combining configuration and Fourier space results into a single consensus result.

3.2 Methodology

3.2.1 Measuring the clustering

We used correlation function and power spectrum multipoles estimated for each of the 1000 EZMOCK realisations. As described in section 2.2.3, the correlation function is measured using the LANDY et SZALAY 1993 estimator, while power spectra were calculated using the YAMAMOTO et al. 2006b estimator, as implemented by BIANCHI et al. 2015b; SCOCCIMARRO 2015b. The reconstruction procedure discussed in section 2.1.3.1 is applied to the catalogues to reduce the impact of non-linear evolution and RSD, and sharpen the BAO feature. Our baseline analysis uses the post-reconstructed catalogues, though in appendix 3.5, we show results for the pre-reconstruction case.

3.2.2 BAO modelling

The code that produces the model and perform the fitting to the data is publicly available[1]. We use the model outlined in section 2.1.3.1 for the multipoles of the power spectrum $P_\ell(k)$ and correlation function $\xi_\ell(r)$. We use scales relevant for the measurement of the BAO feature, typically $0 < k < 0.3\ h\ \mathrm{Mpc}^{-1}$ for power-spectrum and $30 < r < 180\ h^{-1}\mathrm{Mpc}$ for correlation function. In Eq 2.38, the smoothing parameter used for the reconstruction is set to $\Sigma_r = 15 h^{-1}\mathrm{Mpc}$. We use ranges of scales $50 < r < 80$ and $160 < r < 190 h^{-1}\mathrm{Mpc}$ to extract the BAO peak component P_{peak} from the full linear power-spectrum P_{lin}, following the procedure from KIRKBY et al. 2013. We tested polynomial and spline interpolation, and got similar results. The linear power spectrum P_{lin} is obtained from CAMB[2] (LEWIS et al. 2000) with cosmological parameters of our fiducial cosmology defined in table 3.1.

We parametrise the BAO peak position with the parameters scaling the separations across the line of sight, $\alpha_\perp$, and along the line of sight, $\alpha_\parallel$ (Eq 2.23). The rescaling of comoving sound horizon at drag epoch r_d is here to account for the choice of template cosmology used to compute the fixed linear power spectrum. For simplicity, the template cosmology is chosen to match the fiducial cosmology used to estimate distances from redshifts. The observed $\boldsymbol{k}$ is then related to the true $\boldsymbol{k}'$ by $k_\parallel = k'_\parallel/\alpha_\parallel$ and $k_\perp = k'/\alpha_\perp$. Therefore, the observed power spectrum relates to the true power spectrum as

$$P(k_\parallel, k_\perp) = \frac{1}{{\alpha_\perp}^2 \alpha_\parallel} P'\left(\frac{k_\parallel}{\alpha_\parallel}, \frac{k_\perp}{\alpha_\perp}\right). \tag{3.1}$$

[1] https://github.com/julianbautista/baopy
[2] camb.info

TABLE 3.1 : Fiducial cosmologies used in this work. All models are parameterized by their fraction of the total energy density in form of total matter $\Omega_{\rm m}$, cold dark matter $\Omega_{\rm cdm}$, baryons $\Omega_{\rm b}$, and neutrinos Ω_ν, the Hubble constant $h = H_0/(100\ {\rm km\ s^{-1}\ Mpc^{-1}})$, the primordial spectral index n_s and primordial amplitude of power spectrum A_s of scalar perturbations. With these parameters we compute the normalisation of the linear power spectrum σ_8 at $z = 0$ and the comoving sound horizon scale at drag epoch $r_{\rm d}$.

	Baseline	EZmocks
$\Omega_{\rm m}$	0.310	0.307
$\Omega_{\rm cdm}$	0.260	0.259
$\Omega_{\rm b}$	0.048	0.048
Ω_ν	0.0014	0.0
h	0.676	0.678
n_s	0.970	0.961
$A_s[10^{-9}]$	2.041	2.116
$\sigma_8(z=0)$	0.800	0.823
$r_{\rm d}$ [Mpc]	147.78	147.66

We remind that this volume scaling does not apply for the correlation function :

$$\xi(r_\parallel, r_\perp) = \xi'\left(\alpha_\parallel r_\parallel, \alpha_\perp r_\perp\right). \tag{3.2}$$

We use the final multipole models of Eqs 2.54 and 2.55 computed at the effective redshift of the survey $z_{\rm eff}$. Since the polynomial functions of scale added to our model are highly correlated with the Fingers-of-God Lorentzian, we fix the parameter Σ_s to zero. In GIL-MARÍN et al. 2021, the BAO analysis in Fourier space used a polynomial order range $(i_{\rm min}, i_{\rm max}) = (-1, 1)$, while in J. E. BAUTISTA et al. 2021 the configuration space analysis used $(i_{\rm min}, i_{\rm max}) = (-2, 0)$. For both, this corresponds to three free parameters per multipole. In this work, our baseline choice is $(i_{\rm min}, i_{\rm max}) = (-2, 1)$ when performing joint fits unless stated otherwise. Our baseline BAO analysis uses the monopoles P_0, ξ_0 and quadrupoles P_2, ξ_2 of the power spectrum and correlation functions (see section 3.3.4 for results using the hexadecapole). We fix $\beta = 0.35$ and fitting b with a flat prior between $b = 1$ and 4. For all fits, the broadband parameters are free, and both dilation parameters are allowed to vary between 0.5 and 1.5.

3.2.3 Parameter inference

The cosmological parameter inference is performed by means of the maximum likelihood analysis. Assuming Gaussianity of the data vector, we use the likelihood detailed in Eq 2.107, with the vector parameters $\boldsymbol{\theta}$ and a n_p data vector $\boldsymbol{d}$ containing the observed multipoles. We estimate the covariance $\hat{C}$ and precision matrix $\hat{\Psi}$ using the 1000 EZMOCKS. The best-fit BAO parameters $(\alpha_\perp, \alpha_\parallel)$ are determined by minimizing the χ^2 using the quasi-Newton minimum finder algorithm IMINUIT[3] which marginalizes over nuisance parameters while sampling for the parameters of

[3] https://iminuit.readthedocs.io/

interest. The uncertainties in $\alpha_\parallel$ and $\alpha_\perp$ are estimated with the MINOS function provided by iMINUIT. This function computes the intervals where χ^2 increases by unity, which corresponds to a 68% confidence interval. Gaussianity on the parameter posteriors is not assumed in this calculation and uncertainties can be asymmetric with respect to the best-fit value. The two-dimensional confidence contours in $(\alpha_\perp, \alpha_\parallel)$, such as those presented in Figure 3.11, are estimated using the CONTOUR function from iMINUIT. Similarly to MINOS, this function scans χ^2 values in two dimensions, looking for the contours yielding a $\Delta\chi^2 = 2.3$ or 5.9 for 68 and 95 per cent confidence levels, respectively.

3.2.4 Consensus via Gaussian approximation

In previous work, configuration and Fourier space results were combined into a single consensus result using the method presented in SÁNCHEZ et al. 2017.

The idea of the method is to generate a consensus result from M different measurement vectors $\boldsymbol{x}_m$, each containing p elements, and their covariance matrices $\mathbf{C}_{mm}$, each of size $p \times p$. For example, we want to combine $M = 2$ measurements of the vector $\boldsymbol{x}_m = [\alpha_\perp, \alpha_\parallel]$ containing $p = 2$ parameters from Fourier ($m = 1$) and configuration space ($m = 2$), each with its own 2×2 error matrix $\mathbf{C}_{mm}$ derived from their posteriors. The consensus is a single vector $\boldsymbol{x}_c$ with covariance $\mathbf{C}_c$, for which the expressions assume that the χ^2 between individual measurements is the same as the one from the consensus result. The expression for the combined covariance matrix is

$$\mathbf{C}_c \equiv \left(\sum_{m=1}^{M} \sum_{n=1}^{M} \mathbf{C}_{mn}^{-1} \right)^{-1}, \tag{3.3}$$

and the combined data vector is

$$\boldsymbol{x}_c = \mathbf{C}_c \sum_{m=1}^{M} \left(\sum_{n=1}^{M} \mathbf{C}_{nm}^{-1} \right) \boldsymbol{x}_m, \tag{3.4}$$

where $\mathbf{C}_{mn}$ is a $p \times p$ block from the full covariance matrix $\mathscr{C}$ between all parameters and methods containing $pM \times pM$ elements, defined as

$$\mathscr{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1M} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{M1} & \mathbf{C}_{M2} & \cdots & \mathbf{C}_{MM} \end{pmatrix}. \tag{3.5}$$

The diagonal blocks $\mathbf{C}_{mm}$ come from each measurement method M, therefore, assuming Gaussian posteriors for the p parameters. The off-diagonal blocks $\mathbf{C}_{mn}$ with $m \neq n$ is difficult to estimate from the data itself. These off-diagonal blocks are commonly built from mock catalogues. Using a set of many realisations, one can build $\mathscr{C}$ from all the realisations of $\boldsymbol{x}_m$ for each method. We obtain the correlation coefficients $\rho^{\text{mocks}}_{p_1,p_2,m,n}$, that is the covariance $\mathscr{C}$ normalized by its diagonal elements, between parameters p_1, p_2 and methods m, n. We scale these coefficients by the diagonal errors from the data, to obtain the final matrix $\mathscr{C}$ for the data.

It is worth emphasizing that the matrix $\mathscr{C}$ for our data, and more specifically its off-diagonal blocks, depend on that particular realisation of the data in principle. However, the ones derived

from mock measurements are ensemble averages. We account for this fact by scaling the correlations coefficients from the mocks in order to match the maximum correlation coefficient that would be possible with the data (ROSS et al. 2015a). For the same parameter p_1 measured by two different methods m and n, we assume that the maximum correlation between them is given by $\rho_{\max} = \sigma_{p1,m}/\sigma_{p1,n}$, where σ_p is the error of parameter p. This number is computed for the data realisation $\rho^{\rm data}_{\rm max}$ and for the ensemble of mocks $\rho^{\rm mocks}_{\rm max}$. We can write the adjusted correlation coefficients for one particular realisation as

$$\rho^{\rm data}_{p_1,p_1,m,n} = \rho^{\rm mocks}_{p_1,p_1,m,n} \frac{\rho^{\rm data}_{\rm max}}{\rho^{\rm mocks}_{\rm max}}. \tag{3.6}$$

The equation above accounts for the diagonal terms of the off-diagonal block C_{mn}. For the off-diagonal terms, we use

$$\rho^{\rm data}_{p_1,p_2,m,n} = \frac{1}{4}\left(\rho^{\rm data}_{p_1,p_1,m,n} + \rho^{\rm data}_{p_2,p_2,m,n}\right)\left(\rho^{\rm data}_{p_1,p_2,m,m} + \rho^{\rm data}_{p_1,p_2,n,n}\right). \tag{3.7}$$

This adjustment of correlation coefficients is an addition to the method proposed by SÁNCHEZ et al. 2017 and was first implemented in J. E. BAUTISTA et al. 2021 to obtain consensus results.

In summary, the method proposed by SÁNCHEZ et al. 2017 can provide consensus results from strongly correlated measurements of the same quantities. However, it assumes Gaussian likelihoods and yields Gaussian posteriors on the final parameters, which is not a good approximation when measurements are noisy. Also, it relies on mock measurements twice : once to derive the covariance matrix of two-point functions, from which we derive each parameter vector $\boldsymbol{x}_m$; and once again to derive the full parameter-method covariance $\mathscr{C}$, using Eqs. 3.6 and 3.7. We call this method Gaussian approximation (GA). In the following section, we study joint fits in Fourier and configuration space, such that mock catalogues are used only once to derive a consensus result.

3.2.5 Joint analysis

We study a new alternative method to get a unique set of constraints on the parameters $\alpha_\parallel$ and $\alpha_\perp$, taking into account the full correlations between the configuration space (CS) and Fourier space (FS) measurements. We refer to this method as Joint Space (JS) analysis. We concatenate the measured ξ_ℓ multipoles with the P_ℓ multipoles such that we obtain a data vector $[\xi_0, \xi_2, P_0, P_2]$ with $(2 \times r_{bins} + 2 \times k_{bins})$ data points. We then fit the data using a unique set of parameters, except for the nuisance parameters defining the polynomial smooth functions (Eqs. 2.54 and 2.55). In order to perform the joint inference, one needs to estimate the full covariance matrix. We use the measured P_ℓ and ξ_ℓ on the 1000 EZmocks to compute the covariance matrix $\hat{\boldsymbol{C}}$. As in J. E. BAUTISTA et al. 2021 and GIL-MARÍN et al. 2021, we apply the following scale cuts : $r \in [50, 150]$ h^{-1} Mpc and $k \in [0.02, 0.3]$ h Mpc^{-1}.

Figure 3.1 shows the resulting correlation matrix, defined as $\boldsymbol{R} = \boldsymbol{\sigma}^{-1}\hat{\boldsymbol{C}}[\boldsymbol{\sigma}^{-1}]^T$,where $\boldsymbol{\sigma}$ is the vector composed by the square-root of the diagonal elements of $\hat{\mathbf{C}}$. The diagonal blocks show the well-known correlations between the monopole and quadrupole in the same space, while the off diagonal blocks reveal the correlation patterns between the two spaces. The correlations between the two spaces monopole-monopole and quadrupole-quadrupole present the same non-linear features. A given physical scale r is correlated (and anti-correlated) with several k modes, meaning that the statistical information contained in a given scale is spread out over several spectral modes. A physical modelling of these features is proposed in Appendix 3.5.

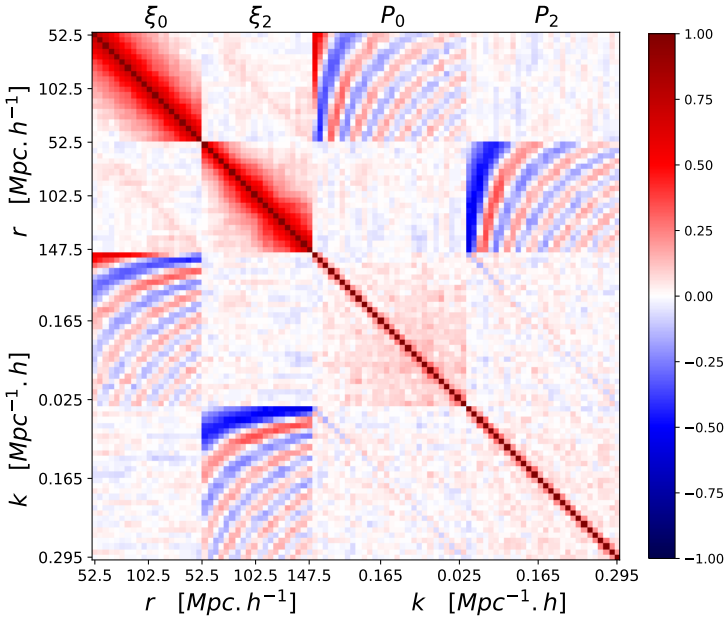


FIGURE 3.1 : Correlation matrix of the correlation function ξ_ℓ and power spectrum multipoles P_ℓ estimated using the 1000 independent measurements of EZmocks.

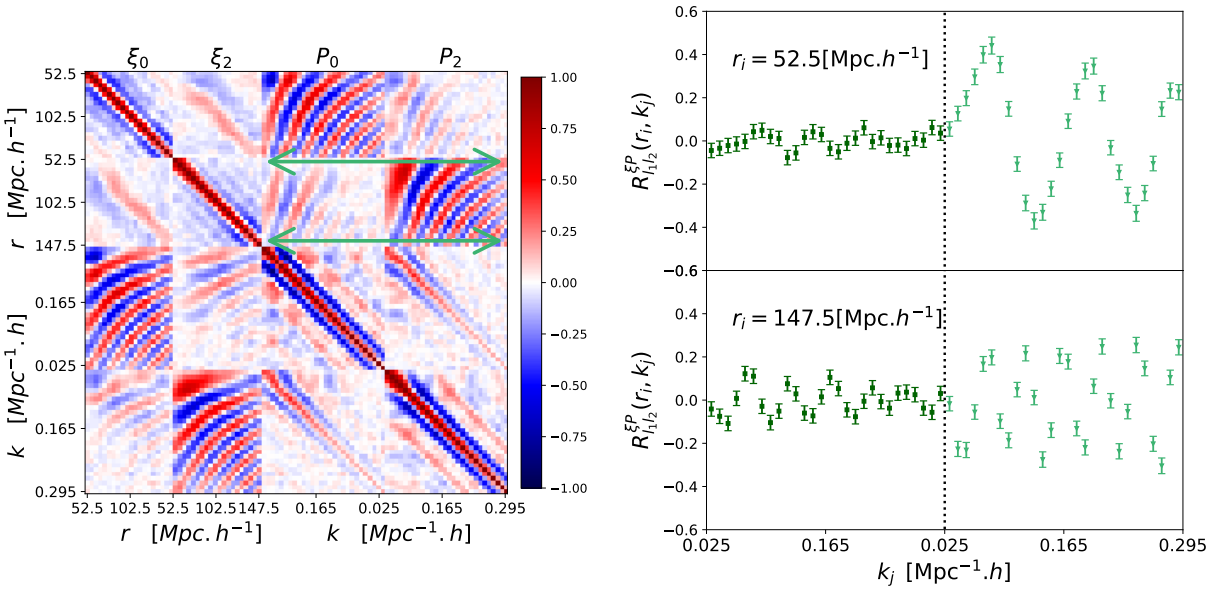


FIGURE 3.2 : Precision matrix of the correlation function ξ_ℓ and power spectrum multipoles P_ℓ.Top panel : normalized estimated precision matrix $R^{\xi P}_{ij} = \Psi_{ij}/\sqrt{\Psi_{ii}\Psi_{jj}}$, obtained from 1000 EZmock realisations. Bottom panel : slices of $R^{\xi P}_{ij}$ showing the cross-correlation between ξ_ℓ and P_ℓ for particular values of scales as shown by the green arrows in the left panel. Errors are estimated from Eq. 29 of TAYLOR et al. 2013.

TABLE 3.2 : Correction factors for the three analysis performed in this work. m_1 is the factor to be applied to the estimated covariance matrix of the parameters and m_2 is the factor that scales the scatter of best-fit parameters of a set of mocks (if these were used in the calculation of the covariance matrix). n_{mocks} is the number of mocks used in the estimation of the covariance matrix, n_p is the total number of parameters fitted and n_d is the total size of the data vector. The derivation of m_1 and m_2 can be found in PERCIVAL et al. 2014.

Analysis	m_1	m_2	n_p	n_d	n_{mocks}
CS	1.018	1.061	11	40	1000
FS	1.035	1.097	11	56	1000
JS	1.062	1.176	19	96	1000

We then need an estimate for the precision matrix $\hat{\mathbf{\Psi}}$. Inverting the covariance mixes the different modes and scales across CS and FS. The top panel of figure 3.2 presents the resulting correlations in the normalized precision matrix (for convenience we call the coefficients of the precision matrix 'correlations'). Green arrows indicate the regions detailed in the bottom panels, where we focus on the correlations between the CS ξ_2 and the FS P_0 and P_2, respectively by plotting the amplitude of the precision matrix at two fixed scaled $r_{\text{min}} = 52.5\ h^{-1}$ Mpc., and $r_{\text{max}} = 147.5\ h^{-1}$ Mpc. The error bars of the precision matrix are computed according to Eq. 29 of TAYLOR et al. 2013. We see that at both scales, ξ_2 is weakly correlated with P_0 while being strongly correlated with P_2. This result is in agreement with the correlations between ξ_0 and ξ_2. Furthermore, for large r, while the number of correlated k modes increases, the amplitude of the correlations decreases distinctly for $\xi_2 - P_2$, and increases for $\xi_2 - P_0$. Note that the correlations between the ξ_0 and P_0 are not shown here but behave similarly.

While capturing additional information about the scale dependence of the joint space correlations, our new methodology has some drawbacks. Indeed, because the covariance matrix $\hat{C}$ we used to infer our set of parameters is an estimate of the real precision matrix drawn from a Wishart distribution (finite sample of mocks), each element is affected by its own uncertainty that should be correctly propagated to the uncertainty of the estimated parameters as explained in section 2.3.3. Note that this is also true for a standard FS or CS analysis but since the correction factors only scale with the number of parameters, data bins and mocks used to estimate the covariance, the enlargement on the constraints is smaller. The factor m_1 is to be directly applied to the estimated parameter covariance matrix for a given measurement, while the factor m_2 scales the standard deviation of a given parameter over a set of mocks. The values of the parameters m_1 and m_2 are given in table 3.2. Because the enlargement of the parameter constraints ($\sqrt{m_1}$) expected in CS is about 1%, 1.7% in FS and 3% in JS, the joint analysis is not guaranteed to give tighter constraints.

3.3 Results on mock catalogues

In this section, we use post-reconstruction EZMOCKS to validate our parameter inference and error estimation (see Appendix 3.5 for a discussion on the analysis of the pre-reconstruction EZMOCKS). The aim is to compare results from configuration space (CS), Fourier space (FS), Joint space (JS) with the consensus results from the Gaussian approximation (GA).

3.3.1 Fits on average correlations

Using the inference methodology described in the section 3.2, we fit the average P_ℓ and ξ_ℓ of the 1000 EZmocks in order to study potential biases and compare the GA and JS methods. At first, we model ξ_ℓ and P_ℓ in the separation ranges $r \in [50, 150]$ h^{-1} Mpc and $k \in [0.02, 0.3]$ h Mpc^{-1}, by setting the broad band smooth polynomial functions from Eqs. 2.54 and 2.55 to the shape $(i^{\min}, i^{\max}) = (-2, 1)$ and letting all fitting parameters free. Then, we assess the robustness of the best fit parameters with respect to variations in $\Sigma_\perp, \Sigma_\parallel$, k ranges, r ranges, number of broadband terms, and template cosmology for the linear power spectrum. Each time we vary one of those settings we keep the other ones fixed in all spaces (CS, FS and JS), facilitating the comparisons of results. Since we fit the mean of the mocks, we normalize the covariance matrix by the total number of mocks $n_{\text{mocks}} = 1000$. As the covariance is estimated with the same 1000 realisations, we do not expect it to be highly accurate, but sufficient for our purposes.

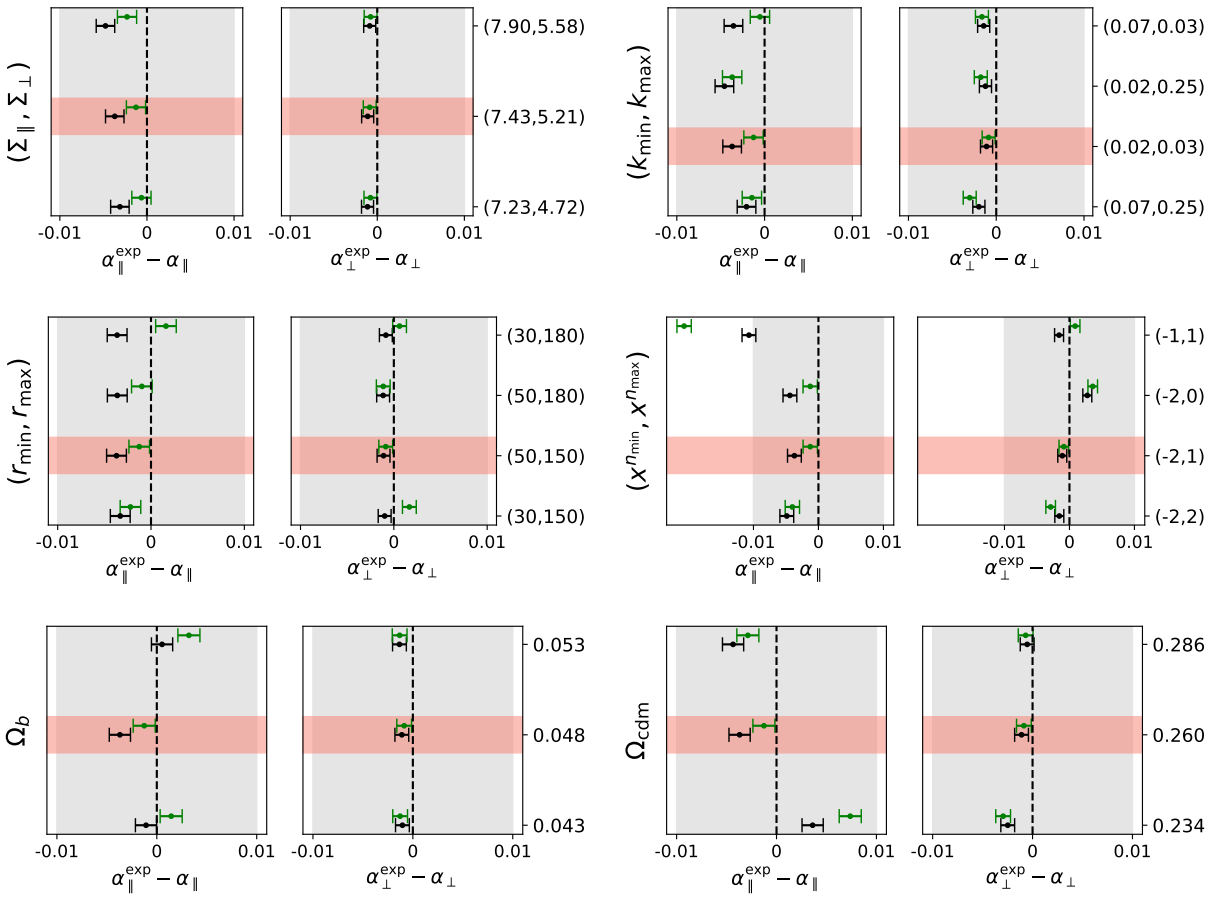


FIGURE 3.3 : Impact of the choice of fitting scales, fiducial cosmology, broadening Σ's parameters and polynomial broadband order in the recovered values of the parameters $\alpha_\parallel$ and $\alpha_\perp$. Each point is the best-fit from the average of the 1000 EZmocks. The JS results are in green and GA in black. The grey shaded areas correspond to a 1 percent error and red shared areas indicate the fiducial choices of our analysis.

Figure 3.3 presents the impact of different analysis settings on the best fit values of $\alpha_\parallel$ and $\alpha_\perp$ for the GA (in black) and JS (in green) analysis. In every plot, a grey shaded area represents the

1 percent deviation from the expected value, which was the tolerance used in previous eBOSS analyses. First, we tested different set of values for the broadening of the BAO peak, $(\Sigma_\parallel, \Sigma_\perp)$, corresponding to the best values found in CS (7.90, 5.58), FS (7.23, 4.72) and JS (7.43, 5.21), all in units of h^{-1}Mpc. Both GA and JS analyses are robust - presenting small and coherent shifts - to changes in those parameters, and for different ranges of scales used in the fit. This indicates that the information is correctly extracted only from the BAO feature. When varying the broadband terms, the JS method shows stable results except for $\alpha_\parallel$ in one setting (-1, 1), where the broadband is not flexible enough to fit the residuals. When increasing the number of polynomial terms, the systematic shifts are notably smaller in JS than GA. The last two panels of Figure 3.3 show the influence of the template cosmology used to compute the linear power spectrum (not the one to convert redshifts into distances). We chose to vary separately the two parameters $\Omega_{\rm b}$ and $\Omega_{\rm cdm}$. Each time, we compute the new expected values for the α parameters by rescaling $r_d^{\rm fid}$ (while the fiducial separations are unchanged as the cosmology used to compute distances does not vary). We find that the systematic shift behave in the same way for the two methods, and do not exceed 1% when varying $\Omega_{\rm cdm}$ or $\Omega_{\rm b}$ by 10% from the baseline cosmology. Note however that we performed the fits while letting free the parameters $\Sigma_\parallel$ and $\Sigma_\perp$, which helps reducing the shift amplitude between the different cosmologies.

Overall, the inferred value for the parameter $\alpha_\perp$ is the more robust than $\alpha_\parallel$ for both methods. Globally, GA and JS behave in the same way when varying the fitting configuration, proving that these are not caused by the new methodology. Best values and uncertainties are consistent between GA and JS. The systematic shifts for $\alpha_\parallel$ are smaller in JS, while no significant difference is seen for $\alpha_\perp$. The chosen baseline for the rest of the analysis is highlighted with a red band in figure 3.3. The systematic errors may partially result from the covariance matrix that would require more realisations, but as they are much smaller than statistical errors budget for the real data ($\sigma_\alpha/\alpha \sim 2\%$), we can safely neglect these.

3.3.2 Varying the number of mocks

We test the robustness of the different analysis against the number of mocks used. We fit the averaged stack of the 1000 mocks using different values of $n_{\rm mocks}$ to estimate the sample covariance matrix.

Figure 3.4 shows the figure of merit defined as Fom $= 1/\sqrt{|C_\theta|}$ with C_θ the covariance matrix of the inferred parameters, and the quadratic shift from the expected parameters $\sqrt{\Delta^2_{\alpha_\parallel} + \Delta^2_{\alpha_\perp}}$ for different $n_{\rm mocks}$. We find that, due to the Whishart bound, the JS precision highly decreases when the number of mocks get small, typically the Fom drops when $n_{\rm mocks} < 4n_d$. However the JS constraint is compatible with the GA at 1σ, and appears to be robust as the Fom does not significantly decrease in the range 1000-600 $n_{\rm mocks}$. While the GA analysis introduces an additional dependence on $n_{\rm mocks}$ through the covariance C_c defined in Eq 3.3 (which has to be calculated for each point using the set of best fit parameters found with the appropriated sample covariance), the Fom is only decreased by a few percent when using 100 realisations. We find that for the JS the quadratic shift from the expected parameters monotonically decreases as the amount of mocks increases. For a large enough amount of realisations, the JS gives the smaller systematic bias.

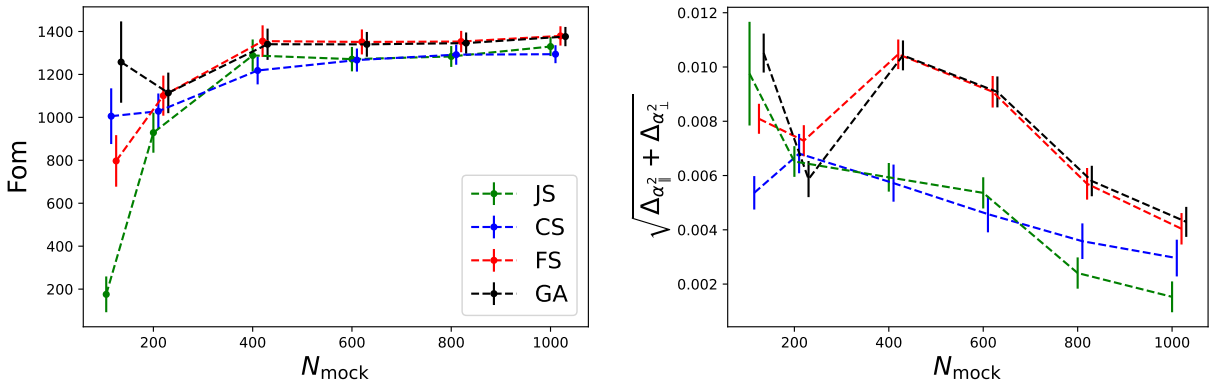


FIGURE 3.4 : Figure of merit (Fom) and quadratic shift from the expected parameters resulting from the fit of the stack of EZmocks for different n_{mocks} used to estimated the sample covariance. Errors for the α's are scaled with $\sqrt{n_{\mathrm{mocks}}}$ and errors for the Fom are derived from Eq 68 of TAYLOR et al. 2013.

3.3.3 Fits on individual mocks

We now focus on the statistical properties of best-fit values and their uncertainties, based on fits of individual realisations of EZmocks. When fitting individual mocks, we fix $(\Sigma_\parallel, \Sigma_\perp)$ to the best-fit values obtained from the joint analysis of the stack of 1000 measurements (see previous section), while letting all other parameters free. Table 3.3 summarises the parameters and flat priors used in these fits. The constraints on $(\alpha_\parallel, \alpha_\perp)$ are then obtained by marginalizing over b and all other nuisance parameters.

TABLE 3.3 : Parameters used in the fits, with their flat priors. See Eq. 2.38 and text for detailed description of the parameters.

Parameter	Prior	Is fixed?
$\alpha_\perp$	[0.5, 1.5]	No
$\alpha_\parallel$	[0.5, 1.5]	No
b	[1., 4.]	No
β	-	0.35
$\Sigma_\parallel$	-	7.428
$\Sigma_\perp$	-	5.210
$a^P_{\ell,i}$	-	No
$a^\xi_{\ell,i}$	-	No

For each method (FS, CS and JS) we remove results from non-converging likelihoods (unsuccessful contour estimation) and extreme best-fit values at 5σ level (σ is defined as half of the range covered by the central 68 per cent values). The best-fit values for which the estimated uncertainties touch the prior boundaries are removed as well. The remaining number of realisations referred as N_{good} is given for each method in table 3.4 and is consistent with the numbers obtained in the official eBOSS analyses (GIL-MARÍN et al. 2021 ; J. E. BAUTISTA et al. 2021).

To obtain GA results, we extract the 1 sigma contour of each mock posterior profile. We fit the likelihood contours corresponding to 68 per cent C.L. with an ellipse, which we translate into a parameter covariance matrix C_{mm}, where m refers to either CS or FS. We scale the resulting covariance with the parameter m_1 (see table 3.2). We then construct the total covariance matrix $\mathscr{C}$ from Eq. 3.5 obtained from the 1000 best-fit $(\alpha_\perp, \alpha_\parallel)$, adjusting each time the coefficients (according to Eqs. 3.6 and 3.7) to account for the observed errors of a given realisation. The corresponding correlation matrix before the individual adjustments is shown in Figure 3.5. We see strong correlations between the parameters in the two spaces. Using the combination method,

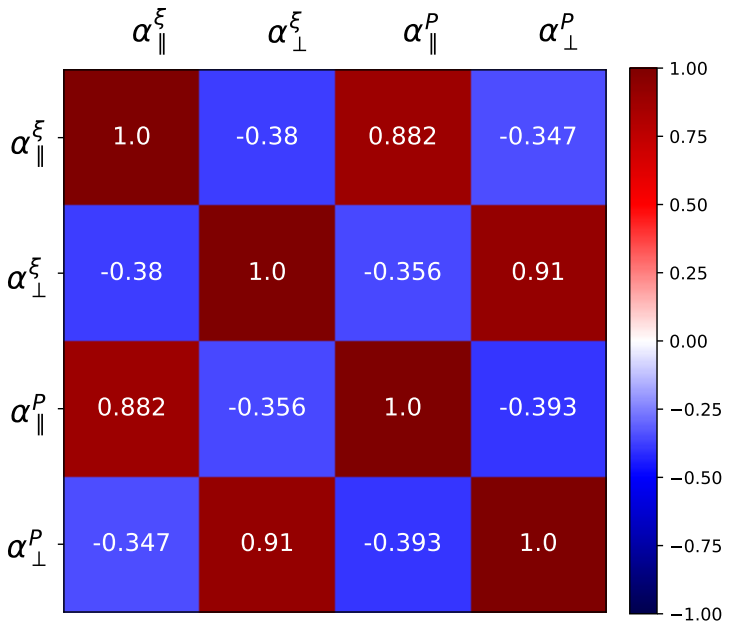


FIGURE 3.5 : Correlation coefficients between $\alpha_{\parallel}$ and $\alpha_{\perp}$ in CS and FS obtained from fits to the 1000 EZmock realisation.

we compute for each mock the consensus data vector $\boldsymbol{x}_c$ and covariance C_c from Eqs. 3.3 and 3.4. Here we emphasize that the strong assumption of Gaussian elliptic contours for the parameters $(\alpha_{\parallel}, \alpha_{\perp})$ of every inference highly depends on the statistic of the considered tracer. The presence of non-Gaussian contours in our sample of mocks introduces biases in the total covariance matrix $\mathscr{C}$ and each individual consensus data vector $\boldsymbol{x}_c$ and covariance $\mathbf{C}_c$. This is one of the limitations of the GA method that we can avoid with a JS fit.

Figure 3.6 compares the distributions of $\alpha_{\parallel}$ and $\alpha_{\perp}$ and their 1σ uncertainties as estimated by GA (top four panels) and JS (bottom four panels) versus the same quantities obtained from CS and FS fits. We see that the distributions are nicely correlated and scattered around the identity line (dashed). We find that the scatter (especially for errors) is less important between FS and JS than CS and JS. This would indicate that the Fourier space information has slightly more weight than the configuration space in the minimization of the likelihood. Moreover, while the GA errors are almost systematically inferior to the CS and FS ones, the JS errors are scattered on both sides of the dashed lines. Due to statistical fluctuations, JS analysis may result in looser constraints than FS or CS. We believe these fluctuations might origin from the finite number of mock realisations used to build the covariance matrix, though it is hard to test this hypothesis without a larger number of mocks.

Table 3.4 summarises the statistical properties of $(\alpha_{\parallel}, \alpha_{\perp})$ for the different analyses, CS, FS, GA and JS, performed on 1000 EZmock realisations. For each parameter, we show six quantities : the average bias $\Delta_{\alpha} = \langle \alpha - \alpha_{\,\mathrm{exp}} \rangle$ with respect to the expected value ($\alpha_{\perp\,,\,\mathrm{exp}} = \alpha_{\parallel\,,\,\mathrm{exp}} = 1$), the standard deviation of best-fit values σ, the mean estimated error $\langle \sigma_i \rangle$, the mean asymmetry of the estimated error distribution $\langle A(\sigma_i) \rangle = \langle 2\left(\sigma_{\mathrm{i}}^{\mathrm{sup}} - \sigma_{\mathrm{i}}^{\mathrm{inf}}\right) / \left(\sigma_{\mathrm{i}}^{\mathrm{sup}} + \sigma_{\mathrm{i}}^{\mathrm{inf}}\right) \rangle$, the mean of the pull $Z_i = (\alpha_i - \langle \alpha_i \rangle) / \sigma_i$ and its standard deviation. If errors are correctly estimated and follow a

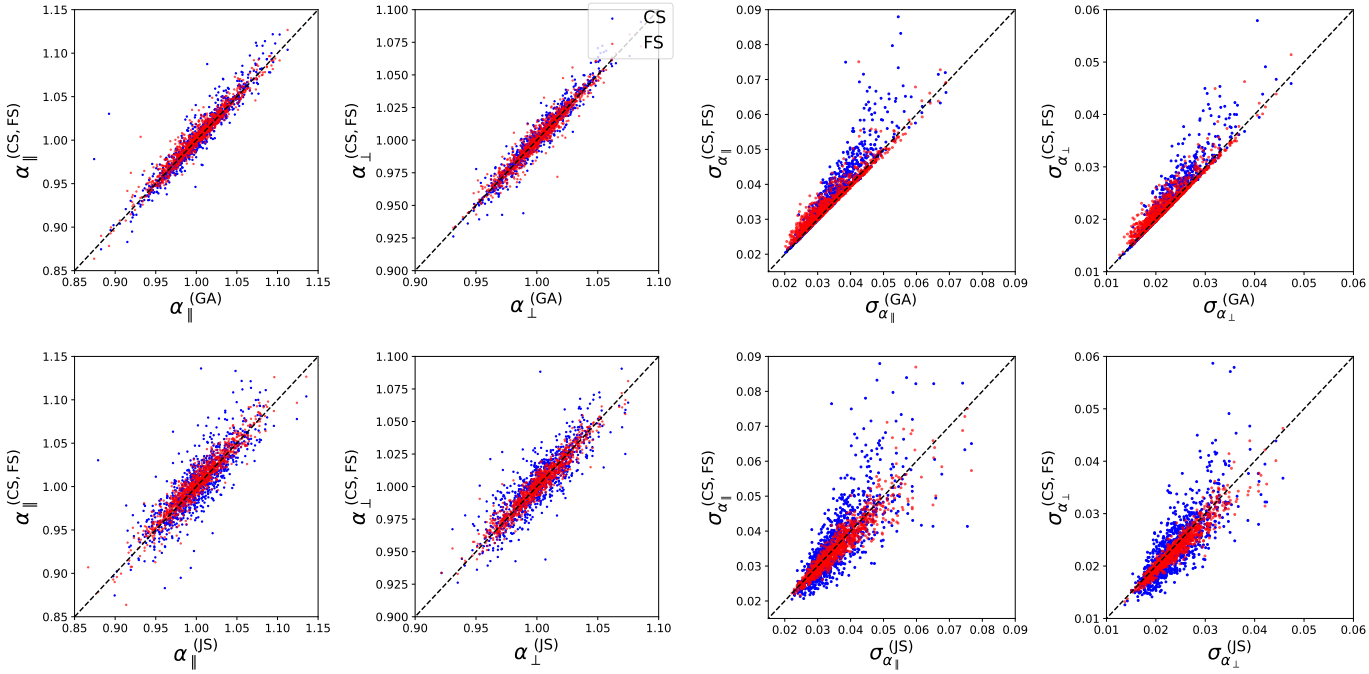


FIGURE 3.6 : Comparison between the distributions of the best-fit BAO parameters $(\alpha_\parallel, \alpha_\perp)$ and their estimated errors obtained by fitting individual mock realisations. The four top (bottom) panels compare the CS (blue), FS (red) distributions with the GA (JS). The dashed lines represent a perfect correlation.

TABLE 3.4 : Statistics on the fit of the 1000 EZmocks realisations. $N_{\rm good}$ is the number of valid realisations after removing undefined contours and extreme values and errors. We show the mean value of the best-fit reduced ${}_r\chi^2_{\rm min}$. For each parameter, we show the average bias $\Delta_\alpha \equiv \langle \alpha_i - \alpha_{\rm exp} \rangle$, the standard deviation of best-fit values $\sigma \equiv \sqrt{\langle \alpha_i^2 \rangle - \langle \alpha_i \rangle^2}$, the average of the per-mock estimated uncertainties $\langle \sigma_i \rangle$, the asymmetry of the estimated error distribution $A(\sigma_i) \equiv \langle 2\left(\sigma_{\rm i}^{\rm sup} - \sigma_{\rm i}^{\rm inf}\right) / \left(\sigma_{\rm i}^{\rm sup} + \sigma_{\rm i}^{\rm inf}\right) \rangle$, where σ_i^{sup} and $\sigma_i^{\rm inf}$ are the superior and inferior one sigma errors estimated from the likelihood profile, the average of the pull $Z_i \equiv (\alpha_i - \langle \alpha_i \rangle)/\sigma_i$ and its standard deviation $\sigma(Z_i)$.

			$\alpha_\perp$						$\alpha_\parallel$					
	$N_{\rm good}$	$\langle {}_r\chi^2_{\rm min} \rangle$	Δ_α [10^{-2}]	σ [10^{-2}]	$\langle \sigma_i \rangle$ [10^{-2}]	$\langle A(\sigma_i) \rangle$ [10^{-2}]	$\langle Z_i \rangle$	$\sigma(Z_i)$	Δ_α [10^{-2}]	σ [10^{-2}]	$\langle \sigma_i \rangle$ [10^{-2}]	$\langle A(\sigma_i) \rangle$ [10^{-2}]	$\langle Z_i \rangle$	$\sigma(Z_i)$
CS	985	0.96	-0.18	2.38	2.35	2.46	-0.03	1.00	-0.26	3.71	3.65	4.63	-0.03	0.96
FS	998	0.99	-0.13	2.29	2.30	1.95	-0.02	0.95	-0.40	3.58	3.54	0.51	0.00	0.93
GA	983	-	-0.12	2.19	2.24	-	-0.03	0.98	-0.19	3.36	3.40	-	-0.01	0.96
JS	994	0.95	-0.12	2.44	2.39	2.40	-0.02	0.93	-0.12	3.64	3.61	0.10	0.00	0.91

Gaussian distribution, we expect that $\sigma = \langle\sigma_i\rangle$, $\langle Z\rangle = 0$ and $\sigma(Z_i) = 1$. Table 3.4 also shows the number $N_{\rm good}$ of valid realisations after removing undetermined likelihoods, extreme values and errors, along with the mean value of the reduced chi-square ${}_r\chi^2_{\rm min}$. Firstly, one can notice that the number of valid realisations (as defined above) for the JS analysis is larger than the one of the combined analysis GA. The GA method requires both FS and CS fits to converge, so the number of valid realisations for GA is necessarily the intersection of valid FS and CS realisations. By joining the information of both spaces, the JS fit is able to constrain the acoustic scale even on noisy mocks, where FS and CS separately fail. For each fitting procedure, we find the minimum reduced chi-square $\langle {}_r\chi^2_{\rm min}\rangle \sim 1$, showing that the majority of the mocks are accurately modeled by our templates. Secondly, we see good agreement between $\langle\sigma_i\rangle$ and σ for both parameters in all analyses. While the JS errors lies in between the CS and the FS errors for $\alpha_\parallel$, the GA results present in average smaller errors. Note that the standard deviation σ of the best-fit values are scaled with the appropriate correction factor $\sqrt{m_2}$ (see table 3.2). We remind that for the GA method we cannot apply any m_2 factor.

Overall, systematic shifts Δ_α for all methods are below half a per cent. The GA and JS methods yield reduced systematic shifts Δ_α than each individual space alone. While the shifts between GA and JS are the same for $\alpha_\perp$, the JS method yields slightly smaller shift for $\alpha_\parallel$ as also observed in the previous section. If we assume that σ is a good estimate of the standard deviation of α, then the uncertainty of Δ_α should be $\sigma_\Delta = \sigma/\sqrt{N_{\rm good}}$. In this case, all systematic shifts are smaller than $3\sigma_\Delta$ except for $\alpha_\parallel$ in FS, which reaches a $3.5\sigma_\Delta$ discrepancy. Shifts for GA are (-1.7, -1.8)σ_Δ for $(\alpha_\perp, \alpha_\parallel)$ and (-1.5, -1.0)σ_Δ for JS fits. Note that our results are slightly different from those reported in J. E. Bautista et al. 2021 and Gil-Marín et al. 2021 due to a few differences in the analyses, such as how we decompose peak and smooth parts of the template, the number of polynomial terms, the values of non-linear damping terms.

Figure 3.7 displays the distributions of the pull Z_i for $\alpha_\parallel$ and $\alpha_\perp$. For all four methods, the mean of the pull $\langle Z_i\rangle$ is centred in zero to a few percent level, suggesting again that there are no systematic bias in the estimated α. Moreover, the standard deviation of the pull $\sigma(Z_i)$ is smaller than unity for both parameters in all analyses indicating a slight overestimation of the errors. This overestimation is more important for the parameter $\alpha_\parallel$, and in JS in general. As the standard deviation of the pulls are only a few percent away from unity, we do not attempt to correct these effects. If any overestimation of uncertainties is real, not correcting for it can be considered as a conservative approach. Yet, we see that for the parameter $\alpha_\parallel$ the overestimation of the errors reaches 10% in JS. While this seems to be significant, this result can be a consequence of the Gaussian assumption for the distribution of the parameters that might not be necessarily true, as we investigate in the following.

To investigate the Gaussian nature of the distribution of the α's parameters, we perform a D'Agostino-Pearson's test (R. B. D'Agostino 1971; R. D'Agostino et Pearson 1973) on the sample of best-fit parameters using the Scipy.stats library[4]. The test combines the high-order statistical moments skewness and kurtosis to test the hypothesis that the two parameters are normally distributed. The statistical quantity K_2 is constructed as a combination of the renormalized skewness and kurtosis such that K_2 asymptotically follows a χ^2 law with two degrees of freedom. When the skewness and kurtosis simultaneously deviate from 0 (Fisher kurtosis) K_2 gets larger. The p-value for the K_2 statistic can then be computed. Table 3.5 shows the results of the Pearson tests on the $\alpha_\perp$, and $\alpha_\parallel$ distributions. Independently of the method, the Gaussian distribution hypothesis is very much unlikely for the parameter $\alpha_\parallel$ as the p-values are

[4] https://docs.scipy.org/doc/scipy/reference/stats.html

TABLE 3.5 : Results of D'Agostino and Pearson's test of normality over the α's distributions. $K_2 = z_k^2 + z_s^2$ is a combination of the skewness and kurtosis normalized coefficients. The p value of the test is a 2-sided chi squared probability for the Gaussianity hypothesis test.

	$\alpha_\perp$		$\alpha_\parallel$	
Analysis	K_2	$p-$value	K_2	$p-$value
CS	5.55	0.06	17.89	1×10^{-4}
FS	2.92	0.23	16.92	2×10^{-4}
GA	3.42	0.18	11.14	3.8×10^{-3}
JS	2.64	0.27	11.91	2.6×10^{-3}

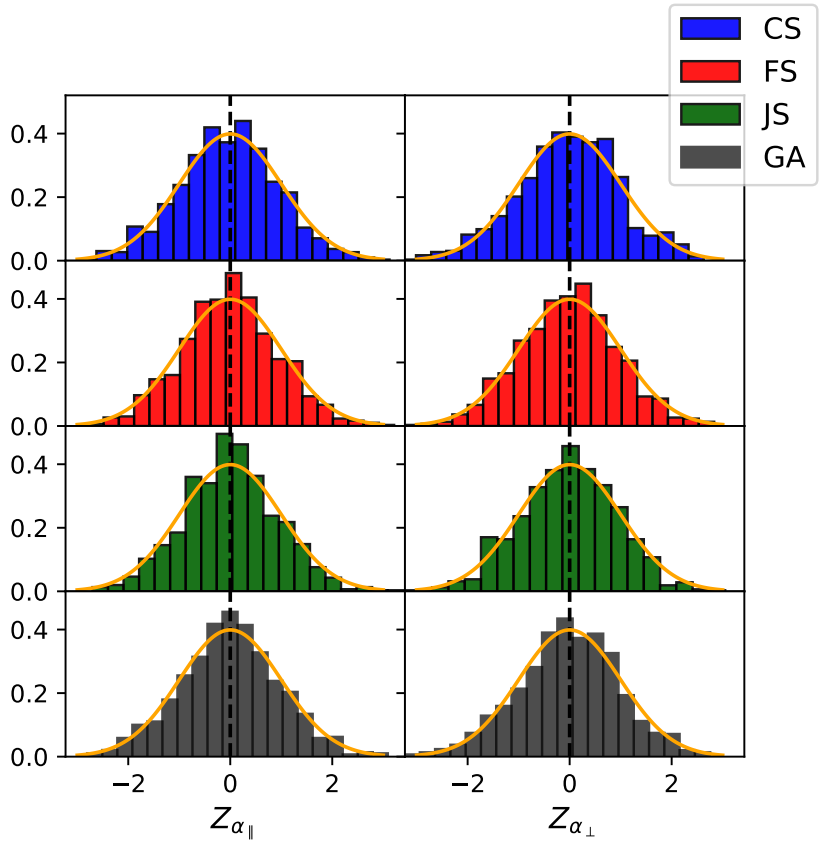


FIGURE 3.7 : Distributions of the pull $Z_i = (x_i - \langle x_i \rangle)/\sigma_i$ for the parameters $(\alpha_\parallel, \alpha_\perp)$ obtained obtained from the CS, FS, GA and JS analyses of the 1000 EZmocks sample. Orange curves describing a Gaussian distribution with zero mean and the standard deviation of the corresponding pull are shown for comparison. The dotted lines are centered on zero.

extremely small. This result indicates that for $\alpha_\parallel$, the smaller values of $\sigma(Z_i)$ observed in Table 3.4 might be partially induced by the non-Gaussianity of the α distributions and not simply by an overestimation of the uncertainties.

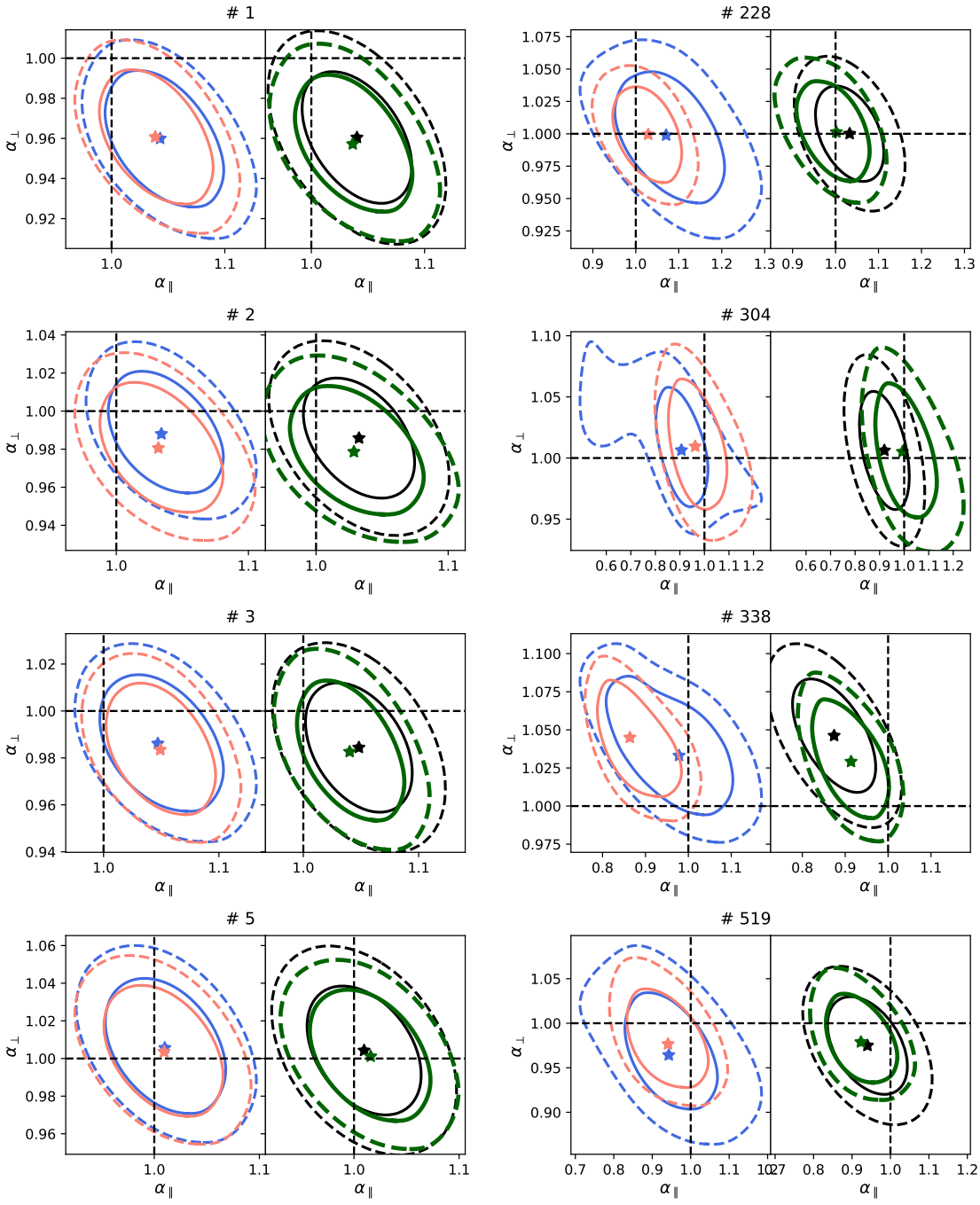


FIGURE 3.8 : Comparison between the different methods, using some EZmock realisations (mock tag above each plot), when the contours found in configuration or Fourier space are Gaussian(left) and non Gaussian(right). Red contours are for FS results, blue for CS, green for JS and black for GA. The expected value is the intersection of the dotted black lines, and the best fit values are described by a star for the JS and GA methods.

Figure 3.8 shows two-dimensional confidence levels for a few realisations of mocks. Red contours are for FS results, blue for CS, green for JS and black for GA. The left panels are cases where the likelihoods can be approximated by a 2D Gaussian distribution while the right panels the opposite. In these examples, we can see how the JS fits are a better description for the non-Gaussian cases, where the final consensus are not necessarily Gaussian.

3.3.4 Including the hexadecapole

Table 3.9 shows the results for the fit of the averaged 1000 EZmocks, with and without using the hexadecapole. For every analysis the use of the additional information results in a slight increase of the systematic bias and of the errors bars. The hexadecapole has more noise than BAO signal (this is true mostly when the fiducial cosmology is the same as the true one from the mocks). The looser constraints also arise from the larger size of the data vector when using hexadecapole. Indeed, a larger amount of data bins results in an increase of the Wishart bound through the rescaling parameter m_1 (especially for the JS analysis). Therefore, we decide not to use the hexadecapole information in this work.

FIGURE 3.9 : Results for the fit of the 1000 EZmocks stack with and without using the hexadecapole information. For each parameter, we show the bias $\Delta_\alpha \equiv (\alpha_i - \alpha_{\rm exp})$ and the estimated uncertainties σ_α.

	$\ell = (0,2)$				$\ell = (0,2,4)$			
	$\alpha_\perp[10^{-2}]$		$\alpha_\parallel[10^{-2}]$		$\alpha_\perp[10^{-2}]$		$\alpha_\parallel[10^{-2}]$	
	Δ_α	σ_α	Δ_α	σ_α	Δ_α	σ_α	Δ_α	σ_α
CS	0.17	2.28	0.24	3.5	0.18	2.3	0.25	3.54
FS	0.09	2.25	0.39	3.42	0.11	2.27	0.39	3.44
GA	0.13	2.23	0.32	3.35	0.14	2.24	0.33	3.36
JS	0.09	2.34	0.13	3.49	0.13	2.4	0.13	3.6

3.4 Application to eBOSS DR16 data

In this section we apply our methodology to the eBOSS LRG sample, described in section 3.1.1, letting free only $\alpha_\parallel$, $\alpha_\perp$, the linear density bias b and the broadband parameters.

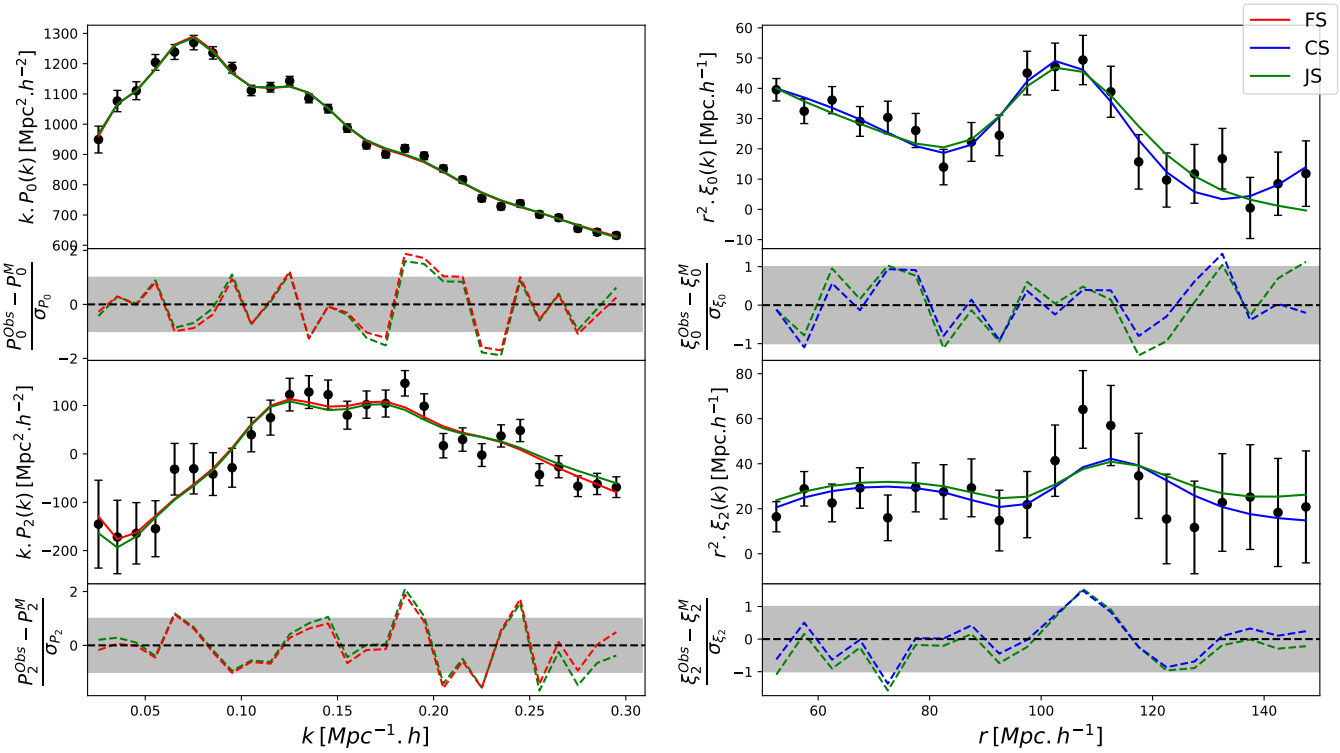


FIGURE 3.10 : Best monopole and quadrupole models for the CS, FS and JS analysis. The residues with respect to the measurements are standardized for each point by the errorbar. The grey shaded area correspond to a 1σ difference.

Figure 3.10 shows the best-fit models for the power spectrum and correlation function multipoles, for three analyses : FS, CS and JS (note that we do not show a GA best-fit model as GA results are derived using the CS and FS best-fit alphas). The residuals in the bottom panels show agreement between the JS model and the individual models for FS and CS, particularly over the BAO features. Furthermore, as the global amplitudes driven by the parameter b are sensibly the same, the small differences between the models arise mainly from the different broadbands. This is specially true for the correlation function as it is composed of fewer data bins.

Figure 3.11 shows the two-dimensional 68 and 95 percent confidence levels for CS, FS, GA and JS. Once again we see an excellent agreement between the four analyses, with the JS giving slightly looser constraints. We also notice small differences in the inclination of the contours. The correlation coefficients for the CS, FS, GA and JS are respectively (-0.395, -0.404, -0.396, -0.445). Those values are consistent with the correlations found with 1000 EZmock realisations : (-0.380, -0.395, -0.383, -0.420). Note that this coefficient for the GA is highly dependent on the quality of the estimated correlation matrix of the parameters estimated from the EZmock sample and shown in Figure 3.5. We find a higher correlation coefficient using the JS analysis, resulting in a thinner and steeper contours. This effect increases errors on the individual parameters without varying much the Figure-of-Merit (see section 3.3.2).

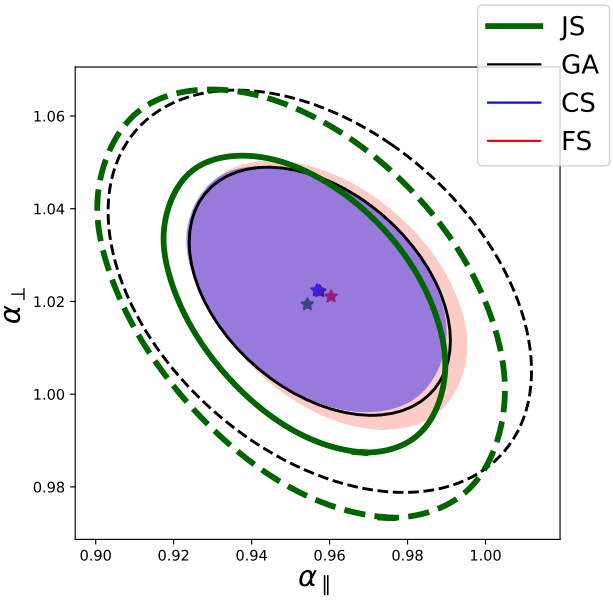


FIGURE 3.11 : Confidence contours of the parameters posterior distributions for the four analysis. For the CS and JS, the 1σ contours are filled in blue and red respectively. For the JS and GA we give the 1σ and 2σ contours for comparison in green and black. The best values are designated with a star.

Table 3.6 summarises the results of the four different analysis. The reduced χ^2 values are 1.28, 1.27 and 1.21 respectively for FS, CS and JS indicating relatively good adjustment to the data without over fitting from the broadband. In each case, the p_{value} indicates a valid fit (if compared with a risk of 5%). The best fit values for the parameters $(\alpha_\perp, \alpha_\parallel)$ are all in agreement according to their respective 1σ error bar. For both parameters, the JS analysis gives the larger uncertainty. Rescaling with the fiducial cosmology, we also derive the corresponding physical quantities D_M/r_d and D_H/r_d. Note that our best fit results do not exactly match the BAO constraints obtained in J. E. BAUTISTA et al. 2021 and GIL-MARÍN et al. 2021, neither in configuration space nor in Fourier space, where the GA consensus gives $D_H/r_d = 19.33 \pm 0.53$ and $D_M/r_d = 17.86 \pm 0.33$ (third entry of Table 14 of J. E. BAUTISTA et al. 2021). Those discrepancies arise from the slightly different non-linear broadening parameters $(\Sigma_\parallel, \Sigma_\perp)$ and a more flexible broadband polynomial expansion $(i_{\text{min}}, i_{\text{max}}) = (-2, 1)$ to be compared to $(-2, 0)$ for J. E. BAUTISTA et al. 2021 and $(-1, 1)$ for GIL-MARÍN et al. 2021.

TABLE 3.6 : Best-fit BAO parameters from the eBOSS DR16 LRG sample for different methods : configuration-space (CS), Fourier-space (FS), the Gaussian combination (GC) and our new joint-space fit (JS). The α values are relative to our fiducial cosmology for which $D_M/r_d(z_{\rm eff}=0.7)=17.436$ and $D_H/r_d(z_{\rm eff}=0.7)=20.194$.

Analysis	$\alpha_\perp$	$\alpha_\parallel$	$\chi^2_{\rm min}/n_{\rm dof}=_r\chi^2_{\rm min}$	$p_{\rm value}$	D_H/r_d	D_M/r_d
CS	1.022 ± 0.018	0.957 ± 0.022	$36.8/(40-11)=1.27$	0.151	19.32 ± 0.44	17.83 ± 0.30
FS	1.021 ± 0.019	0.960 ± 0.024	$57.6/(56-11)=1.28$	0.098	19.39 ± 0.47	17.80 ± 0.33
GA	1.022 ± 0.018	0.957 ± 0.022	-	-	19.33 ± 0.45	17.82 ± 0.31
JS	1.019 ± 0.022	0.954 ± 0.024	$97.7/(96-19)=1.21$	0.056	19.27 ± 0.48	17.77 ± 0.37

3.5 Conclusion

This work introduced a new BAO analysis of the DR16 eBOSS LRG sample, using Fourier and configuration-space information simultaneously. We compared the joint space (JS) analysis with the commonly used Gaussian approximation (GA) method, which combines results from the two spaces (Fourier and configuration) at the parameter level. The main advantage of the JS method is that it does not require any Gaussian assumption for the likelihood profiles. While the GA method is accurate only if the individual likelihoods to be combined are both Gaussian, yielding only Gaussian posterior distributions.

We assessed the systematic biases and errors of both methods by applying JS and GA to a set of 1000 EZMocks, which reproduce the eBOSS DR16 LRG sample properties. Compared to GA, JS provides a more accurate estimation of the acoustic scale by lowering the systematic shift of $\alpha_\parallel$ with respect to its expected value, while GA has by construction a better precision (smaller error bars). Moreover, the JS offers a better control over variations of the analysis. Indeed, the same mocks as the ones used to estimate the covariance are commonly used to perform systematic studies. Because of this, the standard deviation of any parameter should be rescaled with the correction factor m_2 (see Table 3.2), which is not properly defined for the GA. One should note that because of larger size of the data vector, the JS methods requires a sufficiently large mock sample to estimate the covariance. However, we found that the constraints are quite stable, and the number of mocks needed is not much larger than for a regular configuration or Fourier analysis.

We analysed the eBOSS LRG sample using different methods and found consistent results (see Table 3.6). As expected from the statistical study, the JS gives slightly looser constraints and a larger correlation coefficient. Despite providing looser constraints on cosmological parameters than the standard GA, we believe that JS is a more robust and reliable method for modelling the clustering signal by correctly accounting for the correlation between configuration and Fourier space and not relying on the Gaussianity of the parameter likelihoods.

One specific feature of the JS in a BAO analysis is that the broadband terms are independent between spaces. Hence, the JS introduces more nuisance parameters to be fitted simultaneously.

This work can also be extended to joint fits between pre- and post-reconstruction catalogues, as already performed in Fourier space by GIL-MARÍN 2022.

Appendix B : pre-reconstruction results

We perform fits on average correlations and individual mocks for the pre-reconstruction sample. Figure 3.12 presents the impact of different parameters choice on the best fit values of $\alpha_\parallel$ and $\alpha_\perp$ for the GA (in black) and JS (in green) analysis. In every plot, a grey shaded area represents the 1 percent deviation from the expected value. Both analysis appears to be more dependent on the range of scales used (particularly in Fourier space) as the small scales non linearities are not well modelled. However, without the reconstruction procedure that requires a fiducial cosmology, the dependence on Ω_b and $\Omega_{\rm cdm}$ is reduced. The two analysis gives different results for the parameter $\alpha_\parallel$, the JS analysis tends to give a positive bias while the GA gives a negative.

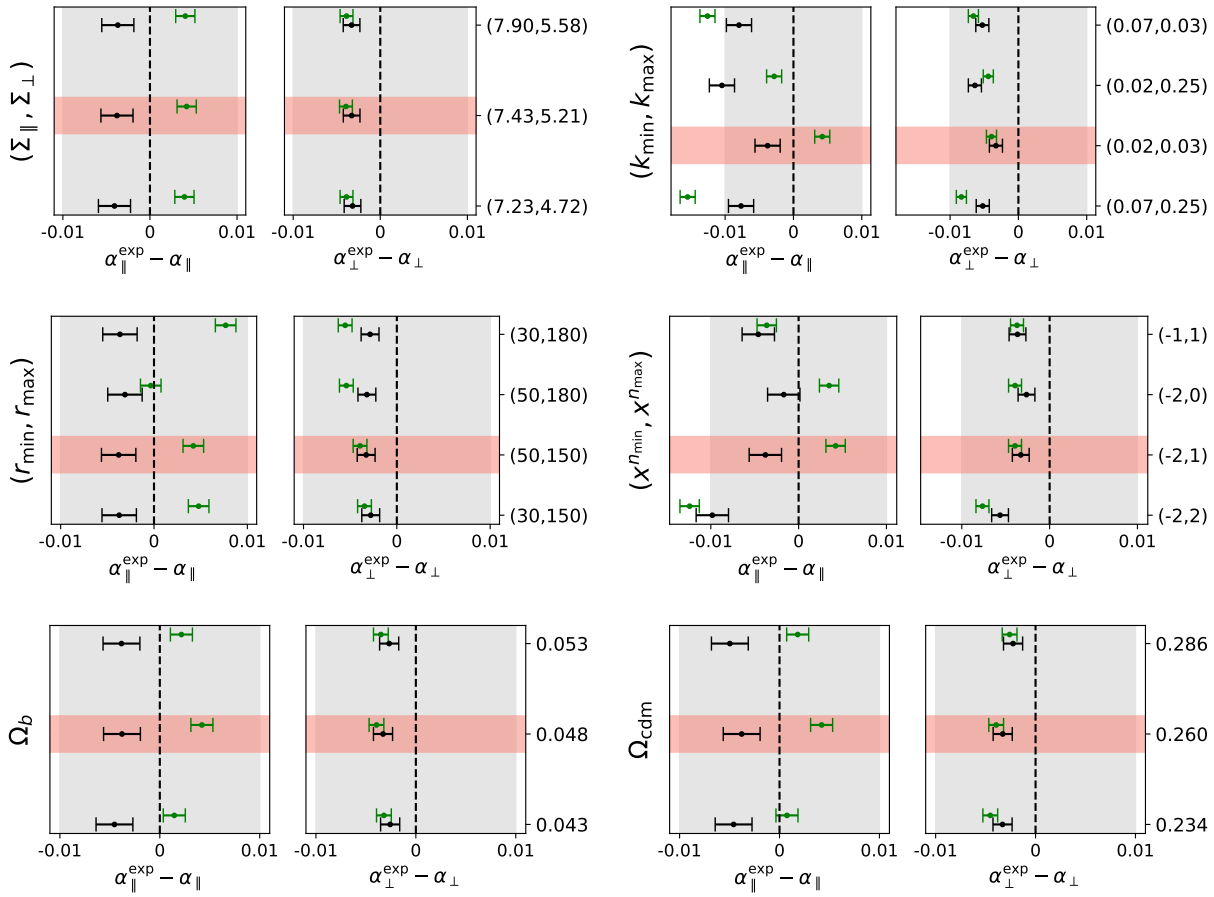


FIGURE 3.12 : Impact of the choice of fitting scales, fiducial cosmology, broadening Σ's parameters and polynomial broadband order in the recovered values of the parameters $\alpha_\parallel$ and $\alpha_\perp$. Each point is the best-fit from the average of the 1000 pre-reconstruction EZmocks. The JS method is in green and GA method is in black. The grey shaded areas correspond to a 1 percent error and red shared areas indicate the fiducial choices of our analysis.

Figure 3.13 shows the correlation matrix between the pre-reconstructed TPCF and PS. The main differences between pre-reconstruction and post-reconstruction covariance matrices are in the off-diagonal blocks within each space (either FS or CS), which are significantly reduced in post-reconstruction. We construct the α's parameters pre-reconstruction total covariance

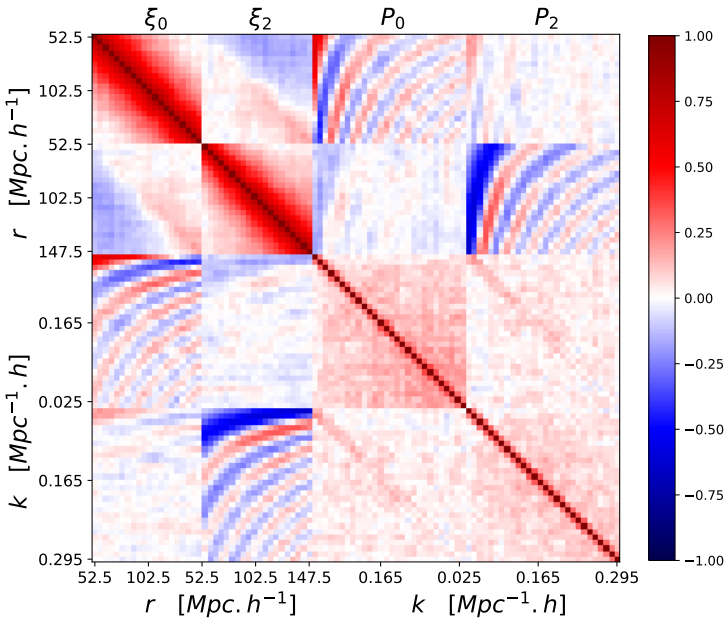


FIGURE 3.13 : Correlation matrix of the correlation function ξ_ℓ and power spectrum multipoles P_ℓ estimated using the 1000 independent measurements of pre-reconstructed EZmocks.

matrix obtained from the 1000 best-fit $(\alpha_\perp, \alpha_\parallel)$. The corresponding correlation matrix before the individual adjustments is shown in Figure 3.14. The reconstruction procedure enhances the correlations for a given parameter in two spaces, but reduces the correlations between the two parameters themselves (both in a given space and across spaces). Overall, the measurements depend on the signal-to-noise ratio of the BAO features, which are lower for pre-reconstruction data.

Table 3.7 summarises the statistical properties of $(\alpha_\parallel, \alpha_\perp)$ for the different analyses, CS, FS, GA and JS, performed on 1000 EZmock realisations. For both parameters, the JS analysis presents larger systematic shifts Δ_α and larger errors than the GA. For the four analysis, $\sigma(Z_i)$ is larger than one, indicating an underestimation of the errors. This effect is more important for the GA. As our modelling is not well suited to describe the non linearities of the pre-reconstruction two point statistics, for every analysis the precision and accuracy are degraded. Here the JS method takes into account the correlations between two observables that are not well modeled in the first place. However, using an appropriate full shape modelling and removing the broadbands nuisance parameters, the CS and FS should be described by the exact same set of parameters. Hence the JS could improve the analysis.

Appendix C : modelisation of the cross correlation pattern

We propose an analytical description of the cross correlations between the correlation function and power spectrum multipoles. The 2-point statistic between the fluctuation field in configura-

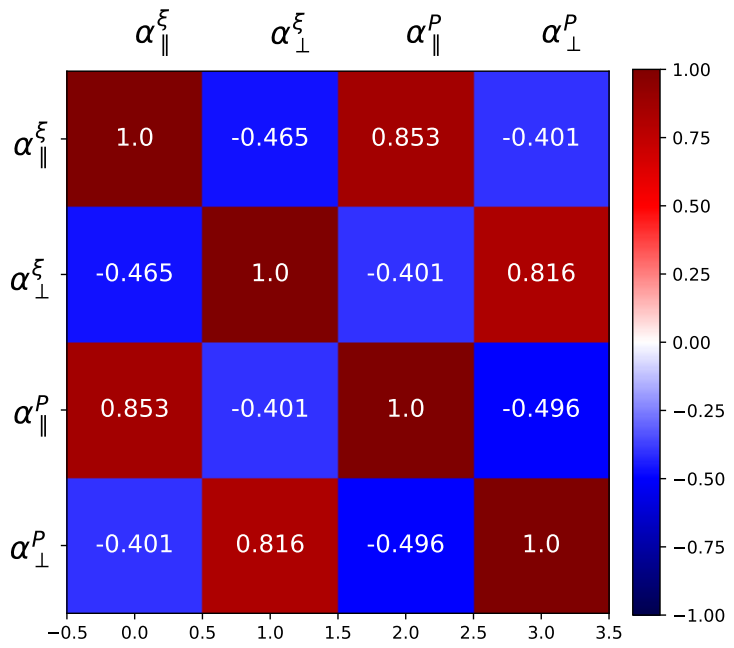


FIGURE 3.14 : Correlation coefficients between $\alpha_\parallel$ and $\alpha_\perp$ in CS and FS obtained from fits to the pre-reconstructed 1000 EZmock realisation.

TABLE 3.7 : Statistics on the fit of the 1000 EZmocks pre-reconstruction realisations. $N_{\rm good}$ is the number of valid realisations after removing undefined contours and extreme values and errors. We show the mean value of the best-fit reduced ${}_r\chi^2_{\rm min}$. For each parameter, we show the average bias $\Delta_\alpha \equiv \langle \alpha_i - \alpha_{\rm exp} \rangle$, the standard deviation of best-fit values $\sigma \equiv \sqrt{\langle \alpha_i^2 \rangle - \langle \alpha_i \rangle^2}$, the average of the per-mock estimated uncertainties $\langle \sigma_i \rangle$, the asymmetry of the estimated error distribution $A(\sigma_i) \equiv \langle 2\left(\sigma_{\rm i}^{\rm sup} - \sigma_{\rm i}^{\rm inf}\right) / \left(\sigma_{\rm i}^{\rm sup} + \sigma_{\rm i}^{\rm inf}\right) \rangle$, where σ_i^{sup} and $\sigma_i^{\rm inf}$ are the superior and inferior one sigma errors estimated from the likelihood profile, the average of the pull $Z_i \equiv (\alpha_i - \langle \alpha_i \rangle)/\sigma_i$ and its standard deviation $\sigma(Z_i)$.

			$\alpha_\perp$						$\alpha_\parallel$					
	$N_{\rm good}$	$\langle {}_r\chi^2_{\rm min} \rangle$	Δ_α [10^{-2}]	σ [10^{-2}]	$\langle \sigma_i \rangle$ [10^{-2}]	$\langle A(\sigma_i) \rangle$ [10^{-2}]	$\langle Z_i \rangle$	$\sigma(Z_i)$	Δ_α [10^{-2}]	σ [10^{-2}]	$\langle \sigma_i \rangle$ [10^{-2}]	$\langle A(\sigma_i) \rangle$ [10^{-2}]	$\langle Z_i \rangle$	$\sigma(Z_i)$
CS	891	0.96	-0.68	3.24	3.08	5.58	-0.05	1.05	-0.66	5.79	5.06	4.24	-0.02	1.11
FS	945	1.01	-0.43	3.26	2.82	3.79	-0.05	1.10	0.11	6.15	4.78	0.65	0.01	1.19
GA	876	-	-0.37	2.95	2.67	-	-0.04	1.12	-0.20	5.26	4.37	-	-0.01	1.21
JS	931	0.96	-0.48	3.37	2.90	5.06	-0.05	1.08	0.58	6.53	4.87	0.21	0.01	1.19

tion and Fourier space is

$$\langle \delta_g(\boldsymbol{k})\delta_g(\boldsymbol{r})\rangle = \frac{1}{(2\pi)^3}\int d^3k' e^{i\boldsymbol{k}'\cdot\boldsymbol{r}}\langle \delta_g(\boldsymbol{k})\delta_g(\boldsymbol{k}')\rangle = e^{i\boldsymbol{k}\cdot\boldsymbol{r}}P(\boldsymbol{k}), \tag{3.8}$$

while assuming Gaussian fields, the cross covariance is given by

$$\begin{aligned}\mathrm{Cov}\,[P(\boldsymbol{k}),\xi(\boldsymbol{r})] &= \langle \delta_g(\boldsymbol{k})\delta_g(-\boldsymbol{k})\delta_g(\boldsymbol{r})\delta_g(-\boldsymbol{r})\rangle \\ &= 2\cos(2\boldsymbol{k}\cdot\boldsymbol{r})|P(\boldsymbol{k})|^2.\end{aligned} \tag{3.9}$$

Then, including the shot noise contribution P_N and the volume rescaling, the cross covariance between the multipoles $P_l(k)$ and $\xi_{l'}(r)$ can be written

$$\begin{aligned}\mathrm{Cov}\,[P_l(k),\xi_{l'}(r)] = \frac{2l+1}{2}\frac{2l'+1}{2}\frac{1}{Vs}\int_{-1}^{1} d\mu_k L_l(\mu_k)|P(k,\mu_k)+P_N|^2 \\ \times\int_{-1}^{1} d\mu_r L_l(\mu_r) 2\cos\left(2rk\mu_k\mu_r + 2kr\sqrt{1-\mu_k^2}\sqrt{1-\mu_r^2}\right).\end{aligned} \tag{3.10}$$

Figure 3.15 shows the modelling of the cross covariance features using Eq 3.10 for a given range of scales. We find this description satisfying as we only aim for a physical understanding of these patterns, and use the full mocks sample to estimate the covariance. A more reliable modelling would require bin averaged integral as performed in Grieb et al. 2016.

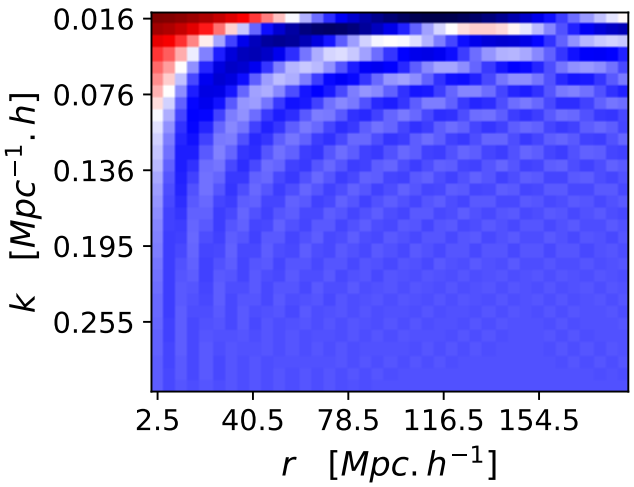


Figure 3.15 : Analytical modelling of the cross correlations between the correlation function and the power spectrum monopoles for a given range of scales. We do not display the colorbar as we set the shot noise to zero and the volume of the survey to one.

4 Simulation based inference for galaxy clustering

Summary

4.1 Emulating clustering from N-body simulations . . . 112
4.1.1 The AbacusSummit Suite of simulations . . . 112
4.1.2 Halo occupation distribution model . . . 113
4.1.3 Observable of interest : The galaxy correlation function . . . 116
4.2 Multiscale Gaussian process model . . . 118
4.2.1 Standard GP regression model . . . 118
4.2.2 Multiscale GP model . . . 124
4.2.3 Application to galaxy clustering . . . 126
4.3 Emulator performance . . . 127
4.3.1 Prediction accuracy . . . 127
4.3.2 Robustness to cosmic variance . . . 128
4.4 Cosmology recovery and constraints . . . 131
4.4.1 Parameter inference using the emulator model . . . 131
4.4.2 Cubic boxes : average fit . . . 133
4.4.3 Cubic boxes : individual fits . . . 134
4.4.4 Observational effects : wide angle and Alcock-Paczynski . . . 137
4.5 Conclusion . . . 142
Appendix A : Emulator covariance choice . . . 144
Appendix B : AP effect on IGP fits . . . 146

The standard analysis of spectroscopic galaxy surveys consists in the measurement and modelling of two main physical features present in the two-point function of the density field : the baryon acoustic oscillations (BAOs) and the redshift space distortions (RSDs). While we have seen that BAO can be well modelled by linear theory, because they manifest as correlations at large linear scales (around 100 h^{-1}Mpc), RSD are more challenging because there is valuable information on small, mildly non-linear scales (below a few tens of h^{-1}Mpc), which are therefore harder to model accurately. The observation of tens of millions of galaxies by ongoing and future surveys (DESI, D. COLLABORATION et al. 2016; DESI COLLABORATION et al. 2022, and Euclid, R. LAUREIJS et al. 2011; E. COLLABORATION et al. 2022), spanning deeper and wider fields of view will significantly reduce statistical uncertainties in the measured statistics. With shrinking errorbars, the cut-off scale of perturbation theories increases, and higher precision models are required to derive unbiased cosmological constraints from the analyses of the non-linear scales of two-point function. As discussed in section 2.1.5, simulation based models provide the highest fidelity description of the matter field dynamics, allowing to extract additional cosmological information from non-linear scales, typically up to $1h^{-1}$Mpc. A recent approach, made possible by increasing computing capabilities, consists in running several N-body simulations for different cosmological models, computing the observable of interest for each, and using them to perform prediction for a new set of cosmological parameters. This process is commonly referred to as emulation. In recent years, several suites of N-body simulations - spanning different cosmologies - have been produced for this purpose (DEROSE et al. 2019; VILLAESCUSA-NAVARRO et al. 2020; MAKSIMOVA et al. 2021). Several emulators have been developed (KOBAYASHI et al. 2020) and some of these have been applied to real data to extract cosmological parameters (KOBAYASHI et al. 2022; CHAPMAN et al. 2022).

Because estimating the expectation value by running many realisations of a simulation is computationally expensive, notably when spanning several cosmologies, any summary statistic estimated from one realisation of the matter field suffers from cosmic variance (section 2.3.3). While this variance can be reduced by increasing the volume of the simulation (especially for small scales), or by using more sophisticated methods (ANGULO et PONTZEN 2016b; HADZHIYSKA et al. 2023a), ultimately every observable measured from a simulation comes with its error budget. Additional noise, coming from numerical specificities of the simulation suite (resolution, time discretisation, gravity solver, etc), contributes to the total error budget and propagates to the predictions of any simulation based model. An emulator should therefore consider the uncertainty of the training set along with the uncertainty of the predictions.

A popular emulation algorithm used in previous works for galaxy clustering (YUAN et al. 2022a; ZHAI et al. 2022; KWAN et al. 2023) is the Gaussian process (GP). Compared to neural networks, GPs have physical hyperparameters and provide a probabilistic distribution around a mean prediction. However, it is well known that GP models scale poorly with the number of training points N as they require the construction and inversion of an $N \times N$ matrix (N is the number of training samples) many times during the training phase. Each of these operations costs $O(N^2)$ memory and $O(N^3)$ time. As a consequence, the dimension of the input space is often restricted to the cosmology and galaxy-halo connection parameters. In particular, the standard approach employed in previous works is to construct an independent emulator for each separation of the correlation function, or for each Fourier mode of the power spectrum, ignoring the well-known correlations between scales.

In this chapter, we propose a new Gaussian process model (DUMERCHAT et J. BAUTISTA 2024), allowing to extend the input parameter space, and therefore to interpolate over a wider range

of parameters, such as scales, redshift bins, wide angles, selection effects, and so on. As a proof of concept, we built an emulator for the galaxy two-point correlation function from a redshift $z = 0.2$ snapshot of the AbacusSummit suite (Maksimova et al. 2021), and extended the input parameter space to include the different separation scales. We implemented our new framework to run on Graphics Processing Units (GPUs) with the Jax library.

This chapter is organised as follows. In section 4.1, we present the AbacusSummit N-body simulations, the galaxy–halo connection, and the observable to emulate. In section 4.2, we describe the GP framework along with our new model. In section 4.3, we validate our emulator performance by comparison to a standard model. Finally, in section 4.4, we explore the constraining power of our model.

4.1 Emulating clustering from N-body simulations

In this section, we describe the suite of N-body simulations used in this work. As these are dark-matter-only, we use a halo occupation distribution (HOD) to populate galaxies. We then discuss the summary statistics we use to evaluate the performance of our emulation technique.

4.1.1 The AbacusSummit Suite of simulations

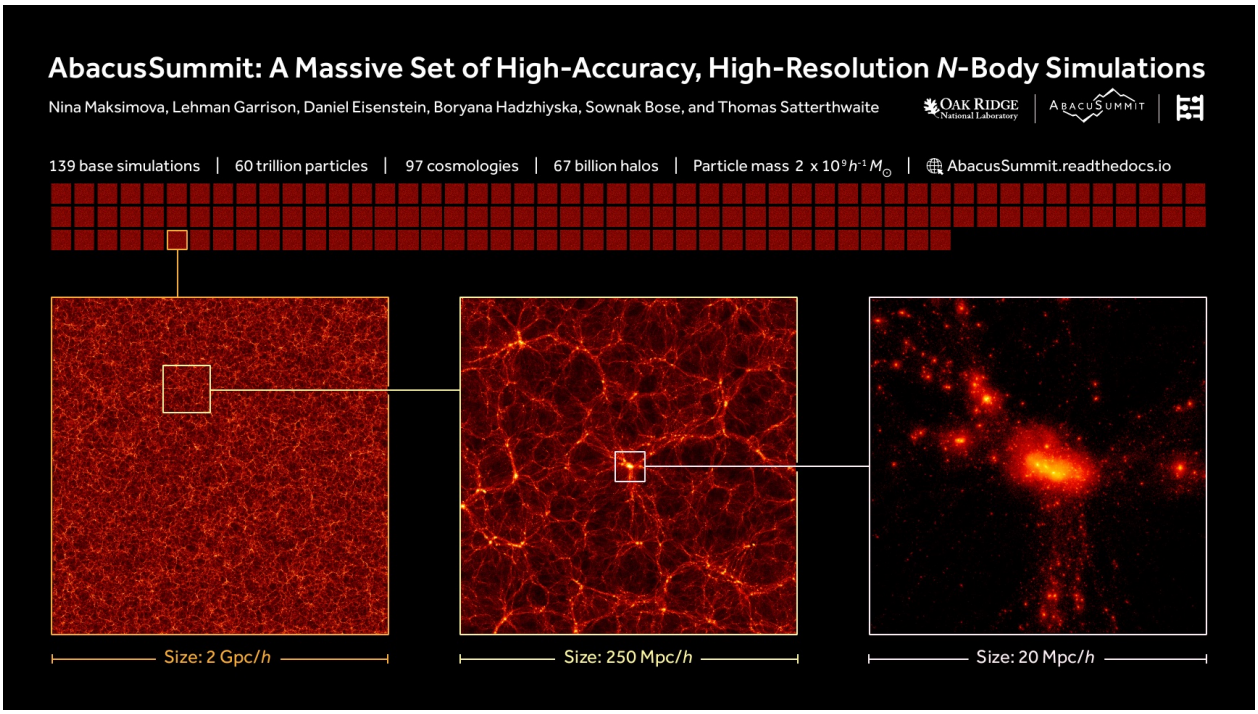


Figure 4.1 : AbacusSummit descriptions of the different data-product run with the base resolution : 6912^3 particles of mass $2 \times 10^9\ h^{-1}M_\odot$ evolved in a $(2\ h^{-1}\mathrm{Gpc})^3$ cubic box. Figure taken from abacussummit.io.

AbacusSummit (Maksimova et al. 2021) is a suite of N-body simulations run with the Abacus N-body code (Garrison et al. 2021) on the Summit supercomputer of the Oak Ridge Leadership Computing Facility. These simulations were designed to match the precision and volume requirements set by the Dark Energy Spectroscopic Survey (DESI). The AbacusSummit suite is depicted

in figure 4.1. It is composed of hundreds of periodic cubic boxes evolved in comoving coordinates to different epochs from redshift $z = 8.0$ to $z = 0.1$. The base simulations consist of 6912^3 particles of mass $2 \times 10^9\ h^{-1}M_\odot$ in a $(2\ h^{-1}\mathrm{Gpc})^3$ volume. The suite provides halo catalogues formed by identifying groups of particles, built using the specialised spherical-overdensity-based halo finder CompaSO (see Hadzhiyska et al. 2021 for more details). AbacusSummit spans several cosmological models sampled around the best-fit ΛCDM model to Planck 2018 data (P. Collaboration et al. 2020) with nine different parameters $\theta_\Omega = \{\omega_{\rm cdm}, \omega_{\rm b}, \sigma_8, w_0, w_a, h, n_{\rm s}, N_{\rm ur}, \alpha_{\rm s}\}$: the physical dark matter density $\omega_{\rm cdm} = \Omega_{\rm cdm,0}h^2$, the physical baryon density $\omega_{\rm b} = \Omega_{b,0}h^2$, the clustering amplitude at redshift zero σ_8, the Dark Energy equation of state $w(a) = w_0 + w_a(1-a)$, the Hubble constant h, the spectral tilt $n_{\rm s}$, the running of the spectral tilt $\alpha_{\rm s}$, and the number of ultra-relativistic species $N_{\rm ur} = N_{\rm eff} - 1$ (the neutrino mass is fixed at 60meV). A flat curvature $\Omega_k = 0$ is assumed for every cosmology. Figure 4.2 shows the sampling over the nine-dimensional space of 88 cosmologies (in blue) used to train the emulator, and 6 cosmologies (in orange) used as a test set, all with the base resolution. The test set includes the centre of the hypercube : the Planck2018 ΛCDM cosmology with a massive neutrino, the same cosmology with massless neutrinos, four secondary cosmologies with low $\omega_{\rm cdm}$ choice based on WMAP7, a wCDM choice, a high $N_{\rm ur}$ choice, and a low σ_8 choice. The training set provides a wider coverage of the parameter space. In the following, we use X_Ω to refer to the 88×9 matrix containing the cosmological parameters of the training set.

All simulations were run with the same initial conditions (ICs), and so changes in the measured clustering are only due to changes in cosmology. To test the cosmic variance, AbacusSummit also provides another set of 25 realisations with the same cosmology and different initial conditions and an additional set of 1400 smaller $(500\ h^{-1}\mathrm{Mpc})^3$ boxes with the same cosmology and mass resolution, but different ICs. The small boxes are used to compute the covariance matrices of our data vectors. All those simulations were run with the base Planck2018 cosmology. While snapshots are available for several redshifts, we only use snapshots at redshift $z = 0.2$ throughout this work[1].

4.1.2 Halo occupation distribution model

We use the HOD framework introduced in section 2.1.5 (Zheng et al. 2007) to assign galaxies to the dark matter halos of the simulations. This framework assumes that the number of galaxies N in a given dark matter halo follows a probabilistic distribution $P(N|\Omega_h)$, where Ω_h can be a set of halo properties. In this work, we assume a vanilla HOD, where such distribution only depends on the halo's mass $\Omega_h = M$. Moreover, the occupation is decomposed into the contribution of central and satellite galaxies $\langle N(M)\rangle = \langle N_{\rm cen}(M)\rangle + \langle N_{\rm sat}(M)\rangle$. The number of central galaxy occupation follows a Bernoulli distribution, while the satellite galaxy occupation follows Poisson distribution. The mean of these distributions are defined as :

$$\langle N_{\rm cen}\rangle = \frac{1}{2}\mathrm{erfc}\left(\frac{\log_{10}(M_{\rm cut}/M)}{\sqrt{2}\sigma}\right), \tag{4.1}$$

$$\langle N_{\rm sat}\rangle = \left[\frac{M - \kappa M_{\rm cut}}{M_1}\right]^\alpha \langle N_{\rm cen}\rangle, \tag{4.2}$$

where $\mathrm{erfc}(x)$ is the complementary error function, and $\theta_{\rm hod} = \{M_{\rm cut}, \sigma, \kappa, M_1, \alpha\}$ the HOD

[1] While a $z = 0.1$ snapshot exists for AbacusSummit, the smaller boxes for covariance matrix estimations were not available.

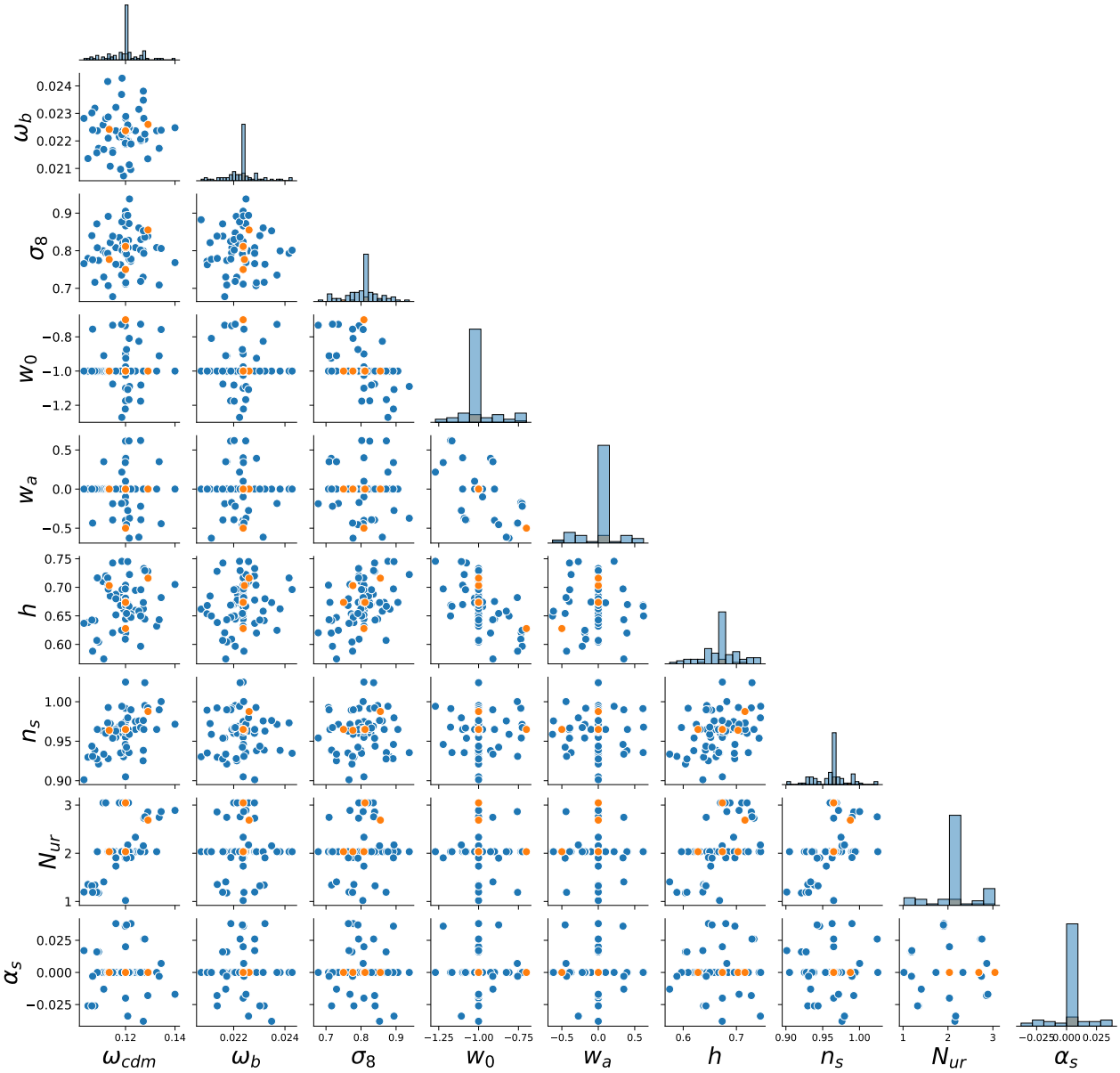


FIGURE 4.2 : Sampling over the nine-dimensional cosmological space around Planck2018 ΛCDM. The 88 cosmologies used to train the emulator are shown in blue, while the 6 cosmologies used as a test set are shown in orange. The dark matter density is denoted $\omega_{\rm cdm}$, the baryon density $\omega_{\rm b}$, the clustering amplitude σ_8, the Dark Energy equation of state $w(a) = w_0 + w_a(1-a)$, the Hubble constant h, the spectral tilt $n_{\rm s}$, the density of massless relics $N_{\rm ur}$, and the running of the spectral tilt α_s (MAKSIMOVA et al. 2021).

parameters. $M_{\rm cut}$ is the minimum mass to host a central galaxy, σ is the width of the cutoff profile (i.e the steepness of the transition 0 to 1 in the number of central galaxies), $\kappa M_{\rm cut}$ is the minimum mass to host a satellite galaxy, M_1 the typical mass to host one satellite galaxy, and α is the power-law index of the satellite occupation.

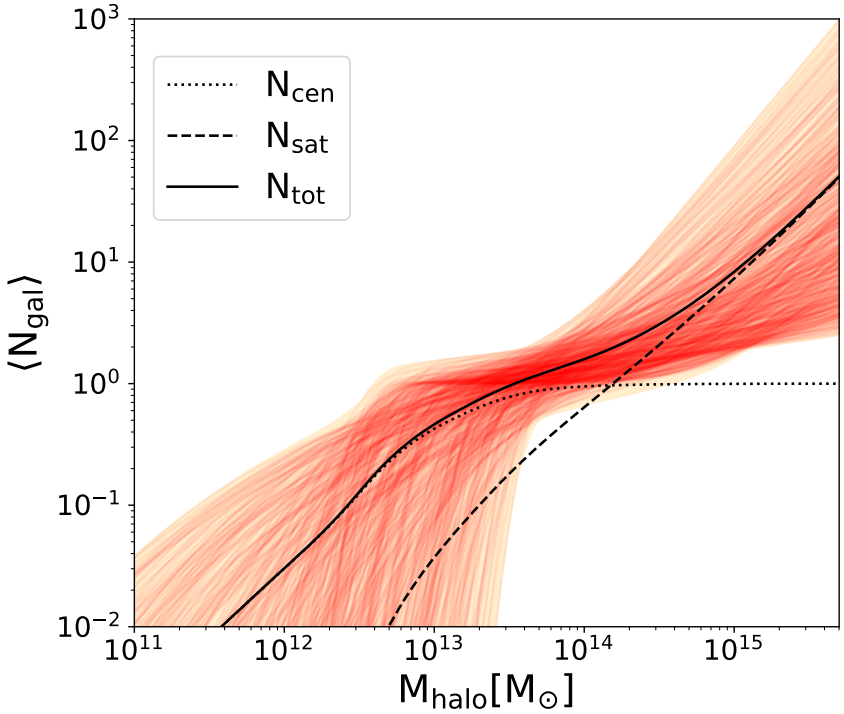


FIGURE 4.3 : Different forms that take the total occupation of the galaxy for the 600 HOD parametrisation used. The black lines show the mean HOD configuration for $\langle N_{\rm cen}(M) \rangle$, $\langle N_{\rm sat}(M) \rangle$, and $\langle N_{\rm tot}(M) \rangle$.

We used the ABACUSHOD implementation (YUAN et al. 2022b), where for each halo, if a random number - drawn from a uniform distribution between 0 and 1 - is smaller than $\langle N_{\rm cen} \rangle$, a central galaxy is assigned to the centre of mass of the halo with the velocity of the halo. Satellite galaxies do not assume a fixed distribution profile such as the Navarro-Frenk-White profile (NAVARRO et al. 1997), but rather follow the distribution of the dark matter particles. Specifically, each particle inside a halo has the same probability of hosting a satellite galaxy $p = \langle N_{\rm sat} \rangle / N_p$ with N_p being the number of particles in a given halo. For every particle, a random number between 0 and 1 is drawn; if it is smaller than p, a galaxy is assigned to its position with the same velocity. The ABACUSUTILS package implements this halo occupation scheme and is publicly available[2].

To build our training set, we populated each of the 88 boxes with 600 different HODs, whose parameter values were sampled from a Latin hypercube (LHS), resulting in 52800 different clustering measurements. For the test set, we used 20 additional HODs configurations to populate six additional test cosmologies. We chose the parameter ranges according to the bright galaxy sample of the DESI one percent survey (PRADA et al. 2023). In the following, we use $X_{\rm HOD}$ to refer to the 600×5 matrix containing the HOD parameters of the training set. Note that using the same HOD sampling for every cosmology results in a grid-structured parameter space $X_\Omega \otimes X_{\rm HOD}$, where $\otimes$ is the Kronecker product. This peculiar data structure is convenient when building the emulator, which we demonstrate in section 4.2. Figure 4.3 illustrates the different forms that the number of galaxies as a function of halo mass takes for the range of HOD parametrisation that we use. The black lines show the mean HOD configuration for $\langle N_{\rm cen}(M) \rangle$, $\langle N_{\rm sat}(M) \rangle$, and $\langle N_{\rm tot}(M) \rangle$. Table 4.1 summarises the sampling range of every cosmological and HOD parameter.

[2] https://github.com/abacusorg/abacusutils

TABLE 4.1 : Cosmological and HOD parameter names and ranges sampled to build the emulator training and testing sets. The cosmology bounds come from the smallest hypercube around the AbacusSummit cosmology grid. The HOD bounds come from the minimum and maximum in each parameter within the 600 HOD.

	Parameter	Range
Cosmology	$\omega_{\rm cdm}$	[0.103 ,0.140]
	$\omega_{\rm b}$	[0.0207,0.0243]
	σ_8	[0.678,0.938]
	w_0	[-1.271,-0.726]
	w_a	[-0.628,0.621]
	h	[0.575 , 0.746]
	n_s	[0.901, 1.025]
	$N_{\rm ur}$	[1.020 , 3.046]
	α_s	[-0.038 , 0.038]
HOD	α	[0.30, 1.48]
	κ	[0.00, 0.99]
	$\log_{10} M_1$	[13.6, 15.1]
	$\log_{10} M_{\rm cut}$	[12.5, 13.7]
	$\log_{10} \sigma$	[-1, 0]

4.1.3 Observable of interest : The galaxy correlation function

To validate our emulation framework, we use the standard galaxy TPCF, measured using the estimators derived in section 2.2.3. Since we have access to the three-dimensional peculiar velocities of our simulated galaxy samples, we can apply the RSD as

$$d_s = d + \frac{v_\parallel}{H(z)}(1+z), \tag{4.3}$$

where d and d_s are the true and distorted radial positions, and $v_\parallel$ is the peculiar velocity along the LoS. Note that we are using simulation snapshots instead of light cones. Hence, the redshift evolution is not simulated and here z has the same value of 0.2 for every galaxy. We further decompose the anisotropic redshift space TPCF into a series of Legendre polynomials

$$\xi(s,\mu) = \sum_\ell L_\ell(\mu)\xi_\ell(s), \tag{4.4}$$

with

$$\xi_\ell(s) = \frac{2l+1}{2}\int \xi(s,\mu)L_\ell(\mu)d\mu. \tag{4.5}$$

As this work is a proof of concept for our new emulation strategy, we focus on the monopole ξ_0 hereafter. The study of additional multipoles is left for the following chapter. The most expensive part of computing the TPCF estimators is the random-random pair counting $SS(s,\mu)$. As we have cubic boxes with periodic boundary conditions, we can use the natural estimator defined in Eq

2.96, so *DS* is not needed and *SS* can be analytically computed following Eq 2.95. Doing so greatly reduces the total computational time, given that we have to compute many correlation functions to build our train and test sets. However, because it assumes a fixed LoS for every pair, wide angle effects (Eq 2.104) could bias the measurements, specially for the largest scales considered. We assess if this effect is significant using the 25 Planck2018 boxes populated with the median HOD of the LHS. Every pair count is done using the package PYCORR[3].

Figure 4.4 shows the mean clustering with its standard deviation computed using Eq. 2.96 with a fixed LoS in red, and using Eq. 2.85 with a varying midpoint LoS in blue. In the latter case, in order to get a realistic clustering, we place the observer at the centre of the box and apply a spherical full-sky footprint from $z = 0.15$ to $z = 0.25$ in radial distance. The bottom panel shows the residual using the cosmic variance of the varying LoS, which is of course larger and more realistic. The shaded area corresponds to a 1σ deviation. We find that, although the deviation seems to increase as the separation s gets larger, the difference is well within the cosmic variance for the scales we probe. Since the wide angle projection mainly depends on the effective redshift of the sample, and the separation bins considered, any projected footprint distribution $P(\theta)$ (see Eq 2.104) should be substantially as peaked around zero as for a full sky spherical footprint. Consequently, we consider that our clustering measurements are unbiased for the redshifts and separations of interest, and we use the Peebles estimator to build our train and test sets.

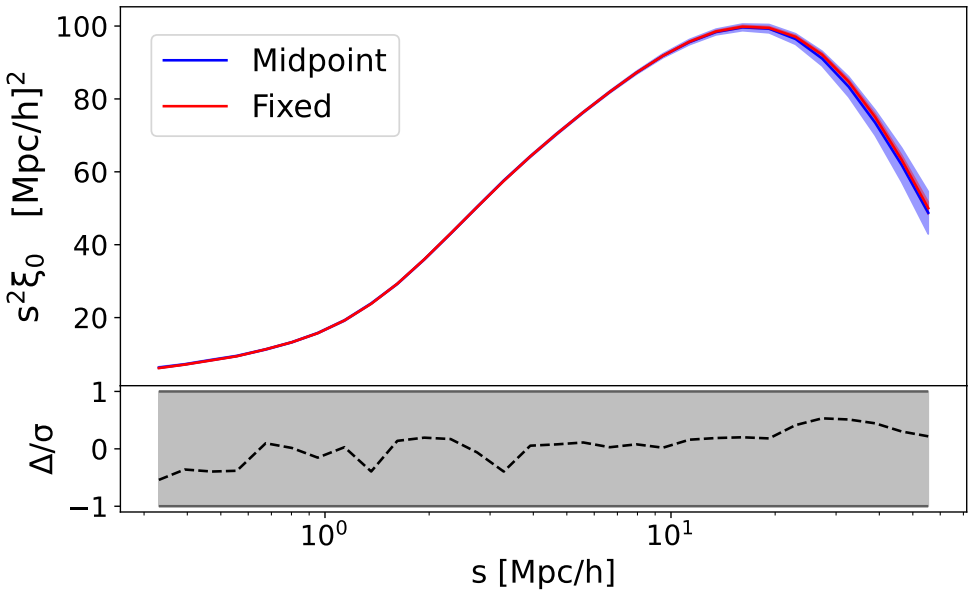


FIGURE 4.4 : Impact of wide-angle effects in the monopole of the correlation function. The mean (solid lines) and standard deviation (shaded areas) of the clustering is computed using the Planck2018 25 boxes with the median HOD. In red, ξ_0 is computed using Eq. 2.96 with a fixed LoS, and in blue using Eq. 2.85 with a varying midpoint LoS. In the latter case, in order to get a realistic clustering, we place the observer at the centre of the boxes, and cut a spherical full-sky footprint from z = 0.15 to z=0.25. The bottom panel shows the residual using the cosmic variance of the (realistic) varying LoS.

[3] https://github.com/cosmodesi/pycorr

Figure 4.5 shows the clustering measured for every cosmology and HOD configurations in the train and test sets, using Eq. 2.96 and applying RSD along a fixed LoS, which we choose to be the z axis of the box. A sample of the training and test sets is shown in blue and orange, respectively. We choose the separation range to be a logarithmic binning going from the resolution of the simulation 0.3 h^{-1}Mpc to 60 h^{-1}Mpc where the flat-sky approximation is still holds. Hereafter, we define X_s as the vector containing those separations (used in section 4.2). To reduce shot noise, we populate five times the halos using the same HOD model, only changing the random seed, and we average their clustering.

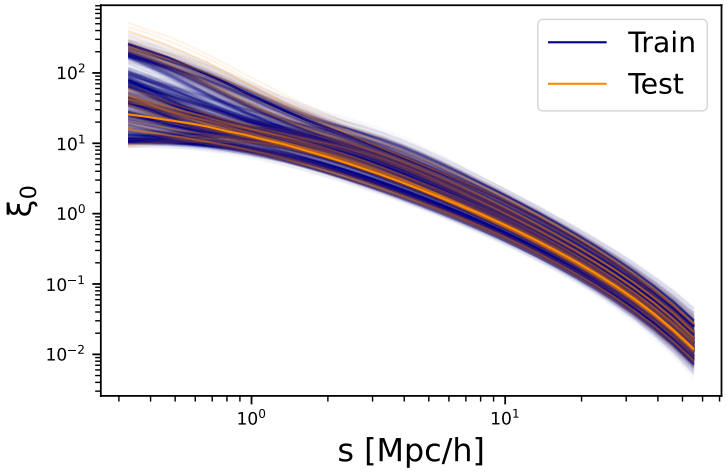


FIGURE 4.5 : Sub-sample of the two-point correlation function measurements used as the emulator train and test sets, in blue and orange respectively. The clustering for every cosmology and HOD is measured using Eq. 2.96 after applying RSD along a fixed LoS, which we choose to be the z axis of the box.

Although ABACUSSUMMIT covers a large volume, the training and test cosmology boxes are run with the same initial phase, and the measured clustering is still subject to cosmic variance. To quantify the uncertainties on our measurements, we measure the TPCF on 1400 of the small Planck2018 boxes using the median HOD and rescaling the volume, to build the covariance matrix C_{cosmic}. The corresponding correlation matrix is shown in figure 4.6. We see that, while the scales below a few h^{-1}Mpc are almost uncorrelated, the correlations for separations above 2.7 h^{-1}Mpc are strong and should not be neglected.

4.2 Multiscale Gaussian process model

In the following, we provide a concise overview of the Gaussian Process (GP) formalism. We also introduce our new model allowing to extend the dimension of the input parameter space, when provided with a gridded sampling for the training set. Our model extends the framework established in SAATCI 2011 to allow for non-diagonal noise kernels. Leveraging this formalism, we emulate every scale of the TPCF simultaneously. For a more complete introduction to GP, see RASMUSSEN et WILLIAMS 2006.

4.2.1 Standard GP regression model

Let $\mathscr{D} = \{(\boldsymbol{x_i}, y_i)|_{i=1}^n\}$ be a training set composed of n pairs of d-dimension input parameter vectors $\boldsymbol{x_i}$ and scalar observations y_i. Let X be the $n \times d$ matrix concatenating every input vector and Y an n column vector. We suppose that the observed values y_i are given by

$$y_i = f(\boldsymbol{x_i}) + \epsilon, \tag{4.6}$$

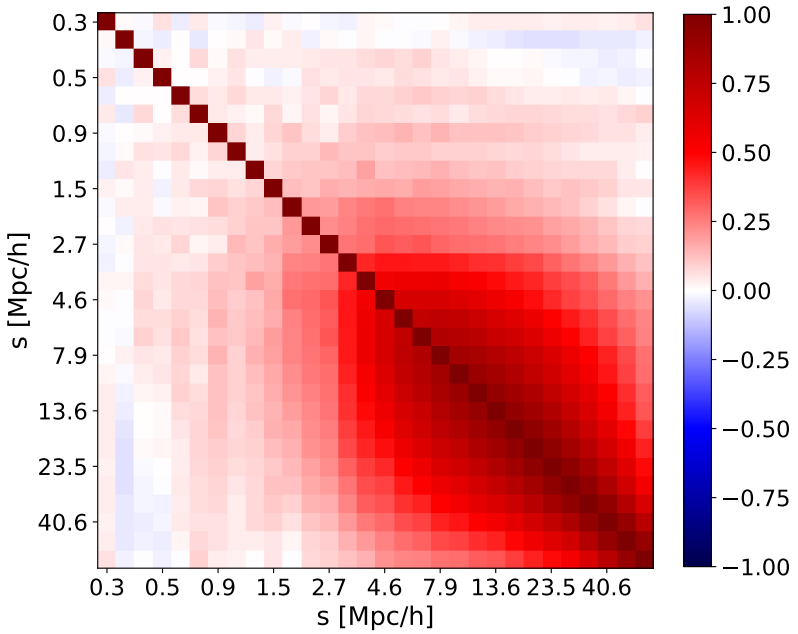


FIGURE 4.6 : Correlation matrix of the TPCF monopole computed from 1400 small box realisations of the Planck2018 cosmology, and using the median HOD. The covariance is rescaled to match the volume of the base boxes.

where f is the true underlying mapping function we wish to learn and ϵ is an additive noise following an independent identically distributed Gaussian distribution with mean zero and variance σ_n^2 :

$$\epsilon \sim \mathcal{N}\left(0, \sigma_n^2 I\right), \tag{4.7}$$

where I is the identity matrix. We assume that the mapping function can be described as

$$f(\boldsymbol{x}) = \Phi(\boldsymbol{x})^\top \boldsymbol{w}, \tag{4.8}$$

where $\Phi(\boldsymbol{x})$ is a set of n_f feature functions (e.g. polynomials $\Phi(\boldsymbol{x}) = (||\boldsymbol{x}||, ||\boldsymbol{x}||^2, ..., ||\boldsymbol{x}||^{n_f})$) mapping the d-dimensional input vector into an n_f-dimensional feature space, and $\boldsymbol{w}$ is an n_f-dimension vector of unknown weight (and with unknown parameters). This model is a linear function of $\boldsymbol{w}$ for any set of feature functions, and is therefore easier to handle analytically in the training process for instance.

The likelihood of observing Y for a set of input parameters X and weights $\boldsymbol{w}$ can then be written as following a Gaussian distribution :

$$p(Y|X, \boldsymbol{w}) \sim \mathcal{N}\left(\Phi(X)^\top \boldsymbol{w}, \sigma_n^2 I.\right) \tag{4.9}$$

Choosing a Gaussian prior with zero mean and covariance Σ for the weights,

$$p(\boldsymbol{w}) = \mathcal{N}(0, \Sigma), \tag{4.10}$$

and following Bayes' theorem, we can write the posterior distribution over the weights as

$$p(\boldsymbol{w}|Y, X) = \frac{p(Y|X, \boldsymbol{w})\, p(\boldsymbol{w})}{p(Y|X)}, \tag{4.11}$$

where $p(Y|X)$ is the evidence or marginal likelihood.

Making n_* predictions f_* for a new set of input vectors X_* (with X_* an $n_* \times d$ matrix) requires that we marginalise over the weight distributions. The evidence is simply a normalisation constant. Using the Gaussian properties of the prior and likelihood, an analytical expression for the posterior distribution of the prediction can be derived. This latter is expressed as :

$$\begin{aligned} p\left(f_*|\boldsymbol{x}_*, X, Y\right) &= \int p\left(f_*|\boldsymbol{x}_*, \boldsymbol{w}\right) p(\boldsymbol{w}|X, Y)\, \mathrm{d}\boldsymbol{w} \\ &\sim \mathcal{N}\left(\mu_f, \Sigma_f\right), \end{aligned} \tag{4.12}$$

where

$$\mu_f = \Phi_*^\top \Sigma \Phi \left(\Phi^\top \Sigma \Phi + \sigma_n^2 I\right)^{-1} Y, \tag{4.13}$$

$$\Sigma_f = \Phi_*^\top \Sigma \Phi_* - \Phi_*^\top \Sigma \Phi \left(\Phi^\top \Sigma \Phi + \sigma_n^2 I\right)^{-1} \Phi^\top \Sigma \Phi_*, \tag{4.14}$$

and $\Phi = \Phi(X)$ and $\Phi_* = \Phi(X_*)$. This posterior is a normal probability distribution over a function space. Note that making a prediction at some new location of the input space is equivalent to performing a Bayesian inference. Thanks to the linearity of the model and the Gaussian assumptions, the posterior probability can be analytically computed.

Let us now define the kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^\top \Sigma \Phi(\boldsymbol{x}')$ (note that $k(\boldsymbol{x}, \boldsymbol{x}')$ returns a scalar). Doing so allows us to switch from the weight-space view to the so-called function-space view. The expressions for the mean and covariance of the model posterior become :

$$\begin{aligned} \mu_f &= K(X_*, X)\left(K(X, X) + \sigma_n^2 I\right)^{-1} Y, \\ \Sigma_f &= K(X_*, X_*) - K(X_*, X)\left(K(X, X) + \sigma_n^2 I\right)^{-1} K(X, X_*), \end{aligned} \tag{4.15}$$

where $K(X, X_*)$ denotes the $n \times n_*$ covariance matrix built by evaluating the kernel $k(\boldsymbol{x}, \boldsymbol{x}_*)$ for the different values of $\boldsymbol{x}$ and $\boldsymbol{x}'$ (i.e $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_{*j})$). Specifying a family of feature functions, and letting their number n_f go to infinity leads to a compact form of kernel. The most popular in the literature (obtained using Gaussian feature functions) is the squared exponential kernel,

$$k_{\exp}(r) = \exp\left(\frac{-r^2}{2}\right), \tag{4.16}$$

or the Matern kernel,

$$\begin{aligned} k_{M_{3/2}}(r) &= \left(1 + \sqrt{3}r\right)\exp\left(-\sqrt{3}r,\right) \\ k_{M_{5/2}}(r) &= \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right)\exp\left(-\sqrt{5}r\right). \end{aligned} \tag{4.17}$$

These are stationary kernels; that is, functions of the distance between two points in the

input space, commonly defined as $r(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{x}-\boldsymbol{x}')^\top M(\boldsymbol{x}-\boldsymbol{x}')$ with $M = \boldsymbol{l}^{-2}I$, where $\boldsymbol{l}$ is a d-dimensional vector containing characteristic length-scales across each dimension of the input space. A large value for some component of $\boldsymbol{l}$ (compared to the input space variance) describes a smoothly varying signal in the corresponding dimension, allowing a larger range on interpolation. In practice, every kernel is also scaled by an additional hyperparameter σ^2 to account for the signal amplitude in the output space. Note that that the sum or product of different kernels also gives a valid kernel.

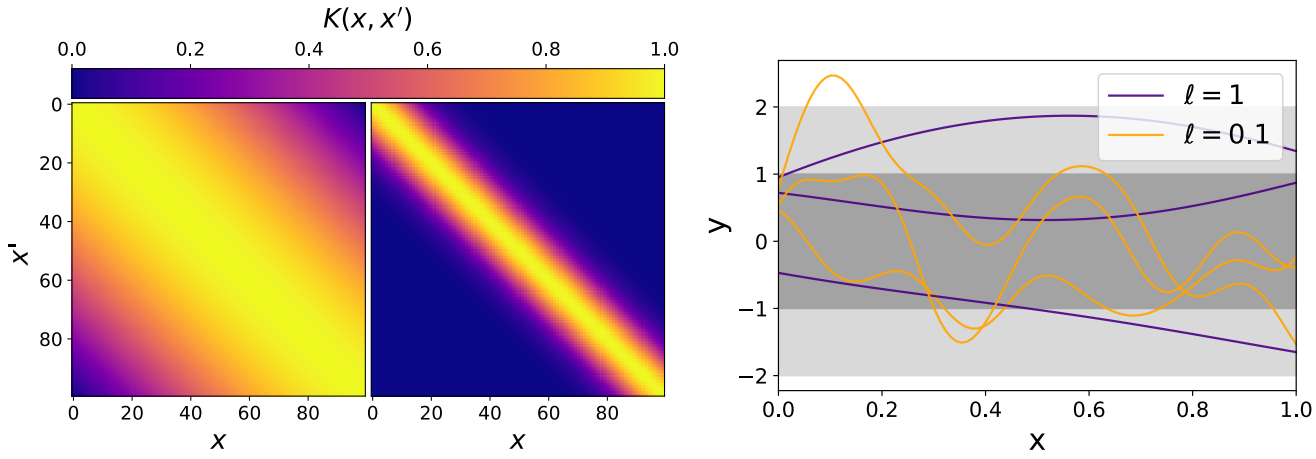


FIGURE 4.7 : Visualisation of squared exponential kernel for a 1D input parameter vector x composed of 100 points linearly sampled from 0 to 1. Left panel : Covariance matrices built using a variance of 1 and characteristic length-scales equal to 1 and 0.1 respectively. Right panel : Random functions sampled from the corresponding Covariance matrices. The grey shaded area indicates one and two times the standard deviation.

Figure 4.7 provides a visual representation of the squared exponential kernel function in the one dimensional case. The left panel depicts two covariance matrices built using characteristic length-scales equal to 1(left) and 0.1(right). In both case the variance σ^2 is set to one. We see that larger length-scales lead to larger spread around the diagonal as more scales are correlated. The right panel show three realisations randomly sampled from their corresponding kernels. The indigo and orange curves correspond to $\ell = 1$ and $\ell = 0.1$ respectively. The grey shaded area indicates one and two σ. We see that large ℓ (with respect to the input range) leads to smoothly varying function, while small ℓ allows for more flexibility. Selecting a kernel with a set of hyperparameters is equivalent to putting a prior $p(\boldsymbol{w}|\boldsymbol{\theta})$ over the weights (or the target function). Those considerations are generalisable to higher dimensions, with different characteristic length-scales. For a more comprehensive coverage we show in figure 4.8 three realisations randomly sampled from a two-dimensional squared exponential kernel with hyperparameters set to $\boldsymbol{\theta} = \{l_1 = 0.1, l_2 = 0.8, \sigma = 1\}$. Those 2D functions illustrate the ability of such kernel to vary quickly or smoothly along the relevant directions.

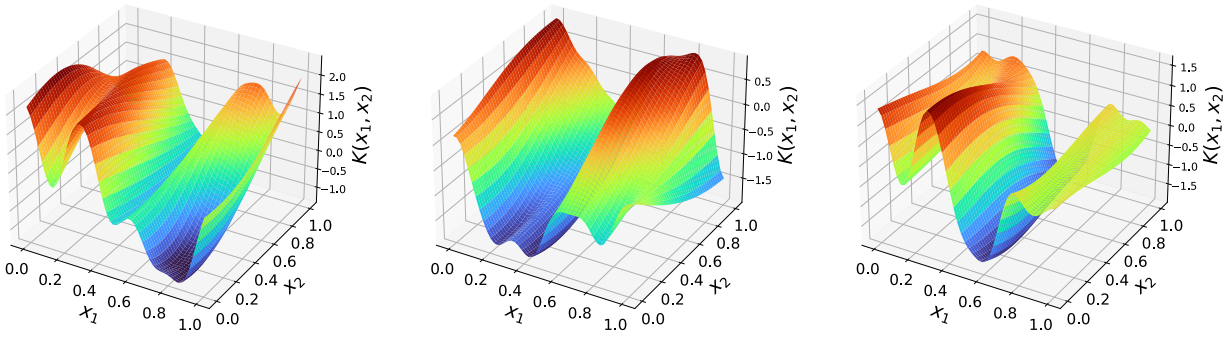


FIGURE 4.8 : Realisations randomly sampled from a 2D squared exponential kernel with hyperparameters $\boldsymbol{\theta} = \{l_1 = 0.1, l_2 = 0.8, \sigma = 1\}$. The 2D input parameter vector is $X = (x_1, x_2)$, where x_1 and x_2 are composed of 50 points linearly sampled from 0 to 1.

The training of the GP consists in choosing one or more kernels and inferring the hyperparameters $\boldsymbol{\theta} = \{\boldsymbol{l}, \sigma\}$ from the input parameter matrix $\boldsymbol{X}$ and their corresponding output observations Y. This is equivalent to a two-level hierarchical Bayesian inference, where Eq. 4.11 is rewritten as

$$p(\boldsymbol{w}|Y, X, \boldsymbol{\theta}) = \frac{p(Y|X, \boldsymbol{w})\, p(\boldsymbol{w}|\boldsymbol{\theta})}{p(Y|X, \boldsymbol{\theta})}. \tag{4.18}$$

Now, maximising the posterior probability for new predictions requires that we maximise the evidence $p(Y|X, \boldsymbol{\theta})$. This term now acts as a likelihood in the second level of the hierarchical Bayesian inference. The posterior over the hyperparameters can then be expressed as

$$p(\boldsymbol{\theta}|Y, X) = \frac{p(Y|X, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(Y|X)}, \tag{4.19}$$

where $p(\boldsymbol{\theta})$ is the hyper-prior and $p(Y|X)$ is the hyper-evidence. This second level inference is usually not analytically tractable and requires algorithmic maximisation of the log-likelihood :

$$\log\left[p(Y|X, \boldsymbol{\theta})\right] = -\frac{1}{2} Y^\top \left(K(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 I\right)^{-1} Y - \frac{1}{2} \log\left|K(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 I\right| - \frac{n \log(2\pi)}{2}. \tag{4.20}$$

The first term on the right-hand side of Eq 4.20 describes the goodness of the fit, while the second term is the complexity penalty (Occam's razor), independent of the observables Y, and allows to reduce overfitting. We recall the standard result for the gradient of the log-likelihood with respect to a given hyperparameter θ_i :

$$\frac{\partial \log p}{\partial \theta_i} = \frac{1}{2}\left(\boldsymbol{\alpha}^\top \frac{\partial K_y}{\partial \theta_i} \boldsymbol{\alpha} - \mathrm{Tr}\left(K_y^{-1} \frac{\partial K_y^\top}{\partial \theta_i}\right)\right), \quad \text{with } K_y = \left(K + \sigma_n^2 I\right) \text{ and } \boldsymbol{\alpha} = K_y^{-1} Y. \tag{4.21}$$

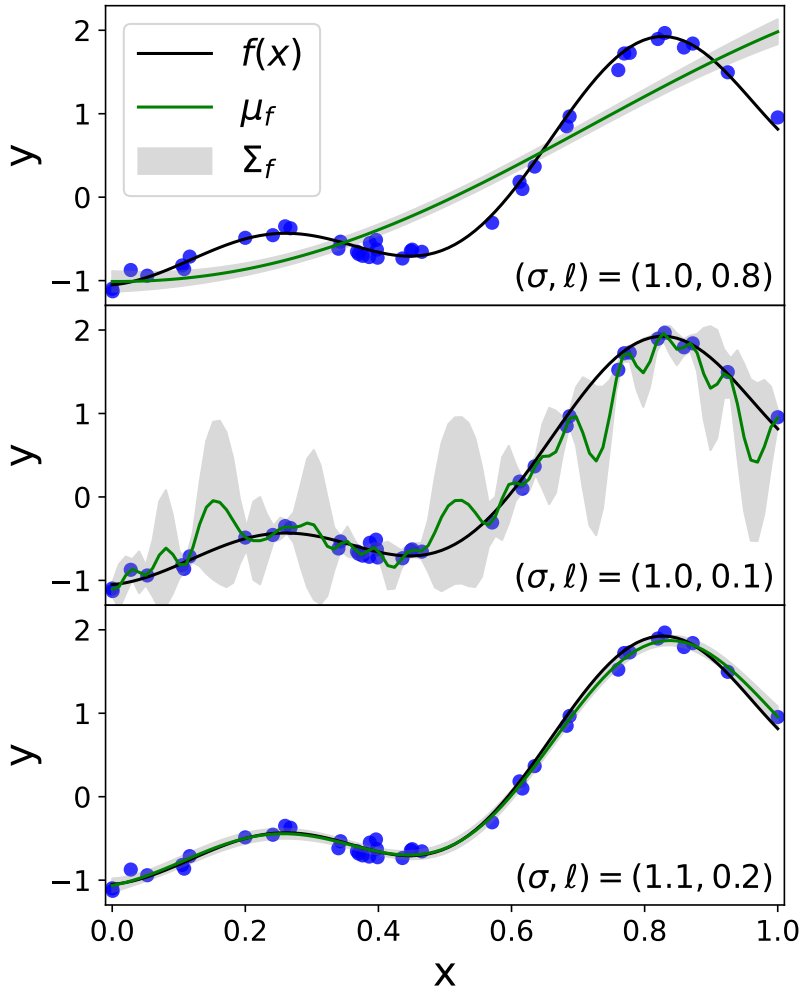


FIGURE 4.9 : Mean and covariance model predictions for different hyperparameter's choice. We sample some input/output training points $\big(x, f(x)\big)$ from a known 1D function $f(x)$ with some random Gaussian noise. The black lines describe the underlying function and the dotted blue points show the training sets. The green lines and the grey shaded area describe the mean and covariance prediction respectively for different hyperparameters value. The bottom panel corresponds to the results obtained with the hyperparameters maximising the log-likelihood.

Figure 4.9 shows a sketch the impact of the hyperparameter's choice on the model prediction. We sample 40 one-dimensional input points x from a known function $f(x)$ with some random Gaussian noise, and infer the mean and covariance predictions from Eq 4.15, using different values for the hyperparameters. The two first panels show the predictions (in green) obtained using deliberately too large and too small length-scales and a variance of one, while the bottom panels shows the results obtained with the hyperparameters maximising the log-likelihood in Eq 4.20. On one hand, too large length-scale leads to smooth functions without enough flexibility

to accurately describe the output values y. On the other hand, too small length-scale allows to accurately describe the training set, interpolation fails and quickly go back to the mean of the output (here 0).

Training a GP (adjusting the hyperparameters) requires that we compute the log-likelihood and its gradient several times, which involves obtaining the inverse and the determinant of the $n \times n$ matrix K_y, for each step of the minimisation. For a training set larger than a few thousand samples, these operations can become computationally prohibitive. For example, considering our case described in section 4.1, this matrix would have $n = n_\Omega \times n_{\rm HOD} \times n_s = 88 \times 600 \times 30 \sim 10^6$, where n_Ω, $n_{\rm HOD}$, and n_s are the number of sampled cosmologies, HODs, and separations, respectively.

4.2.2 Multiscale GP model

To overcome the issue of large matrix inversions for the standard GP, we can use a particular way of sampling the training data input parameters. This is a process known as Kronecker GP modelling, as is described below. Let the input parameter matrix of the training set X be composed of n samples of d dimensions. Suppose it can be decomposed as a Kronecker product, $X = X_1 \otimes X_2$, where X_1 and X_2 have smaller dimensions ($n_1 \times d_1$ and $n_2 \times d_2$ respectively) with $n = n_1 \times n_2$ and $d = d_1 + d_2$. This decomposition is only possible if the training set is grid-sampled (e.g. the same X_2 parameters are repeated for each set of X_1 parameters). This sampling is suboptimal, but as a result the covariance kernel can be computed as a Kronecker product of two lower-dimension matrices $K = K_1 \otimes K_2$.

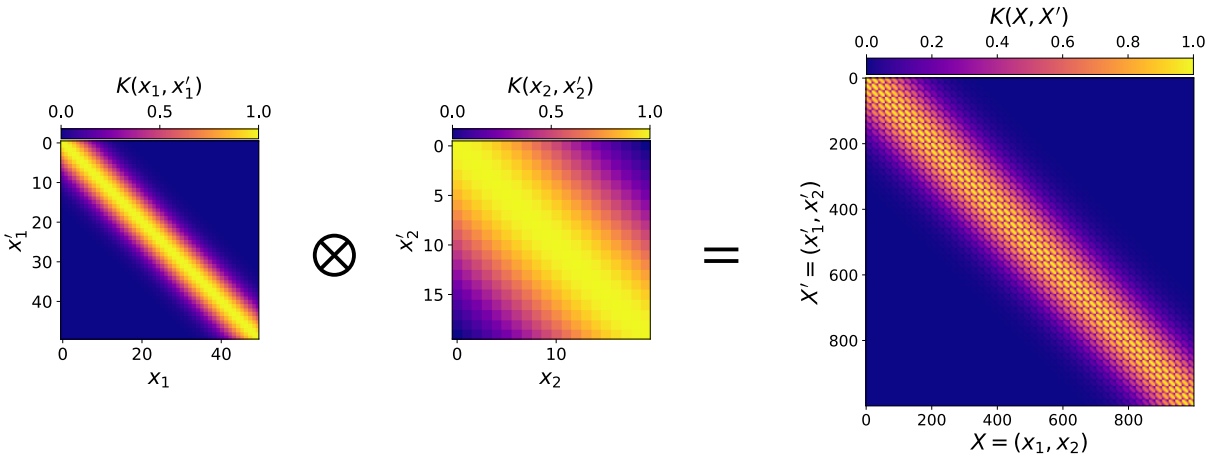


FIGURE 4.10 : Visualisation of Kronecker decomposition in the two-dimensional case using squared exponential kernels. x_1 and x_2 are 1D vectors composed of 50 and 20 points respectively, linearly sampled between 0 and 1. The right panel shows the total covariance matrix $K(X, X')$ built for an input space $X = (x_1, x_2)$, where for each point $x_{1,i}$, the full vector x_2 is repeated. Because of the grid structure of X, $K(X, X')$ can be decomposed as a Kronecker product of two kernels acting independently of the two sub-dimensions $K(x_1, x_1')$ and $K(x_2, x_2')$.

Figure 4.10 illustrates such two-dimensional decomposition. This specific structure allows us to make use of the algebraic properties of Kronecker products. If K is positive definite, it

can be written as a principal component decomposition $K = Q\Lambda Q^\top$, where $Q = Q_1 \otimes Q_2$ and $\Lambda = \Lambda_1 \otimes \Lambda_2$, with $K_i = Q_i \Lambda_i Q_i^\top$, with every Λ matrix diagonal. Thus, every calculation can be performed in sub-dimensional space. Such a model was implemented in the Python library GPy [4] but only for Kronecker decompositions into two subspaces. Saatci 2011 gives a generalisation of this framework to k decompositions $X = \bigotimes_{i=1}^k X_i$, but this latter is only valid for a diagonal uncorrelated noise matrix $N = \sigma^2 I$.

In this work, we extended this framework to general correlated noise covariance matrices of the form $N = \bigotimes_{i=1}^k N_i$. This is important in galaxy clustering because the noise covariance matrices are generally far from being diagonal, as seen in figure 4.6. The essential step is now to compute the inverse and the determinant of $(K + N)$ in Eq. 4.20. Assuming a Kronecker-structured training set and performing the decompositions $N_i = U_i S_i U_i^\top$, we can derive

$$\begin{aligned}
(K+N)^{-1} &= \left(\bigotimes_{i=1}^k K_i + \bigotimes_{i=1}^k N_i\right)^{-1} \\
&= \left(\bigotimes_{i=1}^k K_i + \bigotimes_{i=1}^k U_i S_i U_i^\top\right)^{-1} \\
&= \left(\bigotimes_{i=1}^k K_i + \bigotimes_{i=1}^k U_i \bigotimes_{i=1}^k S_i \bigotimes_{i=1}^k U_i^\top\right)^{-1} \\
&= \bigotimes_{i=1}^k U_i \bigotimes_{i=1}^k S_i^{-\frac{1}{2}} \left(\bigotimes_{i=1}^k \tilde{K}_i + \bigotimes_{i=1}^k I_i\right)^{-1} \bigotimes_{i=1}^k S_i^{-\frac{1}{2}} \bigotimes_{i=1}^k U_i^\top \\
&= \bigotimes_{i=1}^k U_i \bigotimes_{i=1}^k S_i^{-\frac{1}{2}} \bigotimes_{i=1}^k Q_i \left(\bigotimes_{i=1}^k \Lambda_i + \bigotimes_{i=1}^k I_i\right)^{-1} \bigotimes_{i=1}^k Q_i^\top \bigotimes_{i=1}^k S_i^{-\frac{1}{2}} \bigotimes_{i=1}^k U_i^\top,
\end{aligned} \tag{4.22}$$

where we use different properties of the Kronecker product, and define the projected matrices $\tilde{K}_i = S_i^{-\frac{1}{2}} U_i^\top K_i U_i S_i^{-\frac{1}{2}}$, which once built can be decomposed as $\tilde{K}_i = Q_i \Lambda_i Q_i^\top$.

Now, building $\Lambda = \bigotimes_{i=1}^k \Lambda_i$ is trivial as Λ_i are all diagonal matrices. Moreover, $\bigotimes_{i=1}^k I_i = I$, and therefore the $n \times n$ diagonal matrix $(\Lambda + I)$ is easily built and inverted. Additionally, $\left|\bigotimes_{i=1}^k U_i\right| = 1$ and $\left|\bigotimes_{i=1}^k Q_i\right| = 1$, and so the determinant from Eq. 4.20 is simply the product of the eigenvalue products $\left|S^{-1/2}\right| |\Lambda + I| \left|S^{-1/2}\right|$. The goodness-of-fit term is computed by iteratively using the KRONMVPROD algorithm (described in Saatci 2011), which evaluates operations of the form $\boldsymbol{\alpha} = \left(\bigotimes_{i=1}^k A_i\right) \boldsymbol{b}$, where $\boldsymbol{b}$ is a column vector.

We also require the evaluation of log-likelihood gradients. The first term of Eq. 4.21 can be easily computed as the hyperparameters are specific to a particular subdimension d, and therefore to a kernel K_d; that is,

$$\begin{aligned}
\frac{\partial K}{\partial \theta_i} &= \frac{\partial K_i}{\partial \theta_i} \otimes \left(\bigotimes_{j \neq i} K_j\right) \\
\frac{\partial N}{\partial \theta_i} &= \frac{\partial N_i}{\partial \theta_i} \otimes \left(\bigotimes_{j \neq i} N_j\right),
\end{aligned} \tag{4.23}$$

while the second term of Eq. 4.21 can be computed using the cyclic properties of the trace

[4] https://github.com/SheffieldML/GPy

operator :

$$\begin{aligned}\mathrm{Tr}\left(K^{-1}\frac{\partial K^\top}{\partial\theta_i}\right) &= \mathrm{diag}\left(\bigotimes_{i=1}^{k}\Lambda_i+\bigotimes_{i=1}^{k}I_i\right)^\top \mathrm{diag}\left(H^\top\frac{\partial K^\top}{\partial\theta_i}H\right)\\ \mathrm{Tr}\left(N^{-1}\frac{\partial N^\top}{\partial\theta_i}\right) &= \mathrm{diag}\left(\bigotimes_{i=1}^{k}\Lambda_i+\bigotimes_{i=1}^{k}I_i\right)^\top \mathrm{diag}\left(H^\top\frac{\partial N^\top}{\partial\theta_i}H,\right)\end{aligned} \tag{4.24}$$

with $H=\bigotimes_{i=1}^{k}U_i\bigotimes_{i=1}^{k}S_i^{-\frac{1}{2}}\bigotimes_{i=1}^{k}Q_i$. The last terms in the above equations are given by

$$\begin{aligned}H^\top\frac{\partial K^\top}{\partial\theta_i}H &= H_i^\top\frac{\partial K_i^\top}{\partial\theta_i}H_i\otimes\left(\bigotimes_{i\neq j}H_j^\top\frac{\partial K_j^\top}{\partial\theta_i}H_j\right)\\ H^\top\frac{\partial N^\top}{\partial\theta_i}H &= H_i^\top\frac{\partial N_i^\top}{\partial\theta_i}H_i\otimes\left(\bigotimes_{i\neq j}H_j^\top\frac{\partial N_j^\top}{\partial\theta_i}H_j\right).\end{aligned} \tag{4.25}$$

We provide an implementation of this model using the GPy framework and based on JAX GPU parallelisation for algebra operations, which is publicly available as the MKGPy[5] package.

4.2.3 Application to galaxy clustering

To apply this model to our observable, we decompose the parameter space in a gridded structure $X = X_\Omega \otimes X_{\mathrm{HOD}} \otimes X_{\mathrm{s}}$. Specifically, as mentioned in section 4.1.2, we measure the correlation function on the same scales using the same HOD sampling for every cosmology, allowing us to decompose the signal and noise covariances as

$$\begin{aligned}K &= K_\Omega\otimes K_{\mathrm{HOD}}\otimes K_{\mathrm{s}}\\ N &= N_\Omega\otimes N_{\mathrm{HOD}}\otimes N_{\mathrm{s}}.\end{aligned} \tag{4.26}$$

We find that the optimal signal kernels (maximising the log-likelihood) are simply two Matern3/2 kernels for cosmological and HOD spaces and a squared exponential kernel for the scale space. This choice of modelling leads to 16 signal hyperparameters $\theta=\{\boldsymbol{l_\Omega}, \boldsymbol{l}_{\mathrm{HOD}}, l_{\mathrm{s}}, \sigma\}$, which include 9 length-scales for the cosmological parameters $\boldsymbol{l}_\Omega$, 5 for the HOD parameters $\boldsymbol{l}_{\mathrm{HOD}}$, one for the separations l_s, and one overall variance of the signal σ. For the noise kernels, we chose N_{s} to be the fixed covariance matrix of the TPCF monopole computed from the 1400 small boxes (figure 4.6). For the cosmological and HOD noise kernels, we chose diagonal heteroscedastic kernels of the form $N(x,x')=\boldsymbol{\sigma}_n(x)^2\delta_{\mathrm{D}}(x-x')I$, allowing a different variance for every given configuration of HOD and cosmological parameters. We allow $\boldsymbol{\sigma_n}$ to be fully parameterised vectors, leading to 88 noise hyperparameters for each cosmology and 600 for each HOD. Those noise hyperparameters $\left(\boldsymbol{\sigma}_{\boldsymbol{n}_\Omega}, \boldsymbol{\sigma}_{\boldsymbol{n}_{\mathrm{HOD}}}\right)$ are fitted along with the signal hyperparameters θ. To avoid mathematical overflow and to simplify the hyperparameter learning, we normalise each dimension i of the input and output training sets as follows :

[5] https://github.com/TyannDB/mkgpy

$$\tilde{X}^i_\Omega = \frac{X^i_\Omega - \min(X^i_\Omega)}{\max(X^i_\Omega) - \min(X^i_\Omega)}, \quad \tilde{X}^i_{\rm HOD} = \frac{X^i_{\rm HOD} - \min(X^i_{\rm HOD})}{\max(X^i_{\rm HOD}) - \min(X^i_{\rm HOD})},$$
$$\tilde{X}_{\rm s} = \frac{\log(X_{\rm s}) - \log(\min(X_{\rm s}))}{\log(\max(X_{\rm s})) - \log(\min(X_{\rm s}))}, \tag{4.27}$$

where X_Ω and $X_{\rm HOD}$ are normalised to be distributed in the range $[0,1]$. Because we use stationary kernels $k(x,x') = k(|x-x'|)$, we need to make sure that the variance of the signal is roughly constant as a function of the input x, which is not the case along the subspace $X_{\rm s}$. For this input dimension, we use a logarithmic transformation to flatten the signal before applying the normalisation. We also normalise the output training set Y to be normally distributed around zero with a variance of one. Because the TPCF spans a large range of amplitudes, we also apply a logarithmic transformation before doing so :

$$\tilde{Y} = \frac{\log(Y) - \text{mean}(\log(Y))}{\text{std}(\log(Y))}. \tag{4.28}$$

The covariance matrix used as $N_{\rm s}$ must also be projected in the normalised space. To do so, we use the error propagation,

$$\tilde{N}_{\rm s} = J N_{\rm s} J^\top, \tag{4.29}$$

with J the mapping Jacobi matrix $J_{ij} = \frac{\partial \tilde{Y}_i}{\partial Y_j}$ computed for the mean of the 1400 small boxes. Once the emulator is trained, every new input configuration $X_{\rm test}$ should be normalised using Eq. 4.27, and every model prediction μ_f and its covariance Σ_f (Eq. 4.15) should be transformed back into the real space using Eqs. 4.28 and 4.29.

The hyperparameters are optimised using the SCIPY gradient-based algorithm LBFGS with ten random initialisations and selecting the best-fit hyperparameters yielding the highest likelihood. To compare performances, we train both the standard GP, where each separation is treated independently (as separate emulators), and our new multiscale GP. We employ the same kernels, normalisation, and optimisation method.

4.3 Emulator performance

In this section, we compare the performances of two emulators : first, the independent Gaussian process model (IGP), where each scale s of the correlation function $\xi(s)$ has its own independent GP depending on cosmological and HOD parameters; and second, our multiscale Gaussian process model (MGP), which is able to predict the full model for $\xi(s)$ as well as an associated non-diagonal model covariance.

4.3.1 Prediction accuracy

Figure 4.11 compares the IGP and MGP models predictions compared to the test set. The grey lines show the relative difference between the MGP emulator prediction and the expected values from the test set as a percentage. The black dashed line corresponds to a 1% deviation. The blue and green lines are the median of those deviations for the MGP and IGP models, respectively, reaching a sub-percent precision for both models on most of the scales. However, the correlation

functions of our test set suffer from the same noise as the train set, and so we do not expect the difference $\Delta \equiv \xi^{\rm emu} - \xi^{\rm true}$ to be smaller than the estimated uncertainty of $\xi^{\rm true}$. The red dashed line shows the median of $\left|\frac{\sigma}{\xi^{\rm true}}\right|$ as a percentage, with σ being the diagonal of $C_{\rm cosmic}$. A difference smaller than this reference would indicate overfitting of the particular initial phase of the simulations, while larger differences would show that the prediction performance is poor. Both MGP and IGP models seem to be slightly overfitting the initial phase for the large scales, but follow the expected accuracy on small scales. We note that for the very small scales, the MGP is overfitting the test set to a lesser extent than the IGP. We see that the relative error of the full sample in grey varies around the cosmic variance expectation. The dotted red line corresponds to the cosmic variance of the central cosmology and HOD. The wide HOD scatter causes large variations in the clustering amplitude and therefore in the expected cosmic variance.

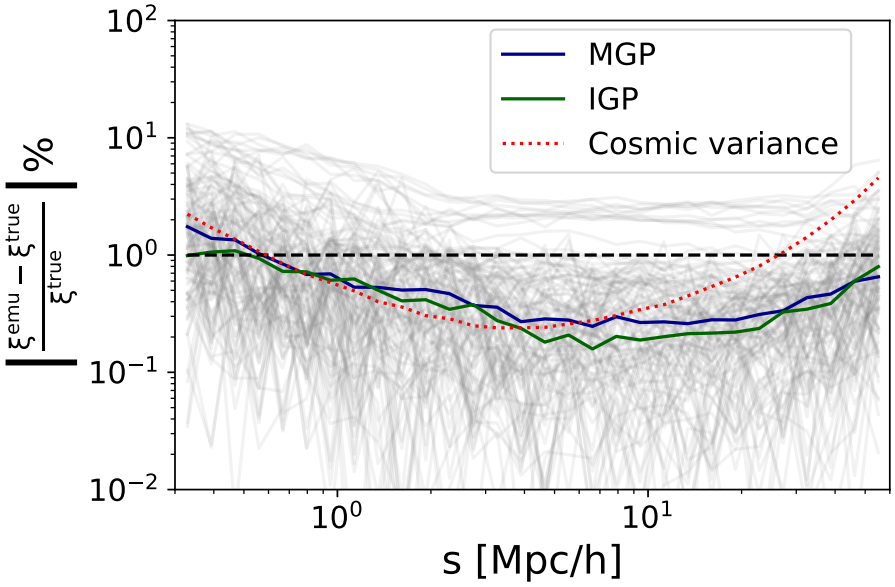


FIGURE 4.11 : Grey lines show the relative difference between the MGP emulator prediction and the expected values for the full test set, as a percentage difference. The blue and green lines are the median of those deviations for the MGP and IGP models, respectively. The black dotted line corresponds to a 1% deviation. The red dotted line shows the median of $|\sigma/\xi^{true}|$ as a percentage. Here, σ^2 is the cosmic variance computed for the Planck 2018 cosmology and median HOD.

4.3.2 Robustness to cosmic variance

Our models were trained on correlation functions from simulations run with different cosmologies and HOD parameters, but all starting from the same IC. Overfitting this particular seed could potentially bias our clustering predictions on simulations with different IC. We tested this issue with 25 phase realisations of AbacusSummit run with the same Planck2018 cosmology and the same set of HOD parameters.

Gaussian processes provide not only a prediction for a model, but also an estimate of the model uncertainty, which depends on the input parameters $\boldsymbol{\theta}$. This uncertainty is expressed as a covariance matrix, denoted $C_{\rm emu}(\boldsymbol{\theta})$. When performing inferences on real data, this model

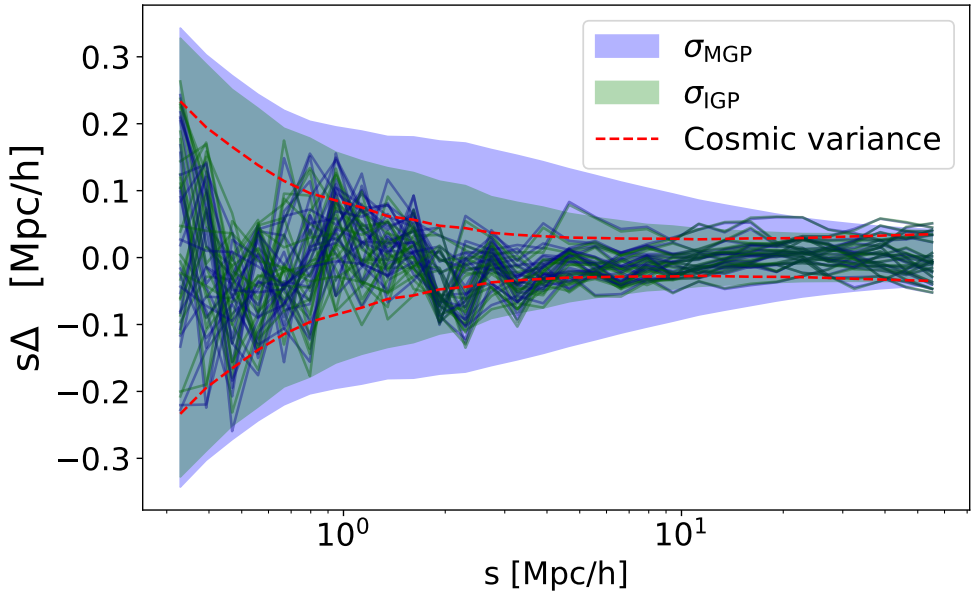


FIGURE 4.12 : Difference between the emulators prediction and the 25 realisations of Planck2018 cosmology with the median HOD in blue and green lines for the MGP and IGP, respectively. The red dashed line describes the cosmic variance level. The total variances, including the emulator variance predictions, are shown with the blue and green shaded area.

covariance has to be added to the measurement uncertainty for the considered scales C_{cosmic} as :

$$C_{\mathrm{tot}}(\theta_\Omega, \theta_{\mathrm{hod}}) = C_{\mathrm{cosmic}} + C_{\mathrm{emu}}(\theta_\Omega, \theta_{\mathrm{hod}}). \tag{4.30}$$

We deliberately dropped the dependency of C_{cosmic} on $\boldsymbol{\theta}$ as it is considered fixed during inferences (as discussed in section 2.3.3). Figure 4.12 shows the difference Δ between the 25 measurements and predictions from both the MGP (blue) and the IGP (green) emulators. The red dashed line displays the cosmic variance level while blue and green lines show the total variance (model + cosmic) according to Eq. 4.30. We see that for intermediate scales, Δ is larger than the cosmic variance level, while using C_{tot} lowers the deviation to less than 1σ. We also note that for this example the variance of the MGP is larger than the one of the IGP. However, Fig. 4.12 only presents the diagonal elements of the covariance matrices.

Figure 4.13 shows the total (model + cosmic) correlation matrix of the MGP. Compared to C_{cosmic} alone (Figure 4.6), the small scales are now significantly correlated. The predicted covariance matrix of the IGP is diagonal by construction, and so cannot account for such correlations, potentially affecting cosmological constraints. We investigate this issue in section 4.4.

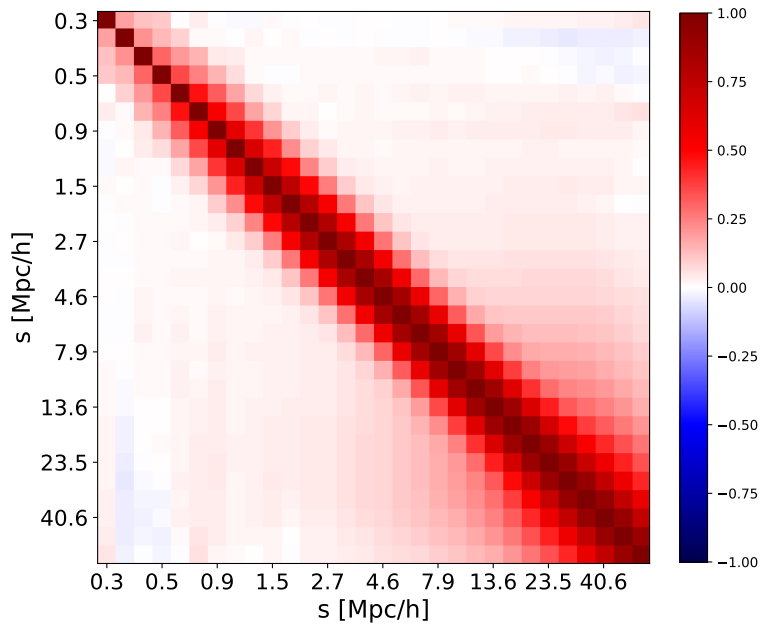


FIGURE 4.13 : Total correlation matrix of the MGP, including the model uncertainty prediction $C_{\text{tot}} = C_{\text{cosmic}} + C_{\text{emu}}(\theta_\Omega, \theta_{\text{hod}})$. The measurement uncertainty C_{cosmic} was evaluated for the Planck2018 cosmology with the median HOD using the small boxes.

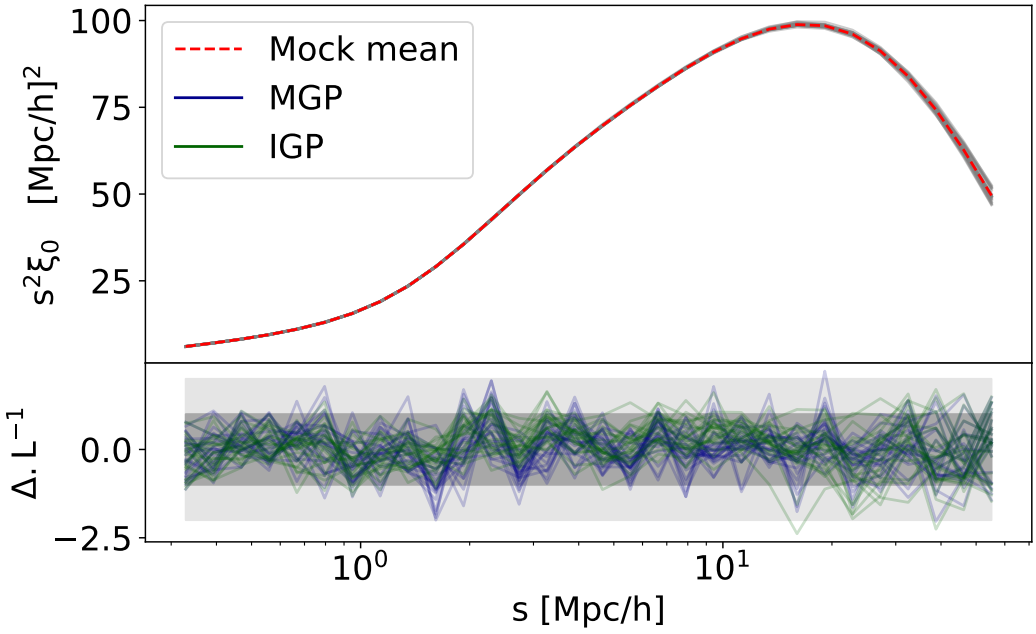


FIGURE 4.14 : Emulator performance on mocks with different phase. In the upper panel the 25 realisations of Planck2018 cosmology using the median HOD are shown in grey. The average is shown in red dotted line. The 25 residuals for both IGP (green) and MGP (blue) predictions are shown in the bottom panel. L^{-1} is the Cholesky decomposition of the inverse of C_{tot}. The grey shaded areas correspond to 1σ and 2σ deviations.

The clustering measurements for the 25 mocks are shown in the top panel of figure 4.14 with the average shown with a red dotted line. The bottom panel describes the 25 residuals for both IGP (green) and MGP (blue) predictions, with $\Delta = \xi^{\rm emu} - \xi^{\rm true}$, and L^{-1} is the Cholesky decomposition of the inverse of the total covariance $C_{\rm tot}$. Because the matrix is symmetric positive-definite, it can be decomposed as

$$C_{\rm tot} = L^\top L, \quad \text{and} \quad C_{\rm tot}^{-1} = L^{-1}(L^{-1})^\top, \tag{4.31}$$

where L is a triangular matrix with real and positive diagonal entries. This residual metric is useful as it takes into account the correlations between scales. The shaded areas correspond to 1σ and 2σ deviations. For most of the scales, both MGP and IGP models are able to reproduce the clustering for the different realisations within a 1σ deviation. For every scale, the residuals are within a 2σ deviation. It seems that for both models the predicted uncertainty $C_{\rm emu}$ solves the initial condition phase problem. In the following section, we study the recovering of the cosmological and HOD parameters.

4.4 Cosmology recovery and constraints

In this section, we describe the performances of our models in recovering unbiased cosmological and HOD parameters. We used the clustering measured on the 25 realisations of the Planck2018 cosmology using the same median HOD parameters.

4.4.1 Parameter inference using the emulator model

The parameter inferences are performed by running MCMC chains with the library EMCEE, using a flat prior corresponding to Table 4.1 and a standard Gaussian log-likelihood defined as

$$\log \mathscr{L}(\boldsymbol{\theta}) \propto -\frac{1}{2}\left(\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})\right) C_{\rm tot}^{-1}(\boldsymbol{\theta}) \left(\boldsymbol{d} - \boldsymbol{\mu}(\boldsymbol{\theta})\right)^\top - \frac{1}{2}\log|C_{\rm tot}(\boldsymbol{\theta})|, \tag{4.32}$$

where $\boldsymbol{d}$ is the data vector, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the model prediction for a given set of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_\Omega, \boldsymbol{\theta}_{\rm hod})$, and $C_{\rm tot}(\boldsymbol{\theta})$ is the total (model + cosmic) covariance defined in Eq. 4.30. Compared to a 'standard' cosmological inference, the terms $C_{\rm tot}^{-1}(\boldsymbol{\theta})$ and $\log|C_{\rm tot}(\boldsymbol{\theta})|$ must be computed for every iteration because of the emulator covariance dependence on the parameters. To do so, we make use of the fast and stable Cholesky decomposition, allowing to compute $\alpha = C_{\rm tot}^{-1}\left(\boldsymbol{d} - \boldsymbol{\mu}\right)$ efficiently. Besides, for a triangular matrix $|L| = \prod_i L_{ii}$, and we can write $|C_{\rm tot}| = |L^\top||L|$. Therefore, this method gives "for free" the determinant contribution as

$$\log|C_{\rm tot}| = 2\sum_i \log L_{ii}. \tag{4.33}$$

This term prevents the maximum likelihood to be located where the model error is the largest. We discuss alternative choice to $C_{\rm emu}$ in appendix 4.5. For a given set of cosmological parameters, we can infer the corresponding growth rate f. The uncertainty on the measured growth rate is simply propagated as $\sigma_f = \frac{\partial f}{\partial \Omega_i} C_{\Omega_{ij}} \frac{\partial f}{\partial \Omega_j}$, where $C_{\Omega_{ij}}$ is the covariance of the inferred cosmological parameters. We use the JACOBI[6] Python library to compute the gradient of our model.

[6] https://hdembinski.github.io/jacobi/

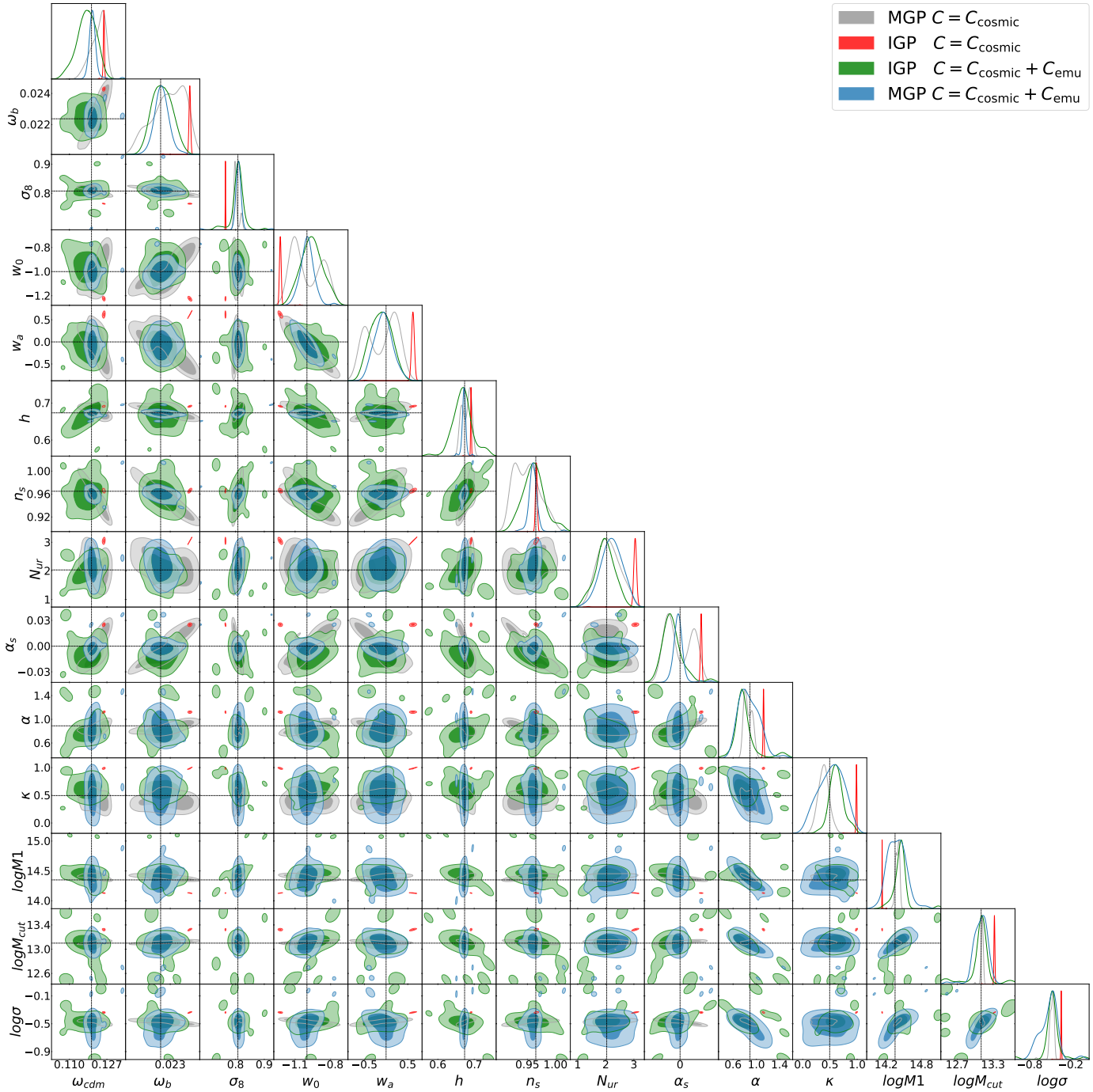


FIGURE 4.15 : Posterior distribution of the cosmological and HOD parameters obtained after running MCMC inference on the average stack of the 25 correlation functions with Planck 2018 cosmology and the median HOD. The results obtained using the IGP model with and without the model covariance prediction are shown in green and red, respectively, and those obtained using the MGP model with and without the model covariance prediction are shown in blue and black, respectively. The true parameters are shown by the black dotted lines.

4.4.2 Cubic boxes : average fit

We first tested the performance of both IGP and MGP models in recovering the parameters of interest when fitting a high-precision measurement, using the average stack of the 25 correlation function monopoles, corresponding to a volume of 200 $(h^{-1}\mathrm{Gpc})^3$. We used the likelihood defined in Eq. 4.32, divided the sample covariance C_{cosmic} by 25, and ran two inferences with and without using the emulator-predicted covariance C_{emu} for each model.

Figure 4.15 shows the 68% and 95% confidence regions obtained for the cosmological and HOD parameters. The true values are indicated by the dotted lines. The contours obtained using the IGP model with(without) the covariance prediction are shown in green(red), and those obtained using the MGP model with(without) the covariance prediction are shown in blue(black). We observe several interesting features. First, neglecting the model covariance leads to strongly biased results for the IGP with very sharp posterior distributions, while the MGP performs significantly better though yielding multimodal posteriors. Second, using the model covariance allows us to recover unbiased results for all cosmological parameters within one sigma for both models. Third, while MGP and IGP contours are consistent, the MGP model gives tighter constraints on cosmological parameters, and the IGP model gives tighter constraints on the HOD parameters. We also see several local maxima in the posterior distributions, especially in the green IGP contours. When sampling cosmological parameters far from those used in the training process of the GPs, the model covariance becomes larger while fitting the data points. This can also be seen as the goodness-of-fit term being counterbalanced by the determinant in the likelihood. Ignoring the model covariance can cause the fitter to converge towards a local maxima of the true log likelihood.

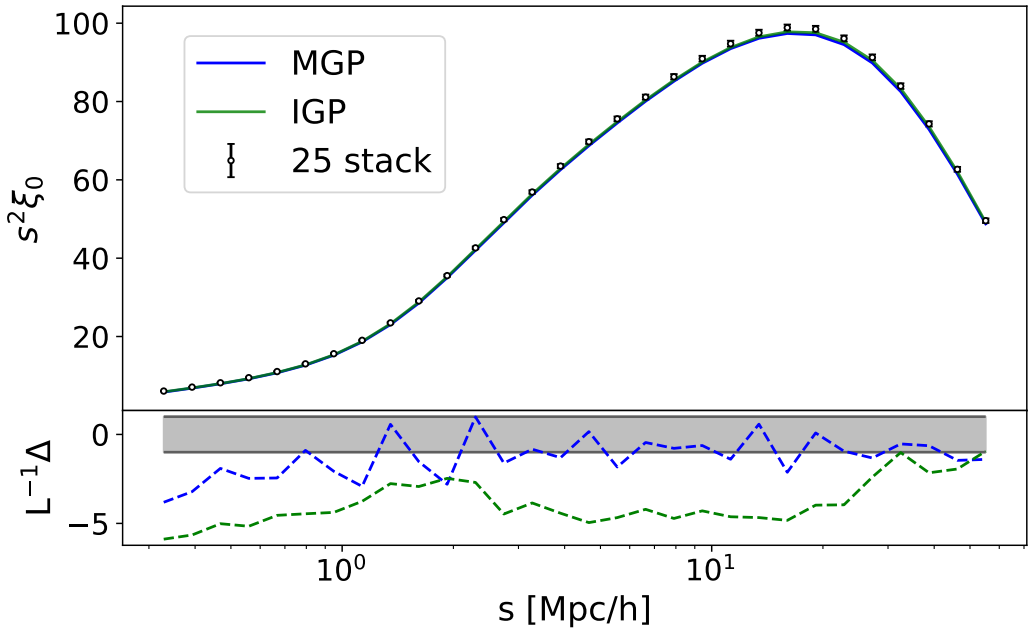


FIGURE 4.16 : The 25 realisations of Planck2018 cosmology with the median HOD in grey. The average is shown with a red dotted line and the MGP prediction is shown in blue. The 25 residuals for both IGP (green) and MGP (blue) predictions are shown in the bottom panel. L^{-1} is the Cholesky decomposition of the inverse of each C_{tot}.

Figure 4.16 compares the best fits of the MGP and IGP models in blue and green, respectively. The bottom panel displays the residuals, using the Cholesky decomposition matrix to account for correlations between scales. The 1σ deviation limit is set by the grey shaded area. For the MGP model, we get a reduced chi squared of $\chi^2_r = 5.6$ and for the IGP we obtain $\chi^2_r = 31$. These large χ^2_r values are due to the large volume probed by the average of 25 simulations. These χ^2_r values and the bottom panel of figure 4.16 show that the data are much better adjusted by the MGP model over all scales, as the accuracy of the model prediction is better estimated.

4.4.3 Cubic boxes : individual fits

In order to obtain a statistically significant measurement of the accuracy and potential biases of our models, we fit each of the 25 boxes separately, corresponding to a probed volume comparable to current surveys. The cosmic covariance matrices were normalised to account for this change in volume. Figure 4.17 shows the reduced chi-squared χ^2_r, the best-fit growth rate parameter f, and their estimated uncertainties σ_f. The derived cosmological parameters along with their inferred uncertainties are shown in Figure 4.18 and the same for the HOD parameters in Figure 4.19.

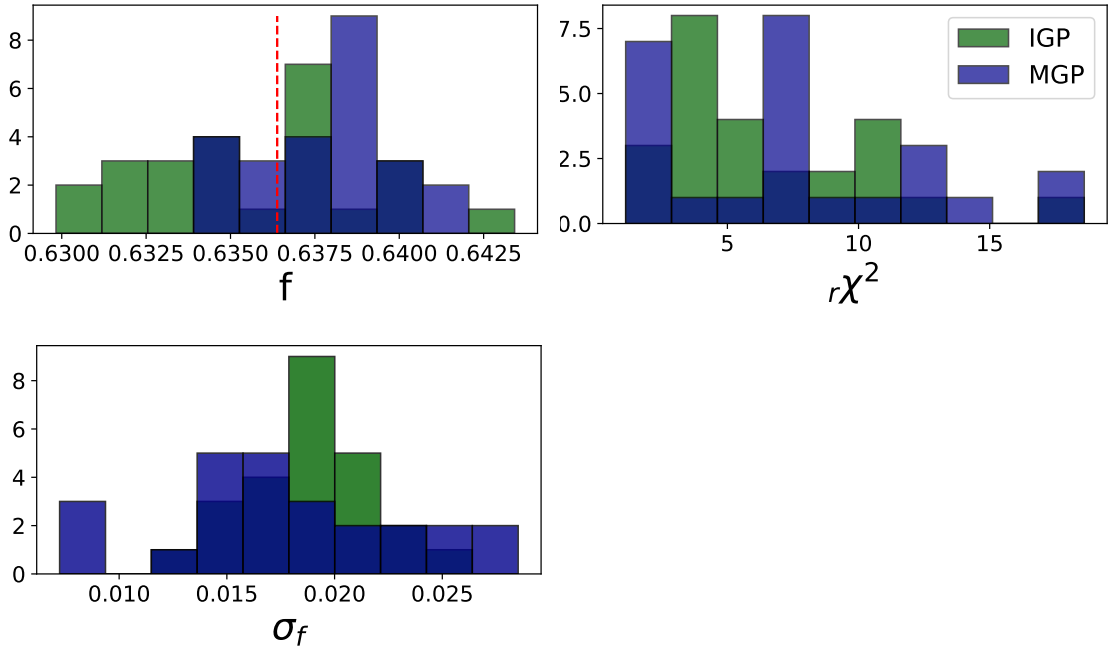


FIGURE 4.17 : Reduced chi squared, best-fit growth rate, and corresponding uncertainties resulting from the separate fits of the 25 cubic mocks with Planck2018 cosmology and the median HOD. Results using the MGP and IGP models are shown in blue and green, respectively. The true parameter values are shown with a red dotted line.

Table 4.2 reports some summary statistics from those results, where $\langle \Delta_p \rangle$ is the difference between the true and best-fit value for parameter p, $\langle \sigma_p \rangle$ is the average over the 25 realisations of the (symmetrised) 1σ confidence level, and std(p) is the standard deviation of the estimated parameter p over the 25 fits. We find that the uncertainty in the growth-rate measurement is reduced with our MGP model, and that both models give consistent results recovering the growth rate within 1σ. While the reduced chi-squared distributions are similar, the average is closer to one for the IGP model, with $\langle \chi^2_r \rangle = 4.93$, while we have $\langle \chi^2_r \rangle = 7.36$ for the MGP and $(30-14) = 16$ degrees of freedom. Those chi-squared give extremely unlikely p-values, less than

TABLE 4.2 : Summary statistics for the fits of the 25 cubic mocks with Planck2018 cosmology and the median HOD. $\langle\Delta_p\rangle$ is the bias, the difference between the true and best-fitted value for any parameter p, $\langle\sigma_p\rangle$ is the average over the 25 realisations of the (symmetrised) 1σ confidence level, and std(p) is the standard deviation of the estimated parameter p over the 25 results. The growth rate f is a derived parameter from the other cosmological parameters.

	IGP [$\times 10^{-2}$]			MGP [$\times 10^{-2}$]		
p	$\langle\Delta_p\rangle$	$\langle\sigma_p\rangle$	std(p)	$\langle\Delta_p\rangle$	$\langle\sigma_p\rangle$	std(p)
f	0.06	1.88	0.33	-0.14	1.80	0.20
ω_m	0.05	0.50	0.14	-0.02	0.16	0.04
ω_b	-0.02	0.09	0.02	-0.01	0.06	0.02
σ_8	-0.28	1.64	0.34	-0.05	0.36	0.04
w_0	-1.61	10.96	2.79	0.21	4.85	0.43
w_a	6.02	31.57	8.11	-0.22	19.04	2.57
h	0.15	2.33	0.45	0.08	0.65	0.12
n_s	0.43	2.53	0.63	0.11	0.72	0.16
N_{ur}	-1.15	41.97	8.31	-5.65	46.56	8.02
α_s	0.56	1.36	0.71	0.02	0.48	0.08
α	11.10	11.24	6.35	-0.38	19.03	4.19
κ	-10.20	20.31	6.42	-6.70	27.70	4.27
$\log M_1$	-7.67	8.86	4.69	0.39	17.39	3.09
$\log M_{cut}$	0.45	4.99	2.24	4.86	10.56	2.83
$\log\sigma$	-1.09	5.70	4.08	3.78	13.50	3.86

10^{-10}, indicating a poor model fitting. One reason could be that, while excluded from the train set, the cosmology of the 25 realisations fitted is located at the centre of the cosmological parameters' hypercube, leading to an underestimation of the model covariance due to the proximity with the training samples. Although it is hard to draw conclusions because of the limited size of the sample, in both cases the mean estimated error $\langle\sigma_f\rangle$ does not exactly match the standard deviation of the best-fit parameters std(p), but is conservatively larger. We also see that our MGP model yields smaller dispersion and results in tighter constraints on f than the IGP model.

Figure 4.18 indicates similar results for the cosmological parameters. Both models are able to retrieve the right parameters within 1σ, but the MGP returns smaller bias, smaller dispersion, and tighter constraints for all parameters except the number of ultra-relativistic species N_{ur}, which is poorly constrained. In particular, the uncertainty on amplitude parameter σ_8 is reduced by a factor of greater than 4, while reducing the bias on the best-fit value. For the HOD parameters illustrated in Figure 4.19, we also get unbiased measurements. Once again, the bias and the dispersion for the best-fit values are overall reduced with the MGP; however, the IGP gets tighter constraints. It therefore seems that the MGP performs better than the IGP because using the correlation between scales helps to constrain the cosmology rather than the HOD parameters, as the HOD signal is mainly in the small scales where the cosmic covariance is almost diagonal (see Figure 4.6), while all scales are affected by cosmological parameters.

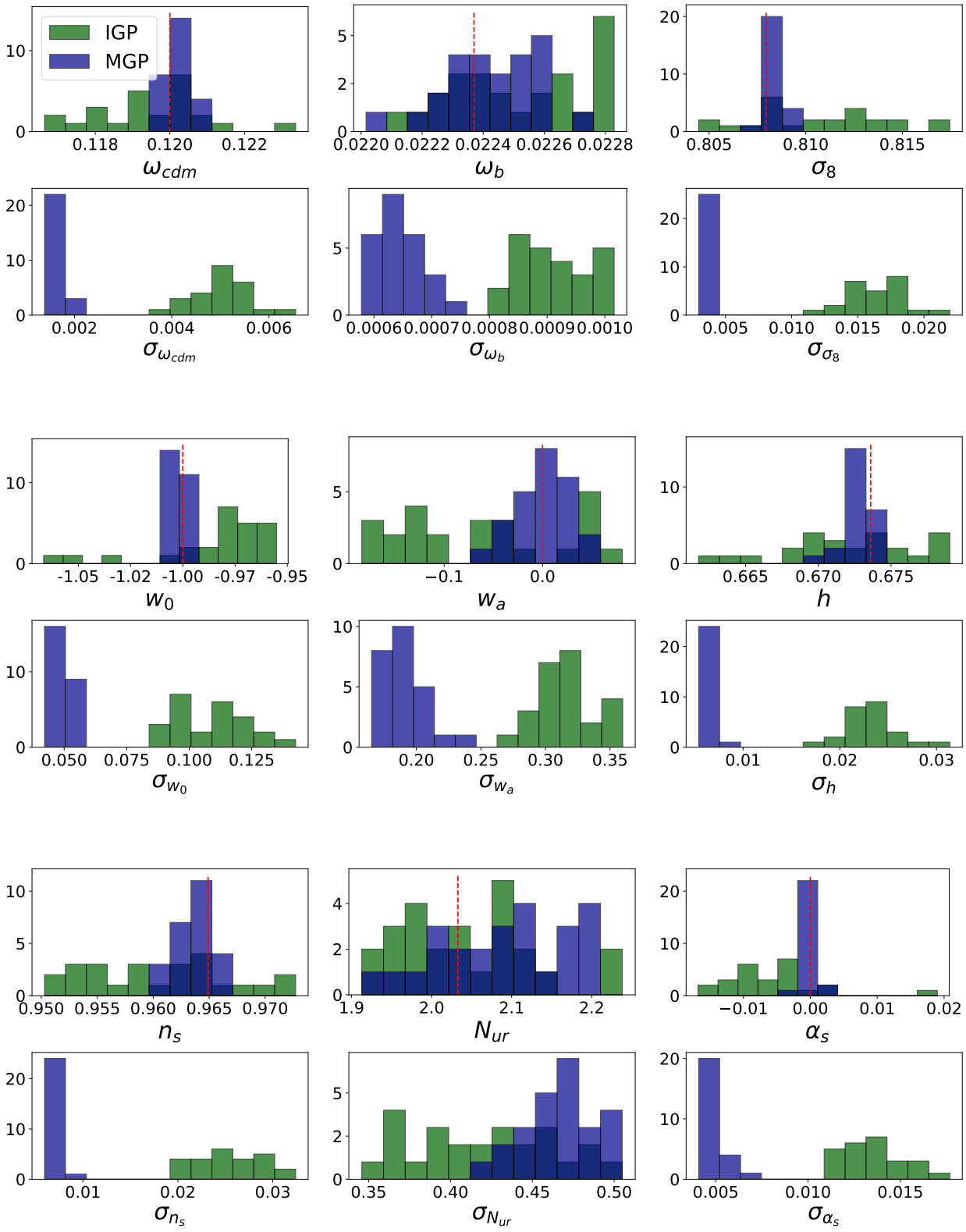


FIGURE 4.18 : Best-fit cosmological parameters along with the corresponding uncertainties on the 25 cubic mocks with Planck2018 cosmology and the median HOD. Results using the MGP and IGP models are shown in blue and green, respectively. The true parameters values are shown by a red dotted line.

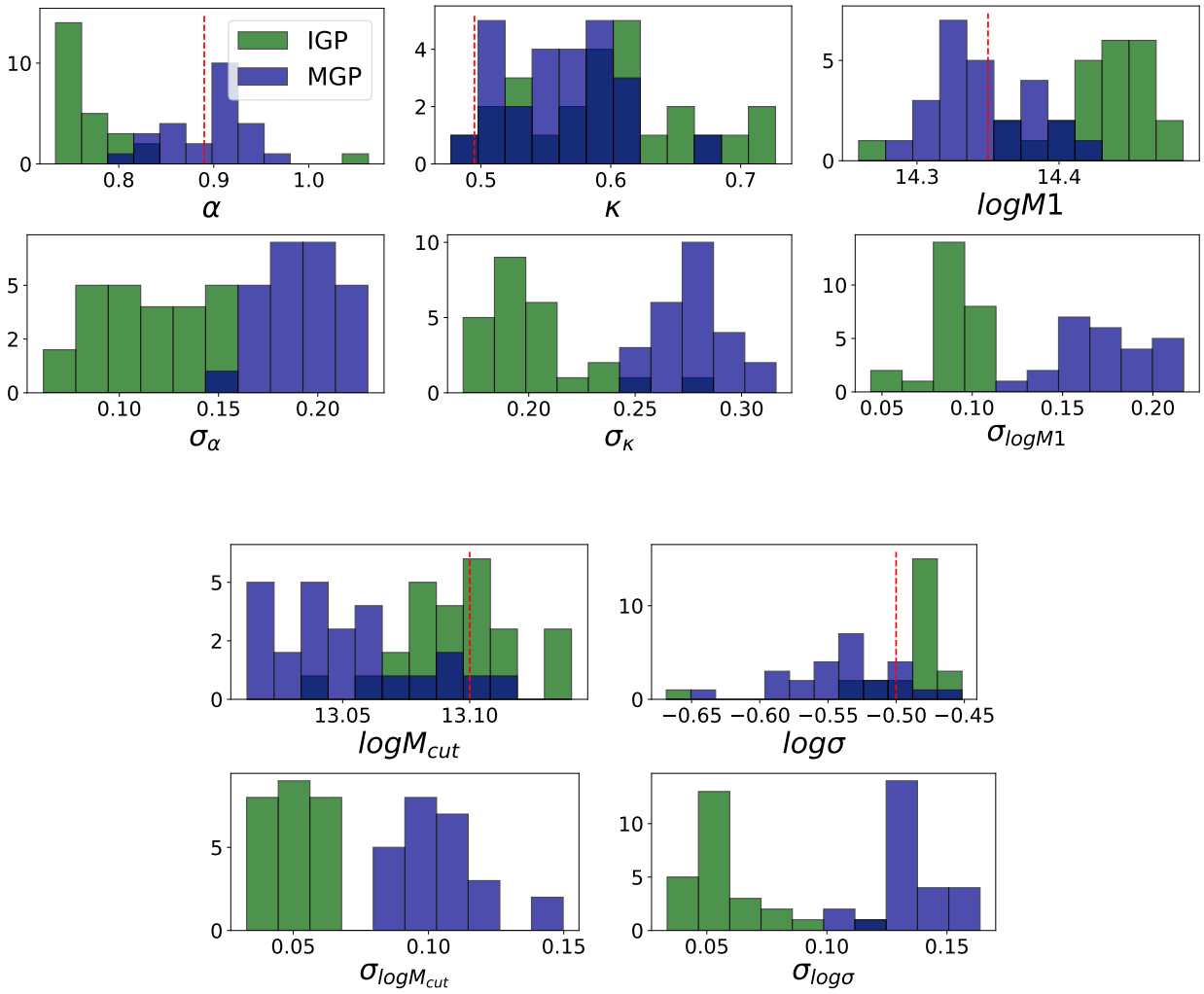


FIGURE 4.19 : Best-fit HOD parameters along with the corresponding uncertainties on the 25 cubic mocks with Planck2018 cosmology and the median HOD. Results using the MGP and IGP models are shown in blue and green, respectively. The true parameters values are shown by a red dotted line.

4.4.4 Observational effects : wide angle and Alcock-Paczynski

We validated the MGP model on cubic mocks similar to the simulations used to train the emulators and show that it performs significantly better than the standard IGP. Now we test the constraining power and accuracy of the MGP model on more realistic measurements. For each of the 25 realisations, we place the observer at the centre of the box.

In order to reproduce a survey-like geometry with a full-sky spherical footprint, we only select galaxies in the range redshifts $0.15 < z < 0.25$ from the observer, since the snapshot describes the clustering at redshift $z = 0.2$. We consider each survey to cover the full sky. We build random catalogues with similar distribution, and use the Landy-Szalay estimator of Eq. 2.85 to compute the correlation functions. The measurement of the TPCF requires the choice of a specific fiducial cosmology to convert redshifts into comoving distances. As mentioned in section 2.1.2.1, differences between the fiducial and the true cosmologies lead to AP distortions (ALCOCK et PACZYNSKI 1979) in the measured clustering. This effect is often neglected when building an emulator (YUAN et al. 2022a; ZHAI et al. 2022) as it is small on very small scales. Moreover,

modelling the AP distortion requires either one extra parameter in the training input space or a continuous model over different scales for interpolation. By construction, our MGP model is more adapted to building a continuous model than the IGP, since the MGP model can be evaluated at any scale within the training range. If the choice of fiducial cosmology is not far from the true cosmology, the AP effect can be modelled with two parameters, which scale the radial and angular separations. However, as we are only using the monopole of the TPCF, we only have access to isotropic information. We use the isotropic dilatation parameter α_V defined in Eq 2.24 to scale separations as

$$\xi_0^{\rm obs}(s) = \xi_0^{\rm mod}(\alpha_V s). \tag{4.34}$$

When running an inference, rather than adding one extra free parameter, α_V will be computed for each log-likelihood estimation following Eq 2.25. We select cosmology c177 of the AbacusSummit suite as a fiducial cosmology. Its deviation from the true one (Planck18), results in a $\sim 2\%$ variation in the isotropic dilatation parameter $\alpha_v \approx 1.02$. We measure the clustering of the 25 realisations of our survey-like mocks using both the true and c177 fiducial cosmologies, and rescale the covariance matrix $C_{\rm cosmic}$ to match the volume considered.

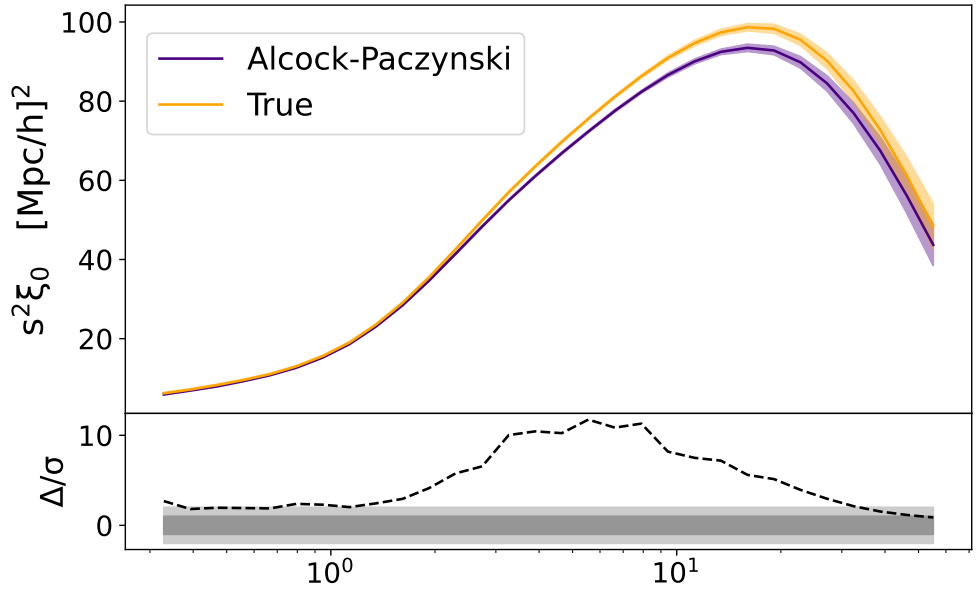


FIGURE 4.20 : AP effect on the observable. Top panel : Mean clustering with its standard deviation of the 25 mocks with Planck2018 cosmology, the median HOD, and a spherical full-sky footprint from $z = 0.15$ to $z = 0.25$. The true and AP-distorted clustering are shown in orange and indigo, respectively. The bottom panel shows the residual between both curves using the cosmic variance of the varying LoS clustering.

The average monopoles along with the residuals are shown in Figure 4.20. Although the cosmic variance is larger for the survey-like mocks compared to the cubic box measurements - from the reduced volume - the AP distortions lead to significant deviations for every scale, up to 10 σ for scales of a few h^{-1}Mpc. To assess whether ignoring the AP distortions can lead to biased cosmological constraints, we fit the 25 realisations of the AP distorted clustering separately using the MGP model with and without accounting for AP distortions, i.e., fitting for α_V or fixing α_V to 1. We refer to the former as MGP$_{\alpha_V}$. In both cases, we rescale the sample covariance matrix with

the corresponding spherical volume. We also compare the impact of ignoring the AP effect using MGP and IGP in Appendix 4.5.

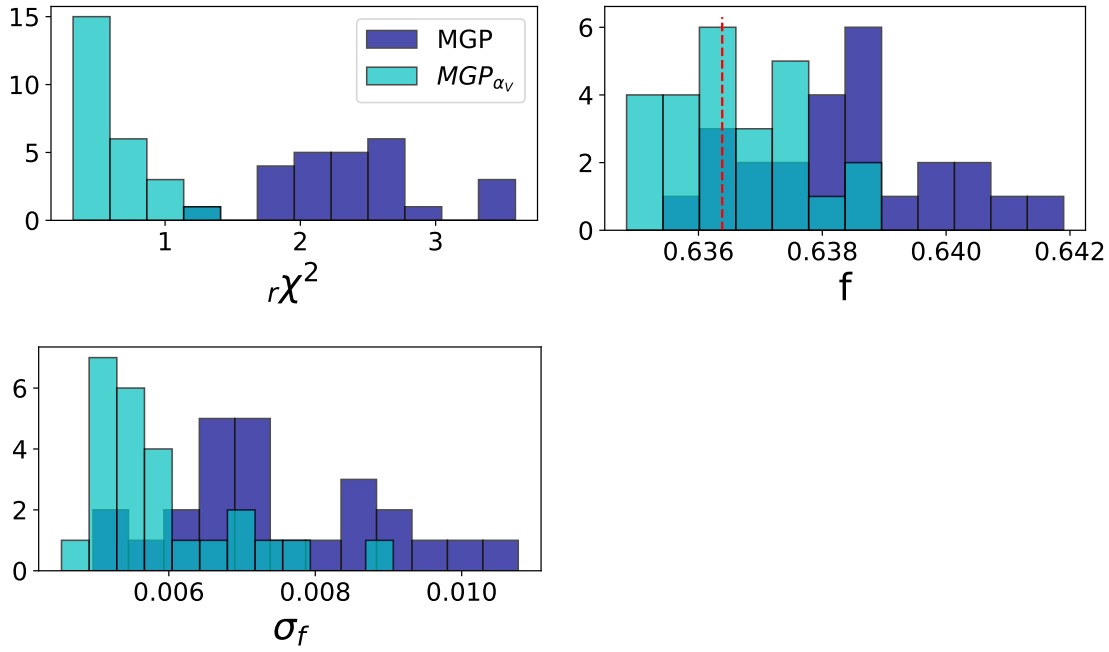


FIGURE 4.21 : Reduced chi squared, best-fit growth rate, and corresponding uncertainties resulting from the separate fits of the 25 mocks with spherical full-sky footprint from $z = 0.15$ to $z = 0.25$, AP distortion, Planck2018 cosmology, and the median HOD. Results using the MGP with and without modelling AP distortions are shown in cyan and blue, respectively. The true growth-rate value is shown by a red dotted line.

Figure 4.21 displays the distribution of reduced chi-squared, best-fit growth rate f, and their estimated uncertainties σ_f. The best-fit cosmological and HOD parameters with their uncertainties are shown in Figures 4.22 and 4.23, respectively. Table 4.3 shows the summary statistics as in Table 4.2, but now for the survey-like mocks. With a significantly larger cosmic variance, the models better fit the data, and therefore the reduced chi-squared distributions are less spread and more centred on one compared to the cubic box fits. The data points are also better described when accounting for AP distortions; the mean of the distribution is $\langle\chi^2_r\rangle = 0.55$ for MGP_{α_V} and $\langle\chi^2_r\rangle = 2.43$ for MGP for $(30-14) = 16$ degrees of freedom, corresponding to p-values of 0.9 and 10^{-3} respectively. In both cases, we recover the expected growth rate within 1σ, though ignoring the AP effect increases the bias by a factor of ten and leads to larger error bars, and also with best-fit values for f that are more scattered. We also note that compared to the previous cubic box fits, although the cosmic variance of the survey-like mocks is larger, the growth rate is more constrained in both cases. We see similar results in Figure 4.22 for the cosmological parameters : when considering AP distortion in the modelling, we reduce the bias and obtain tighter constraints. However, the Dark Energy parameters w_0 and w_a appear as an exception, where although MGP_{α_V} gives tighter constraints, the bias is larger than when using MGP. Yet, as those parameters are weakly constrained, the difference in the growth-rate measurement accuracy between the two models mainly comes from the ω_m and h parameters. For the HOD parameters showcased in figure 4.23, in both cases we have unbiased results for every parameter. However, as HOD parameters are mostly constrained at small scales, where the AP effect is less

important, the MGP_{α_V} model does not bring any substantial improvement, and even gives looser constraints than the MGP without AP modelling on average.

TABLE 4.3 : Summary statistics for the fits of the 25 mocks with spherical full-sky footprint from $z = 0.15$ to $z = 0.25$, Planck2018 cosmology, and the median HOD, where the clustering was measured using cosmology c177 of the ABACUSSUMMIT suite as fiducial cosmology resulting in an isotropic dilatation parameter $\alpha_\nu \approx 1.02$. $\langle\Delta_\mathrm{p}\rangle$ is the bias, the difference between the true and best-fit value for any parameter p, $\langle\sigma_\mathrm{p}\rangle$ is the average over the 25 realisations of the (symmetrised) one sigma confidence level, and $\mathrm{std}(\mathrm{p})$ is the standard deviation of the estimated parameter p over the 25 results. The growth rate f is a derived parameter from the other cosmological parameters.

	MGP_{α_V} $[\times 10^{-2}]$ accounting α_V			$\mathrm{MGP}[\times 10^{-2}]$ neglecting for α_V		
p	$\langle\Delta_\mathrm{p}\rangle$	$\langle\sigma_\mathrm{p}\rangle$	$\mathrm{std}(\mathrm{p})$	$\langle\Delta_\mathrm{p}\rangle$	$\langle\sigma_\mathrm{p}\rangle$	$\mathrm{std}(\mathrm{p})$
f	-0.02	0.59	0.10	-0.20	0.75	0.15
ω_m	-0.00	0.10	0.02	-0.04	0.13	0.03
ω_b	-0.00	0.04	0.00	-0.00	0.05	0.01
σ_8	-0.04	0.19	0.01	-0.04	0.23	0.02
w_0	-0.13	2.93	0.28	-0.03	3.27	0.41
w_a	0.56	11.77	0.80	0.34	12.68	1.67
h	0.02	0.42	0.08	0.15	0.52	0.09
n_s	0.01	0.45	0.07	0.01	0.50	0.08
N_ur	-1.34	30.49	2.05	-2.94	31.75	5.01
α_s	0.00	0.22	0.02	0.01	0.28	0.04
α	5.66	32.97	5.97	10.54	23.24	4.75
κ	-1.57	26.37	1.23	0.04	22.84	4.53
$\log M_1$	-0.51	32.20	3.99	-1.27	30.42	8.47
$\log M_\mathrm{cut}$	5.06	11.15	2.09	8.95	11.12	2.87
$\log\sigma$	6.96	15.94	2.92	4.74	13.97	2.56

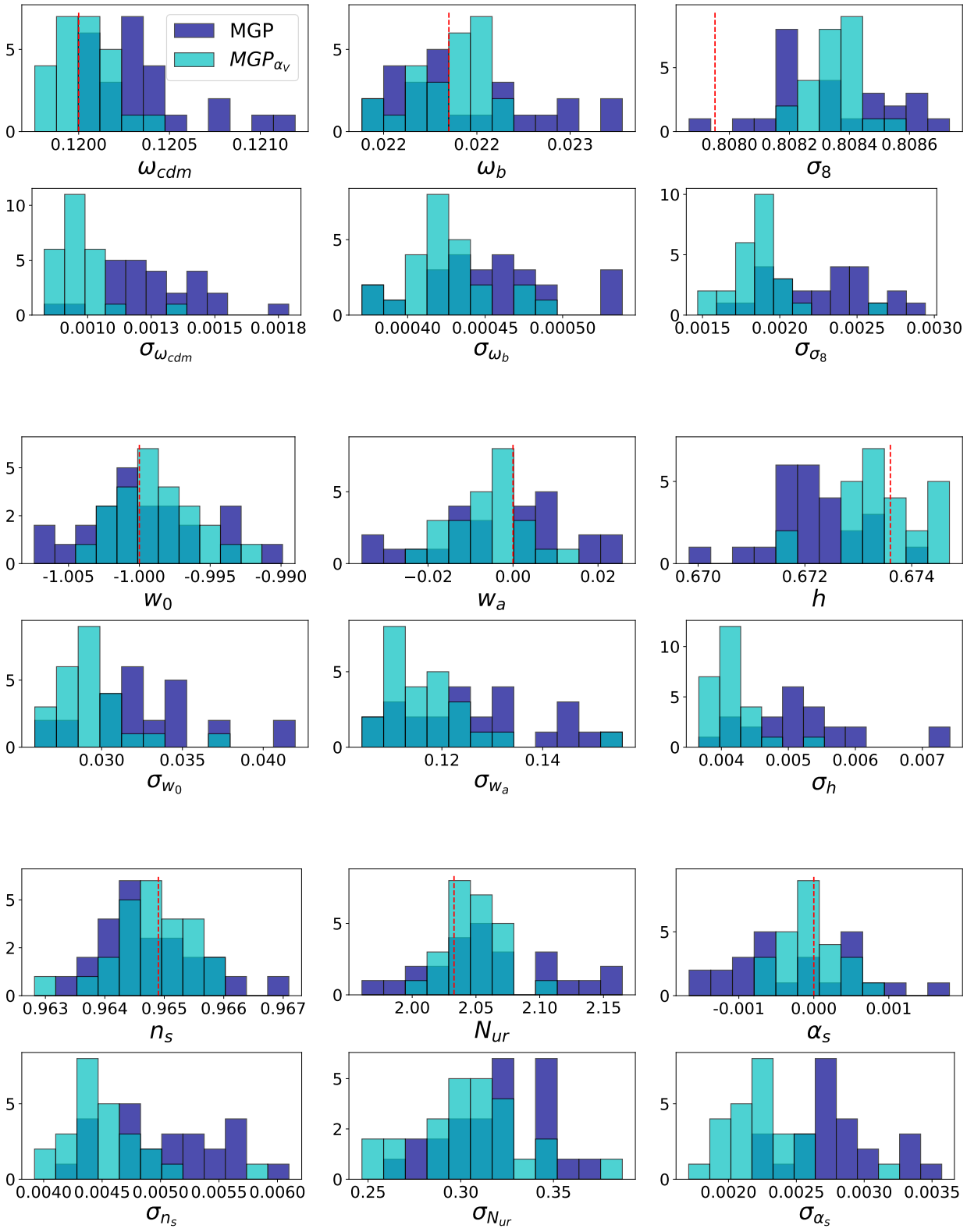


FIGURE 4.22 : Best-fit cosmological parameters along with the corresponding uncertainties on the 25 mocks with spherical full sky footprint from $z = 0.15$ to $z = 0.25$, AP distortion, Planck2018 cosmology, and the median HOD. Results using the MGP with and without modelling AP distortions are shown in cyan and blue, respectively. The true parameter values are pointed by a red dotted line.

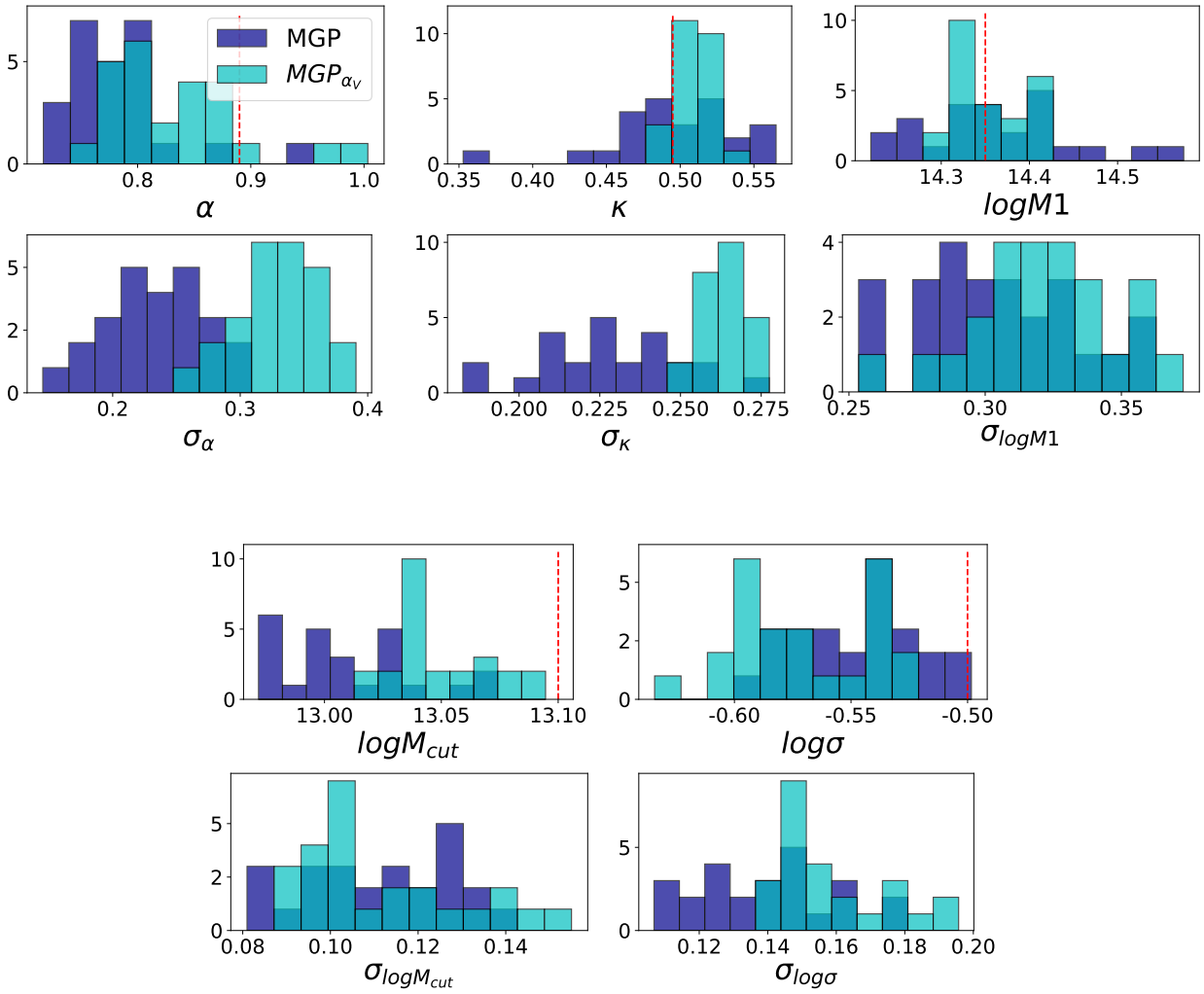


FIGURE 4.23 : Best-fit HOD parameters along with the corresponding uncertainties on the 25 mocks with spherical full-sky footprint from $z = 0.15$ to $z = 0.25$, AP distortion, Planck2018 cosmology, and the median HOD. Results using the MGP with and without modelling AP distortions are shown in cyan and blue, respectively. The true parameter values are shown by a red dotted line.

4.5 Conclusion

In this chapter, we reviewed the standard methodology for training an emulator model for galaxy clustering at non-linear scales from a suite of dark matter simulations. We introduced a new Gaussian process model that allows efficient extension of the input parameter space dimension and accounts for correlated noise. As a proof of concept for our model, we emulated the redshift space TPCF of galaxies over the input space $X = X_\Omega \otimes X_{\rm HOD} \otimes X_{\rm s}$, allowing additional interpolation over a continuous range of scales and making use of the well-known correlations between them. We implemented our multiscale Gaussian process model (MGP) as well as the standard approach, which consists of constructing independent Gaussian processes (IGPs) for each separation. We used 88 cosmologies, 600 HODs, and 30 separation scales as a training set, all measured in the AbacusSummit suite of n-body simulations. The input parameter space of the resulting models is composed of nine cosmological parameters, five HOD parameters, and the set of separations

of the TPCF.

After validating the predictive accuracy of our trained model on a test set with six cosmologies and 20 HODs, we ran inferences using MCMC to assess the robustness and constraining power of our models. In a high-precision setting, where we used the average TPCF of 25 realisations, ignoring the predicted uncertainties of our GP models can lead to biased results, especially with the IGP model. We also demonstrate that both models are able to recover the expected HOD parameters. Marginalising over such parameters, we are able to obtain unbiased cosmological constraints. Our MGP model returns more accurate and precise constraints than the IGP on most of the cosmological parameters. Moreover, we show that Alcock-Paczynski distortions can become significant — compared to the cosmic variance — at intermediate scales for the redshifts we are testing. Neglecting this effect in an emulator model, as has been done in previous works, contributes to loosening the constraints and increasing the bias on cosmological parameters.

The parameter space of such emulator models could also be extended to higher dimensions to be able to interpolate and marginalise over systematic quantities, extended HOD models, selection effects, different redshift bins, and more. There remain some drawbacks to using a training set with many dimensions. First, building the training set is expensive. Second, although a Kronecker GP model with a very large number of training points can be trained relatively quickly, the predictions of the model covariance can be slow. It is also important to reiterate that for galaxies at lower redshift, building a model requires more realistic simulations that reproduce redshift evolution (light cone) to properly account for wild angle effects. Such a model would also need to be validated on realistic mocks using different galaxy-halo connections before application to real measurements. Our MGP model could be used in future works to emulate a larger set of galaxy clustering observables and their cross correlations, such as galaxy and peculiar velocity auto- and cross-correlations. Accurate models of these correlations are essential in the era of high-precision measurements from surveys such as DESI or Euclid.

Appendix A : Emulator covariance choice

We discuss the choice of an appropriate covariance matrix associated with the emulator uncertainties. When the emulator model does not give an uncertainty associated with the prediction, the covariance matrix used to account for the modelling systematics is usually derived from the test set and decomposed into two terms

$$C_{\text{test}} = C_{\text{perf}} + C_{\text{cosmic}}. \tag{4.35}$$

Where C_{cosmic} (shown in figure 4.6) accounts for the training set cosmic variance and C_{perf}, accounting for the emulator performance to predict the test set, is computed as

$$C_{\text{perf}} = \frac{1}{N_{\text{test}} - 1} \sum_{k=1}^{N_{\text{test}}} (\Delta_k - \overline{\Delta})(\Delta_k - \overline{\Delta})^\top, \tag{4.36}$$

where Δ_k denotes the difference between the model prediction and the k-th element of the test set, and $\overline{\Delta}$ is averaged over the complete test set. While it does not depend on the parameters $\boldsymbol{\theta}$, the model covariance C_{test} is analogue to C_{emu}. When performing inferences on external data, it must be added to the measurement uncertainty as previously $C_{\text{tot}} = C_{\text{data}} + C_{\text{test}}$.

We compare these two choices of model covariance by fitting the average stack of the 25 correlation functions with Planck 2018 cosmology and the median HOD. Figure 4.24 shows the Posterior distributions of the cosmological and HOD parameters obtained. We find that the different methods are consistent, and all parameters are well recovered within one sigma. However, we see that using the parameter dependent $C_{\text{emu}}(\theta_\Omega, \theta_{hod})$ leads to tighter constraints on the cosmological parameters and looser constraints for most of the HOD parameters. Indeed, C_{test} accounts twice for the cosmic variance as the test set is not the ground truth. This overestimation of the covariance is more important on the large scales, thus, does not affect the HOD parameters, mostly constrained by the smallest scales.

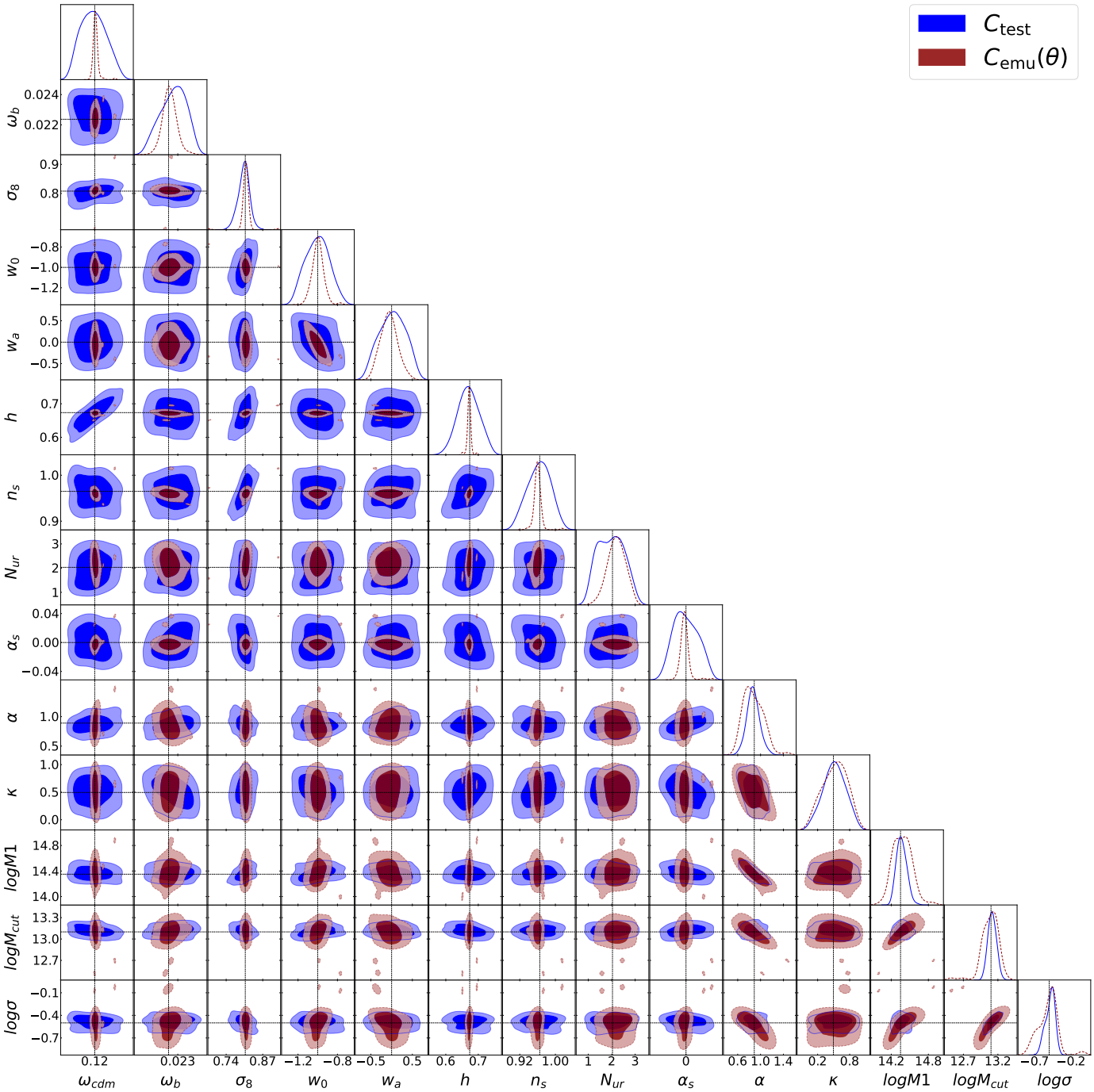


FIGURE 4.24 : Posterior distribution of the cosmological and HOD parameters obtained after running MCMC inference with the MGP model on the average stack of the 25 correlation functions with Planck 2018 cosmology and the median HOD. The results obtained using the covariance derived from the test set are shown in blue. The results obtained using the emulator covariance evaluated for each sampling point are shown in red. The true parameters are pointed by the black dotted lines.

TABLE 4.4 : Summary statistics for the fits of the 25 mocks with a spherical full-sky footprint from $z = 0.15$ to $z = 0.25$, Planck2018 cosmology, and the median HOD. $\langle \Delta_p \rangle$ is the bias, the difference between the true and best-fit value for any parameter p, $\langle \sigma_p \rangle$ is the average over the 25 realisations of the (symmetrised) one sigma confidence level, and std(p) is the standard deviation of the estimated parameter p over the 25 results.

$(10^2\times)$	IGP			MGP		
p	$\langle \Delta_p \rangle$	$\langle \sigma_p \rangle$	std(p)	$\langle \Delta_p \rangle$	$\langle \sigma_p \rangle$	std(p)
f	-0.79	2.05	0.52	-0.20	0.75	0.15
ω_m	-0.18	0.74	0.19	-0.04	0.13	0.03
ω_b	-0.01	0.11	0.02	-0.00	0.05	0.01
σ_8	2.25	1.97	0.70	-0.04	0.23	0.02
w_0	-2.03	14.14	3.17	-0.03	3.27	0.41
w_a	3.25	37.54	8.42	0.34	12.68	1.67
h	0.71	3.04	0.76	0.15	0.52	0.09
n_s	-0.01	3.2	0.72	0.01	0.50	0.08
N_{ur}	-2.46	57.32	13.49	-2.94	31.75	5.01
α_s	0.40	1.84	0.86	0.01	0.28	0.04
α	6.02	15.82	5.15	10.54	23.24	4.75
κ	-7.31	29.72	4.8	0.04	22.84	4.53
$\log M_1$	0.66	12.0	4.63	-1.27	30.42	8.47
$\log M_{cut}$	4.99	6.89	2.43	8.95	11.12	2.87
$\log\sigma$	2.33	9.27	2.96	4.74	13.97	2.56

Appendix B : AP effect on IGP fits

We show in section 4.4.4 that ignoring the AP distortions can lead to a loss of accuracy and precision, especially for the two cosmological parameters of interest, f and σ_8. Here, we compare the impact of ignoring the AP distortions using the MGP and IGP models. Note that as the IGP is not able to interpolate over scales, the AP effect is usually ignored in standard analysis (YUAN et al. 2022a; ZHAI et al. 2022; KWAN et al. 2023). While the AP effect could also be modelled using the IGP, this would require that an extra approximated interpolation be performed between the trained scales, both for model and uncertainty estimates. We fit the 25 realisations of the AP distorted clustering separately using both models without modelling the AP distortions. In both cases, we rescale the sample covariance matrix with the corresponding spherical volume. The best-fit values for f and σ_8 along with their estimated uncertainties are described in figure 4.21. The best-fit cosmological and HOD parameters with their uncertainties are shown in figure 4.25. Table 4.4 shows summary statistics for all cosmological and HOD parameters. We see that when ignoring the AP distortions, although not the case for HOD parameters (the very small scales remain unaffected by the AP effect), the MGP model performs significantly better than the IGP for every cosmological parameter. The IGP model in particular gives constraints on σ_8 that are biased by more than 1σ.

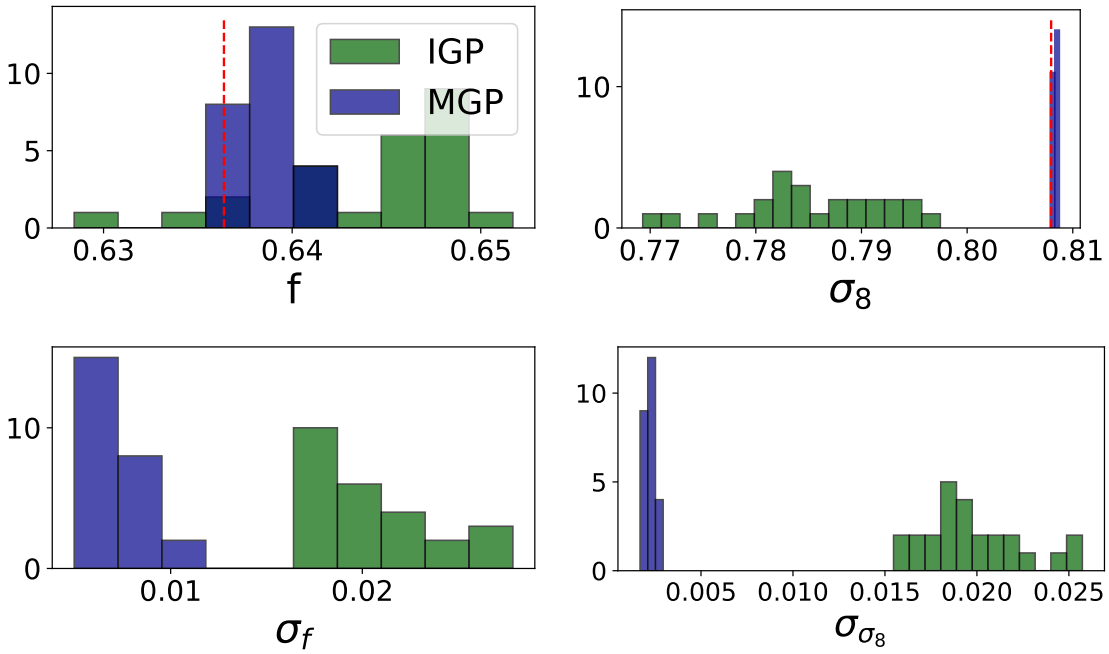


FIGURE 4.25 : Best-fit growth rate and corresponding uncertainties resulting from the separate fits of the 25 mocks with spherical full-sky footprint from $z = 0.15$ to $z = 0.25$, AP distortion, Planck2018 cosmology, and the median HOD. Results using the MGP and IGP models are shown in blue and green, respectively. The true parameter values are shown by the red dotted line.

5 Combining non-linear peculiar velocity and galaxy clustering

Summary

5.1 Peculiar velocity and galaxy clustering from N-body simulations . . . 150
5.1.1 Observable . . . 150
5.1.2 Realistic peculiar velocity catalogue . . . 152
5.2 Building the emulators . . . 156
5.2.1 Multiscale Gaussian process . . . 156
5.2.2 Prediction accuracy and precision . . . 157
5.3 Parameters recovery . . . 160
5.3.1 Inference strategy . . . 160
5.3.2 Results . . . 160
5.4 Inference on noisy measurements . . . 164
5.4.1 Constraining power . . . 164
5.4.2 Varying the minimum fitting scale . . . 166
5.5 Conclusion . . . 168
Appendix A : Emulator dynamical range . . . 170
Appendix B : Covariance template scaling . . . 174
Appendix C : Recovery mocks best-fitted parameters . . . 176
Appendix D : Realistic mocks best-fitted parameters . . . 177

The distribution of galaxies traces the non-linear and non-Gaussian matter field. This deviation from linearity and Gaussianity is enhanced at small scales and low redshift. As we are mapping tens of millions of galaxies (A. LAUREIJS et ARDUINI 2011; E. COLLABORATION et al. 2022; DESI COLLABORATION et al. 2022), the constraining power on the cosmological model is no longer statistically limited, but restricted by the use of our standard observables. Addressing this challenge requires to :

1. Model the non-linear clustering to extract additional information from the small scales.

2. Go beyond the information contained in the two point statistics.

We have seen that the first point can be achieved through SBI, while modelling of the non-linear scales is still limited by the resolution of the simulations, and the complexity of baryonic physics driving the galaxy formation process. Furthermore, since two-point statistics do not hold the full cosmological information contained in the non-linear matter distribution (section 2.1.4), alternative methods introducing summary statistics beyond the standard observables have been developed. A promising approach, known as field-level inference (STADLER et al. 2023), consists in directly fitting the full distribution of the perturbed fields, without using any compressed statistics and marginalising over the initial conditions. Recently, different alternative methods based on various statistics such as bispectrum analysis (YANKELEVICH et PORCIANI 2019; AGARWAL et al. 2021), wavelet scattering transform (VALOGIANNIS et DVORKIN 2022), and k-th nearest neighbour statistics (YUAN et al. 2023), have been developed to extract information beyond the standard two-point function. Additional information can also be extracted from the standard two-point statistics by cross correlating different probes of the matter filed in an overlapping volume such as cosmic voids, weak lensing, galaxy multi-tracer analysis (Y. WANG et G.-B. ZHAO 2020), density splits (PAILLAS et al. 2023), or direct measurements of the galaxy peculiar velocities. As it was demonstrated in some recent work (ADAMS et BLAKE 2017; TURNER et al. 2023), illustrated on figure 2.3, the joint analysis of galaxy and velocity clustering at low redshift increases the constraints on growth rate measurements, leveraging the reduction of cosmic variance. While scales typically above 50 to 60 h^{-1}Mpc are well modelled by linear theory, small scales - sensitive to cosmology and with high SNR - are challenging to model analytically (Figure 2.4).

In this Chapter we apply our emulation framework introduced in Chapter 4 to the 3×2 points correlations function estimators - derived in Chapter 2 - to jointly describe the clustering of galaxies and peculiar velocities in redshift space. We use the ABACUSSUMMIT simulation suite to train our models, and build realistic mock catalogues to assess the constraining power of the 3×2 points analysis. This chapter is organised as follows. In section 5.1, we present the observables to emulate, and the different data sets that we will be using. In section 5.2, we describe the procedure for training the different models and validate their accuracy and precision. In section 5.3, we assess the parameters' recovery when running statistical inference. Finally, in section 5.4, we explore the constraining power of the joint analysis on a set of realistic mocks.

5.1 Peculiar velocity and galaxy clustering from N-body simulations

We follow the same methodology used in the previous chapter, using the AbacusSummit simulations combined with the HOD formalism to generate galaxy samples and train Multiscale Gaussian Process (MGP) models for multipoles of the galaxy-galaxy, velocity-velocity and velocity-galaxy correlation functions in redshift space.

5.1.1 Observable

We exploit the same redshift space galaxy catalogues described in the previous chapter to define our train and test sets (see table 4.1). Once again, because we are working with cubic boxes with periodic boundary conditions, we use Peebles estimators under the flat sky approximation (Eq 2.96) with analytical random-random pair counting, to quickly measure ξ_{gg}, ξ_{vv} and ξ_{vg} on 30 logarithmic separation bins going from 0.3 h^{-1}Mpc to 60 h^{-1}Mpc. For a given cosmology and HOD, every measurement is performed five times to reduce shot noise. Once measured, the TPCF are decomposed into multipoles following Eq 4.5. In order to balance between compression and loss of information, we only use the monopole and quadrupole for both ξ_{gg} and ξ_{vv} and the dipole for ξ_{vg}. Note that while ξ_{vg} is neither an odd nor even function of μ, we find that the dipole gives the stronger signal. We verified that the flat sky approximation does not yield biased measurements of the clustering. From the 25 Planck2018 boxes with the median HOD, we measured our five multipoles [1] both using the Peebles estimators on the full boxes(similar to our baseline to build train and test set), and using the Landy-Szalay-like estimators (Eqs 2.85,2.92,2.93) in a spherical footprint with a radial cut corresponding to a redshift range of $z \in [0;0.1]$. In the following, we refer to the latest measurement as the recovery set.

In figure 5.1 the solid lines show the expected mean clustering of the recovery set, with the full-sky spherical footprint. The dashed black lines show the mean clustering of the 25 boxes computed using the fixed LoS Peebles estimator. We see that for the scales considered here, the deviation from the expected value of the flat-sky approximation measurements is significantly smaller than cosmic variance.

Although the flat-sky approximation enable to use the full volume of the cubic boxes, the train and test cosmology boxes are run with the same IC. Hence, our measurements are subject to cosmic variance. As previously, we build the covariance matrix $C_{\rm cosmic}$ from the 1400 small boxes for Planck2018 cosmology with the median HOD, using the same estimators. We then rescale the covariance to match the volume of the base boxes. Figure 5.2 gives the corresponding correlation matrix between the different observables for our selected separations bins. We find strong correlation for the velocity statistics, over a wide range of scales, especially for the velocity autocorrelation monopole leading to smooth variation of the signal.

[1] Pair counts are performed using the package pycorr https://github.com/cosmodesi/pycorr

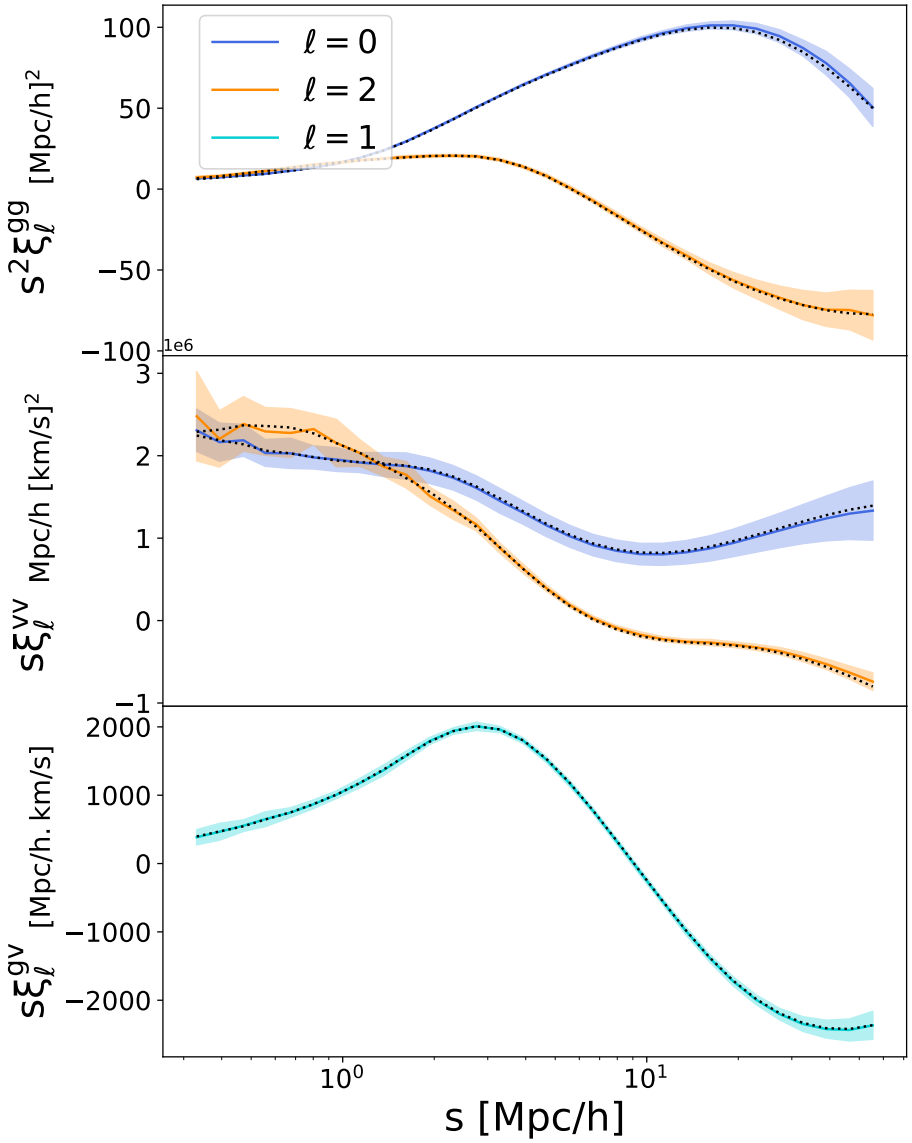


FIGURE 5.1 : Comparison between fixed and varying LoS estimators using of the 25 Planck2018 boxes with the median HOD. The solid lines show the expected mean clustering of the 25 boxes computed using Landy-Szalay-like estimators with a full-sky spherical footprint in a redshift range of $z \in [0; 0.1]$. The shaded area are the corresponding standard deviation of the realisations. The dashed black lines show the mean of the flat-sky approximation clustering.

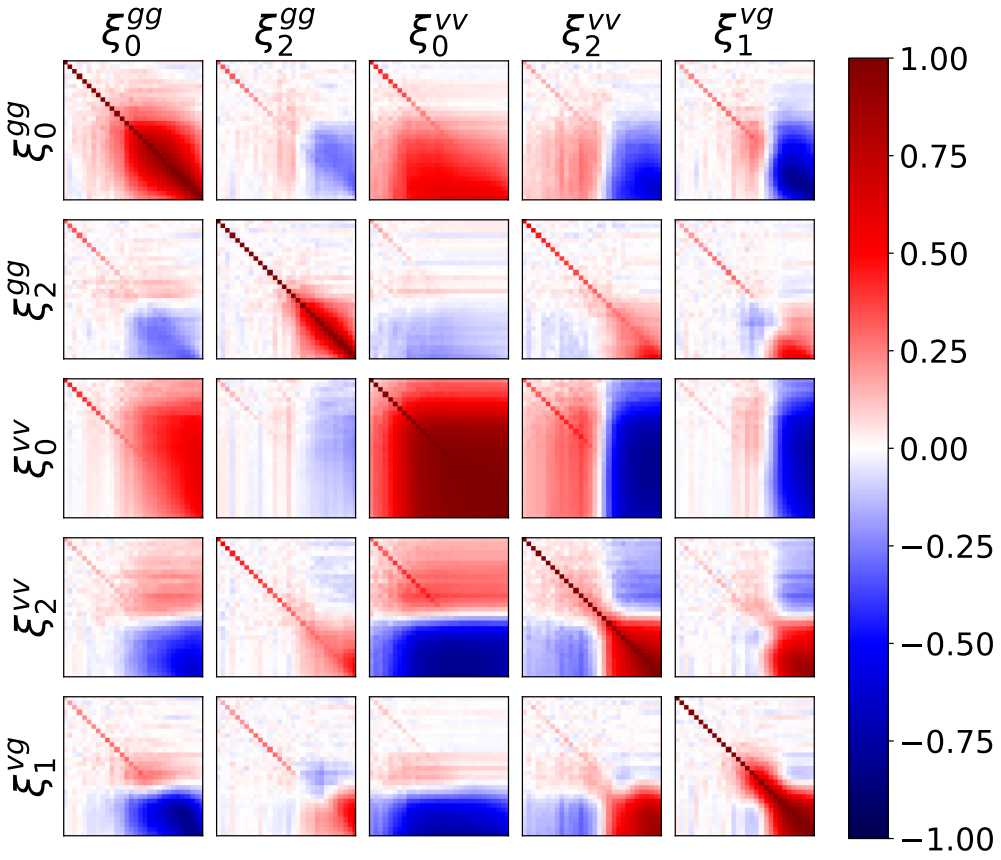


FIGURE 5.2 : Correlation matrix computed from 1400 small boxes realisations of the Planck2018 cosmology, using the median HOD. The different panels show the auto and cross correlation between the TPCF multipoles of ξ_{gg} , ξ_{vv} and ξ_{gv}. Each of the subplot axis correspond to the scale s with a logarithmic binning going from 0.3 h^{-1}Mpc to 60 h^{-1}Mpc.

5.1.2 Realistic peculiar velocity catalogue

In every TPCF measurement considered so far, to build our train/test/recovery sets, every galaxy position was associated with a radial peculiar velocity. Yet it is well known that measuring peculiar velocities in real data is a more involved process, requiring both an independent distance and redshift measurements, with their associated uncertainties. Therefore, peculiar velocity surveys are typically less complete than galaxy surveys. In section 2.2, we pointed out the main techniques to measure peculiar velocity in large scale survey. Combined with spectroscopic redshift measurements, the upcoming Rubin Observatory and Zwicky Transient Facility (ZTF) measurements (BELLM et al. 2019), observations of type Ia Supernovae (SnIa) will provide precise peculiar velocity catalogues at low redshift. Tully-Fisher(TF) and Fundamental Plane (FP) (CAMPBELL et al. 2014; HONG et al. 2019; HOWLETT et al. 2022; SAULDER et al. 2023a), velocity catalogues provide additional distance estimates, with larger uncertainties but in greater number. We want to create a set of 'realistic' mocks to study the impact of distance uncertainties and number densities of peculiar velocity tracers on the cosmological constraints.

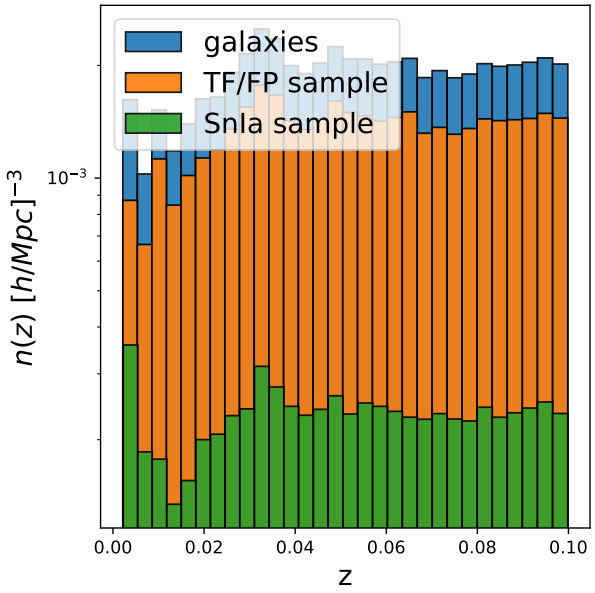


FIGURE 5.3 : Redshift density distribution of the galaxy, TF/FP velocity, and SnIa velocity samples in one realisation of our realistic mocks set. The samples are randomly selected from the complete galaxy catalogue of one of the 25 Planck2018 cosmology boxes, populated using the realistic HOD.

We use the 25 boxes with Planck2018 cosmology to create three distinct catalogues : a galaxy sample tracing the density field only, a joint TF and FP velocity sample (similar uncertainties), and a separate SnIa velocity sample. We consider the SnIa and TF/FP samples as a single velocity catalogue hereafter. We choose an HOD that roughly matches the expected densities $n(z)$ within $z \in [0, 0.1]$. We get a galaxy density around $\bar{n}(z) = 2 \times 10^{-3} h^{-3}/\mathrm{Mpc}^3$ (corresponding to a subsample of the DESI Bright Galaxy Survey). From this sample, we randomly select galaxies to get a TF/FP sample with a density around $\bar{n}(z) = 10^{-3} h^{-3}/\mathrm{Mpc}^3$, and a SnIa sample with a density around $\bar{n}(z) = 10^{-4} h^{-3}/\mathrm{Mpc}^3$. The corresponding HOD parameters used to populate the galaxy sample are $\alpha = 1.3$, $\kappa = 0.8$, $\log_{10} M_1 = 13.7$, $\log_{10} M_{\mathrm{cut}} = 12.6$ and $\log_{10} \sigma = -0.35$. Moreover, we chose a fiducial cosmology(c003 in AbacusSummit) different from the true one to convert redshifts into comoving distances, inducing Alcock-Paczynski (AP) distortions (Alcock et Paczynski 1979) in the measured clustering. Figure 5.3 show the $n(z)$ radial distribution obtained for one realisation of our realistic mocks set.

In section 2.2 we described how peculiar velocities are derived from log-distance ratio measurement η following Eqs 2.74, 2.75. We use one unique type of conversion factor $\alpha(z_{\mathrm{obs}})$, following the approximation $\eta \approx \eta_L \approx \eta_A$. Because of the intrinsic scatter of standardisable objects, the log-distance ratio measurements have a typical error $\sigma_\eta = \left(f_{\mathrm{err}}/\ln 10\right)$ with f_{err} around 20 % for TF/FP, and 7% for SnIa (Boruah et al. 2020). After applying the redshift cut, the peculiar velocities are converted to log distance ratio. A Gaussian random error is then applied on η with zero mean and a standard deviation corresponding to a f_{err} of 7% and 20 % for the SnIa and TF/FP samples respectively. Gaussian errors on η, yield a log-normal distribution for the peculiar velocity errors. This uncorrelated noise is an approximation. In real analyses, the distances errors might be correlated due to calibration. Figure 5.4 illustrates this effect for one of the 25 mocks, where the black dashed histograms show the true v and η distribution where the SnIa and TF/FP samples are depicted in orange and green respectively, compared to the true distribution in black.

We see that the velocity measurements are strongly dominated by the noise and that the peculiar velocity distribution becomes asymmetric.

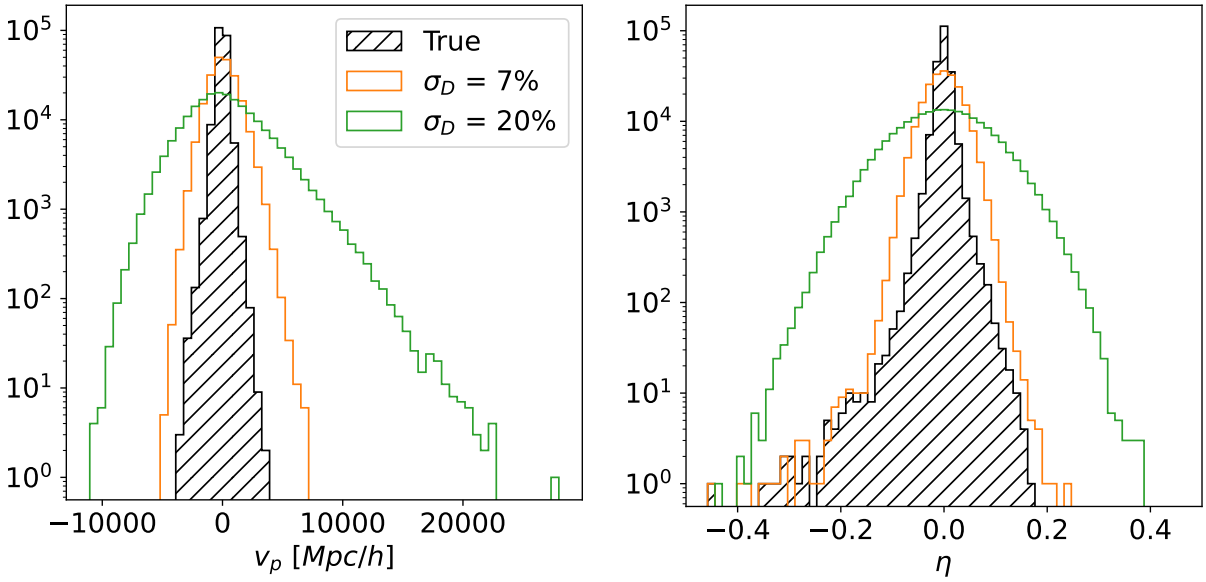


FIGURE 5.4 : Velocity and log-distance ratio distribution from one mock. The black dashed histograms show the true v and η distribution while the simulated SnIa and TF/FP samples are depicted in orange and green respectively.

We use the FKP weights defined in Eqs 2.79 and 2.80 to avoid biasing the clustering measurements, chosing $P_g = 10^4\ h^{-3}\text{Mpc}^3$ and $P_v = 10^9\ h^{-3}\text{Mpc}^3\ \text{km}^2\text{s}^{-2}$. However, the weights w_v consider Gaussian errors on the velocities. Since the Gaussian errors are applied to η, we can define some weights for the log distance ratio from the approximation $\eta = \alpha(z)v$ (Eq 2.74) :

$$w_\eta = \frac{1}{\alpha^2(z)\ P_v\ \bar{n}(z) + \sigma_\eta^2}, \tag{5.1}$$

and use $\left(\eta_i\, w_{\eta,i}/\alpha_i\right)$ as a proxy for the weighted peculiar velocities v_i in the pair-counts. Note that $\alpha(z)$ has nothing to do with the HOD parameter α. We test our estimators by measuring the multipoles of ξ^{vv} and ξ^{vg} on our 25 mocks using the weighted noisy measurements, and using the true peculiar velocities without any noise $\sigma_v = 0$.

Figure 5.5 compares the two measurements, with the average of the 25 "true" clustering shown in orange and our estimator shown in blue. The shaded area is the standard deviation of the 25 mocks. We find that our estimator deviation from the noise-free measurements is always well within one sigma. While the noisy measurements increase the variance of the signal on small scales (as expected), ξ^{vv} multipoles variance remains unaffected for the largest scales. We further build a covariance matrix for our realistic mocks applying the same procedure to the 1400 small boxes, duplicating them to get the proper volume corresponding to the redshift range used.

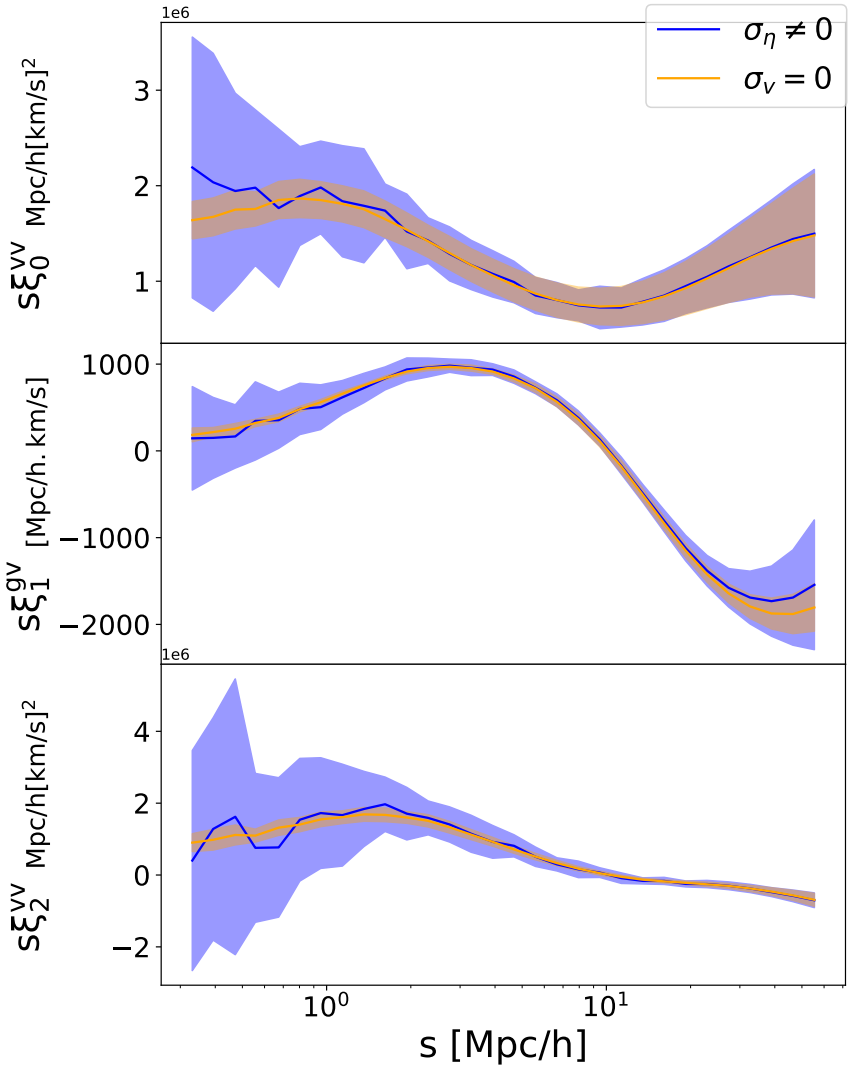


FIGURE 5.5 : Comparison of the noisy estimator with the noise-free measurements for peculiar velocity clustering using our set of 25 realistic mocks. The average of the 25 "true" clustering is shown in orange and our estimator is shown in blue. The shaded area is the standard deviation of the 25 mocks.

We summarise the different clustering measurements that will be used in the following sections :

- Training set : $N_{\text{cosmo}} \times N_{\text{HOD}} = 88 \times 600$ different clustering with fixed ICs. Used to train the emulators.
- Test set : $N_{\text{cosmo}} \times N_{\text{HOD}} = 6 \times 20$ different clustering with fixed ICs. Used to assess the emulators' performance.
- Recovery set : 25 realisations of the Planck2018 cosmology with the median HOD and a spherical full sky footprint for a redshift range $z \in [0, 0.1]$. Used to run MCMC and validate the recovery on cosmological and HOD parameters using the different models.
- Realistic mocks : 25 realisations of the Planck2018 cosmology with Gaussian log-distance

ratio errors and a HOD matching the expected $n(z)$.Full sky footprint for a redshift range $z \in [0, 0.1]$ and AP effect. Used to test the constraining power of the emulator.

- Small Boxes : 1400 realisations of the Planck2018 cosmology. Used to build covariance matrices for the training set and the realistic mocks.

5.2 Building the emulators

We want to learn, for each multipole, the mapping function between the input X_Ω, $X_{\rm HOD}$, X_s and the corresponding output ξ_ℓ^{ab} using our training set. It is important to bear in mind that this mapping function is dependent of the numerical treatment of the underlying physical process (gravity solver, time discretisation, halo finder, frozen IC,HOD shot noise etc.) and is therefore tuned to the specificities of the simulation suite. One of the main challenge of SBI is to correctly consider all these sources of errors when performing predictions. Gaussian Process formalism allows to estimate the uncertainties in the training set, and to get both a mean prediction for some new input values and the model covariance between the newly predicted points. In the following we describe the emulation procedure and discuss the performance of the trained models for the different observables.

5.2.1 Multiscale Gaussian process

We build independent emulators for our five statistics $\left[\xi_0^{gg}, \xi_2^{gg}, \xi_0^{vv}, \xi_2^{vv}, \xi_1^{vg}\right]$ using the MGP process algorithm described in chapter 4.The grid sampling of the parameter space allows use to decompose the signal and noise kernels as Kronecker products

$$K = K_\Omega \otimes K_{\rm HOD} \otimes K_{\rm s}, \quad \text{and} \quad N = N_\Omega \otimes N_{\rm HOD} \otimes N_{\rm s}, \tag{5.2}$$

where $K_\Omega, K_{\rm HOD}$ and $K_{\rm s}$ are the kernels used to respectively model the cosmological, HOD parameters and scale dependence of the signal. The kernels $N_\Omega, N_{\rm HOD}$ and $N_{\rm s}$ are analogously used to describe the parameter dependence of the noise in the training set. We fix $N_{\rm s}$ to be equal to $C_{\rm cosmic}$ described previously in figure 5.2, and for the cosmological and HOD noise kernels we chose diagonal heteroscedastic kernels allowing to rescale the variance for every HOD and cosmological configuration. This leads to $88 + 600 = 688$ noise hyperparameters for every multipole. For the signal kernels, we use standard squared exponential kernels defined in Eq 4.16 (as we found that more sophisticated kernels do not increase the emulator performance), leading to 16 signal hyperparameters composed of 9 length-scales for the cosmological parameters, 5 for the HOD parameters, one for the separations and one overall variance of the signal. We take the logarithm of the separation $\log_{10} X_s$ to train our emulators over a more smoothly varying signal. To avoid overflow, we further normalise each input dimension X_Ω, $X_{\rm HOD}$ and $\log_{10} X_s$ between 0 and 1. To reduce the dynamical range of the output multipole, we apply a bi-symmetric log transformation (Webber 2012) suited for datasets with both negative and positive values, defined as

$$\text{slog}_S(x) = \text{sign}(x) \cdot \log_{10}(1 + |(x/S)|), \tag{5.3}$$

where the constant S adjusts the slope near the origin. We then normalise the output of our training sets to be centred on zero and have a standard deviation of one. The S values are chosen to fine tune the model. We set it to 10^1 for ξ_0^{gg} and ξ_2^{gg}, 10^3 for ξ_0^{vv} and ξ_1^{vg}, and 10^5 for ξ_2^{gg}. The

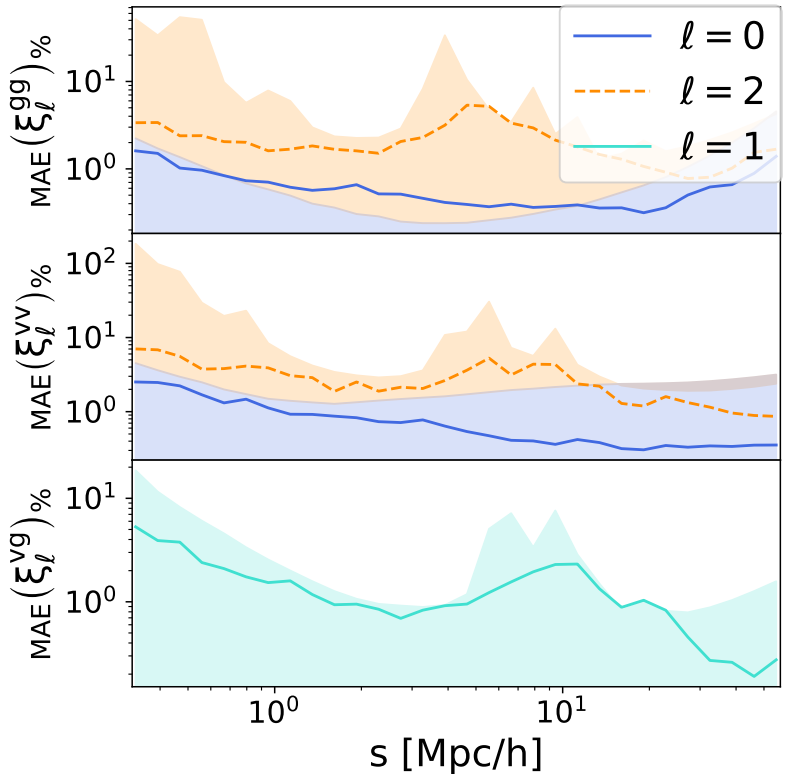


FIGURE 5.6 : Median Absolute Error (MAE) evaluated of the complete test set. The blue and orange solid lines in the top (middle) panel show the performance of the ξ_0^{gg} and ξ_2^{gg} (ξ_0^{vv} and ξ_2^{vv}) models. The turquoise solid line in the bottom panel show the performance for ξ_1^{vg}. The shaded area estimate the expected accuracy on the test set as the median of $|\sigma/\xi^{\text{test}}|$.

covariance C_{cosmic} used as the N_s kernel is transformed and normalised as well following the same steps. Once the input and output of the train set are normalised, the hyperparameters are optimised using the SCIPY gradient based algorithm LBFGS with 10 random initialisations. The dynamical range of the resulting models is depicted in appendix 5.5.

5.2.2 Prediction accuracy and precision

For every prediction, the mean and covariance are transformed back into physical space. We first test our models' accuracy on the complete test set by measuring the Median Absolute Error (MAE) for the five multipoles, defined as $\text{MAE}(\xi) = \text{median}\left|\left(\xi^{\text{emu}} - \xi^{\text{test}}\right)/\xi^{\text{test}}\right|$. Note that every multipole with $\ell \neq 0$ crosses 0 on scales around $10\,h^{-1}$ Mpc and is therefore locally ill-defined. Figure 5.6 presents the MAE in percent for each multipole. The blue solid and orange dashed lines in the top panel show the performance of the ξ_0^{gg} and ξ_2^{gg} models. The same colours are used for ξ_0^{vv} and ξ_2^{vv} in the middle panel. The turquoise solid line show the performance for ξ_1^{vg} in the bottom panel. The shaded areas describe the expected accuracy on the test set defined as the median of $|\sigma/\xi^{\text{test}}|$ in %, with σ^2 the diagonal of C_{cosmic}. While we find good performance for every model, as the MAEs roughly follow the expected accuracies, we observe some discrepancies. For ξ_0^{gg}, on scales between 1-10 h^{-1}Mpc, the MAE is slightly larger than the expected accuracy.

We also see that the models are sometimes overfitting the test set, especially on ξ_0^{vv} large scales as they are highly correlated and signal is smooth (see figure 5.2). The model does not properly evaluate the amount of noise in the training set. We argue that both these under and over-fitting should be taken into account by the emulator predicted covariance.

We verify this statement using the recovery set as it is no longer fixed to the same IC. For each of the 25 mocks we compute both the difference $\Delta\xi = (\xi^{\text{emu}} - \xi^{\text{reco}})$ and the total covariance matrix $C_{\text{tot}} = C_{\text{cosmic}} + C_{\text{emu}}$ with C_{cosmic} rescaled to match the recovery set volume. Figure 5.7 shows the residuals $\Delta\xi \cdot L^{-1}$ with L^{-1} the Cholesky decomposition (Eq 4.31) of the inverse of the total covariance (allowing to consider the correlations between scales in the residuals). As previously, the blue and orange solid lines in the top panel show the performance of the ξ_0^{gg} and ξ_2^{gg} models, the same colours are used for ξ_0^{vv} and ξ_2^{vv} in the middle panel, and the turquoise lines show the performance for ξ_1^{vg} in the bottom panel. The grey shaded areas mark the one and two sigma deviations. We find that the different models are performing reasonably well, although for ξ_0^{vv} the largest scales residuals can larger than four sigmas, which, out of 25 realisations is statistically significant. In the following we test the effect of this inconsistency when performing an inference on the cosmological and HOD parameters. On should remind that, the considered covariance does not contain any observational noise and is therefore underestimated compared to real data.

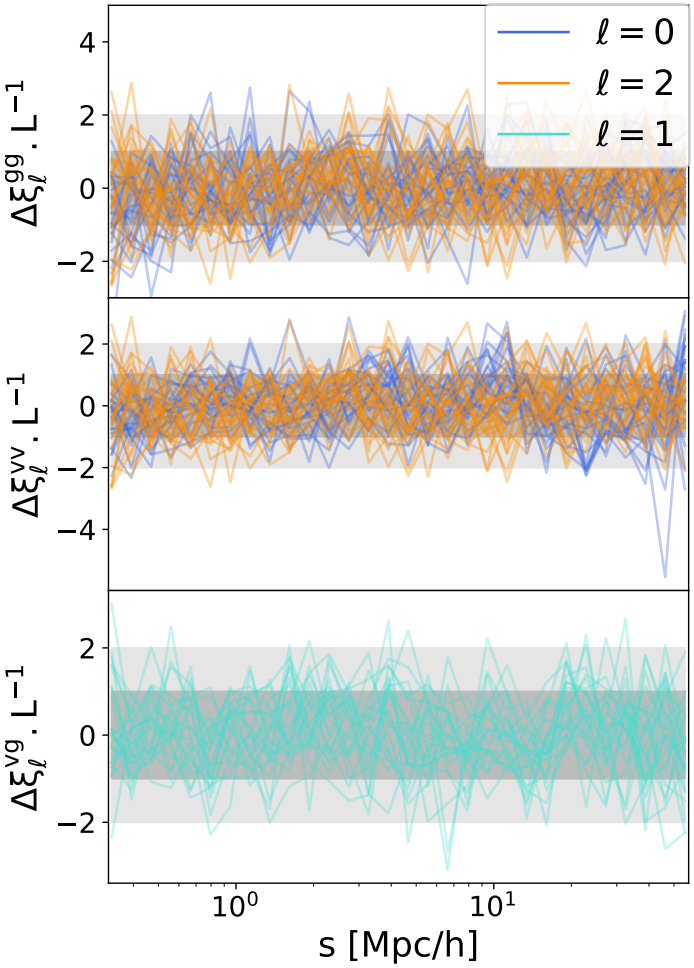


FIGURE 5.7 : Residuals $\Delta\xi \cdot L^{-1}$ with L^{-1} the Cholesky decomposition of the inverse of the total covariance $C_{\text{tot}} = C_{\text{cosmic}} + C_{\text{emu}}$. The blue and orange solid lines in the top panel show the performance of the ξ_0^{gg} and ξ_2^{gg} models. The same colours are used for ξ_0^{vv} and ξ_2^{vv} in the middle panel. The turquoise lines show the performance for ξ_1^{vg} in the bottom panel. The grey shaded areas are the one and two sigma deviations.

5.3 Parameters recovery

In order to validate our models, we use the recovery set to perform statistical inferences over the cosmological and HOD parameters, for different combination of observables.

5.3.1 Inference strategy

We perform the parameter inference by running MCMC chains with the library EMCEE, using the flat prior corresponding to the train set described in table 4.1, and the Gaussian log-likelihood previously defined in Eq 4.32. To cross-check each observable and evaluate the amount of additional information contained in the velocity statistics, we use different concatenation of the multipoles ξ_ℓ^{ab} for the data vector $\boldsymbol{d}$. When using the full set of observable to perform the 3×2 point analysis, the data vector becomes significantly large ($n_d = 150$) with respect to the number of mocks used to estimate the covariance matrix ($n_{\rm mock} = 1400$). Consequently, to account for the noise in the covariance and precision matrix estimates, we rescale $C_{\rm cosmic}$ as described in section 2.3.3, with the revisited factor derived in PERCIVAL et al. 2022

$$m_1 = \frac{(n_{\rm mock}-1)\left[1+B(n_d-n_p)\right]}{n_{\rm mock}-n_d+n_p-1}. \tag{5.4}$$

Since the estimation of the model covariance is time-consuming, we reduce the computational cost of a likelihood estimation by fixing the emulator covariance to its expected value, evaluated for the true parameters of the simulation. We test this approximation in appendix 5.5, and find it consistent with the results obtained when varying the covariance for each likelihood estimation.

5.3.2 Results

We perform four different inferences on the 25 mocks of the recovery set. We use the galaxy information only with the data vector $[\xi_0^{gg},\xi_2^{gg}]$, the velocity information only with the data vector $[\xi_0^{vv},\xi_2^{vv}]$, the cross information only with the data vector$[\xi_1^{vg}]$, and the total information with the data vector $[\xi_0^{gg},\xi_2^{gg},\xi_0^{vv},\xi_2^{vv},\xi_1^{vg}]$. Figure 5.8 shows the one and two sigma confidence levels of the posterior distribution averaged over the 25 realisations of the recovery set. The true values of the parameters are indicated with the black lines. The four methods give consistent results, both for the cosmological and HOD parameters, recovering every fitted parameters within one sigma. The contours obtained using the different observables have slightly different shapes and inclinations. Except for the parameters that weakly affect the clustering (see appendix 5.5) and are therefore unconstrained, jointly fitting the galaxy and peculiar velocity statistics reduces parameter degeneracy and yields tighter constraints, especially for w_0 and σ_8.

Table 5.1 presents the summary statistics computed from the 25 MCMC chains. The corresponding histograms of the best-fitted values and estimated errorbars are shown is appendix 5.5. We find that averaging over the recovery set for every parameter θ, the difference between the expected and best fitted parameter value $\langle\Delta_\theta\rangle$ is well within the estimated one sigma $\langle\sigma_\theta\rangle$. However, the standard deviation of the pull $\sigma(Z_\theta)$ - with $Z_\theta \equiv (\theta - \langle\theta\rangle)/\sigma_\theta$ - is smaller than one for every parameter and every method. While it is challenging to conclude from such a small sample, this might indicate a systematic overestimation of the uncertainties, or a deviation from Gaussianity in the fitted parameter distributions (see histograms appendix 5.5). The reason could

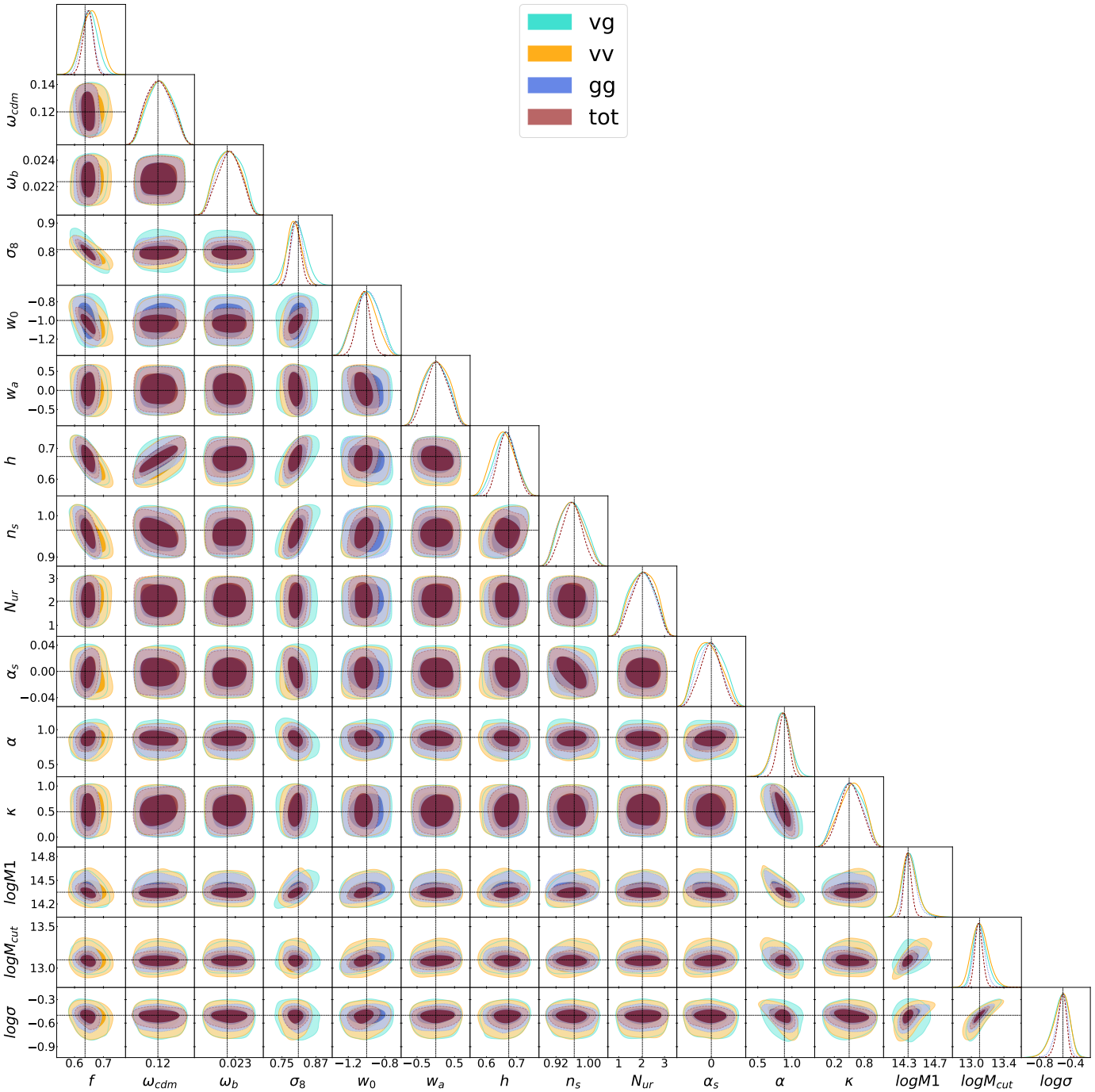


FIGURE 5.8 : Average posterior distribution of the inferred parameters, from the fits of the recovery set. The contours correspond to 1 and 2σ confidence levels. The results obtained using the galaxy, velocity, cross, and total data vectors are shown in blue, orange, cyan and red respectively.

TABLE 5.1 : Summary statistics for the fits of the 25 spherical full sky footprint mocks with Planck2018 cosmology and the median HOD. $\langle\Delta_\theta\rangle$ is the bias, the difference between the true and best fitted value for any parameter θ scaled by 10^2. $\langle\sigma_\theta\rangle$ is the average over the 25 realisations of the (symmetrized) one sigma confidence level scaled by 10^2. $\sigma(Z_\theta)$ is the standard deviation of the pull for any parameter θ. Note that f is a derived parameter from the other cosmological ones.

	gg			vv			vg			tot		
p	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$
f	-0.85	1.94	0.38	-2.08	2.65	0.79	-1.43	2.70	0.31	-1.11	1.54	0.54
ω_m	-0.02	0.95	0.27	-0.12	0.90	0.41	-0.20	0.93	0.31	-0.07	0.93	0.43
ω_b	-0.01	0.09	0.33	-0.01	0.10	0.29	-0.01	0.10	0.23	-0.02	0.08	0.40
σ_8	0.56	2.06	0.35	1.44	2.16	0.54	0.66	2.98	0.40	1.05	1.60	0.47
w_0	-0.10	12.83	0.33	1.67	11.40	0.44	-0.39	13.60	0.27	3.34	6.31	0.50
w_a	0.40	32.77	0.30	-1.38	32.96	0.44	0.52	32.57	0.29	-2.52	28.36	0.32
h	0.86	3.15	0.33	1.82	3.29	0.58	1.13	3.52	0.26	0.66	2.64	0.43
n_s	0.87	2.87	0.41	0.80	2.86	0.44	0.42	3.08	0.23	0.68	2.19	0.51
N_ur	-0.43	52.12	0.21	-3.26	57.55	0.22	3.44	56.77	0.22	-4.68	50.45	0.28
α_s	0.31	1.70	0.28	0.52	1.85	0.52	0.16	2.12	0.22	0.11	1.43	0.56
α	3.24	9.94	0.53	3.84	9.96	0.69	3.54	10.99	0.49	1.73	6.12	0.84
κ	-0.33	23.51	0.39	-6.79	23.37	0.41	-3.27	25.70	0.25	-3.74	18.68	0.69
$\log M_1$	-2.31	6.98	0.55	-1.63	8.38	0.57	-2.31	7.97	0.62	0.43	3.94	0.61
$\log M_\mathrm{cut}$	0.35	6.24	0.40	-0.35	8.77	0.58	-0.18	8.05	0.30	1.13	3.99	0.62
$\log\sigma$	0.79	6.80	0.43	1.42	9.51	0.49	1.94	10.24	0.35	1.24	5.00	0.60

be a wrong estimation of the emulator error, as we have seen that the transformation Eq 5.3 induces non-Gaussian predictions for the GP.

Yet, the results obtained are conservative. The mean reduced chi-squared are equal to 1.1 for the "gg" analysis, 0.96 for "vv", 1.49 for "vg", and 0.85 for the complete joint analysis. While those values deviates from one, the corresponding p-value $[\mathrm{gg},\mathrm{vv},\mathrm{vg},\mathrm{tot}]_\text{p-value} = [0.3, 0.6, 0.1, 0.9]$ indicate reasonably good fits. Assuming the models to be valid, the data points and associated covariance are quite likely.

Compared to the inference using only the galaxy auto-correlation, using the velocity auto-correlation or the cross-correlation yield looser constraints on the parameters of interest, namely the amplitude σ_8 and growth rate f. This is expected as the velocity perturbations spans larger perturbation modes. We get similar results for the HOD parameters which only affect the velocity clustering on scales typically below $\sim 10h^{-1}$Mpc(dynamical range appendix 5.5). While less information is contained in the velocity auto and cross-correlations than in the galaxy density, combining the two tracers increases the constraints obtained on every parameter. Nevertheless, the velocity statistics yield globally larger mean parameter biases. Therefore, combining the galaxy and velocity information result in an overall increase of the systematic shifts - notably for f and σ_8 - compared to the galaxy information alone

While those biases are well within the estimated uncertainties for individual parameters, we

can evaluate the statistical significance on the total cosmological and HOD parameters biases. Accounting for the different correlations between parameters, we can compress the posterior information using the Figure of Merit (FoM) and Figure of Bias (FoB) defined as

$$\mathrm{FoB} = \sqrt{(\theta_{\mathrm{best}} - \theta_{\mathrm{true}})^{\top} C_{\theta}^{-1} (\theta_{\mathrm{best}} - \theta_{\mathrm{true}})}, \quad \text{and} \quad \mathrm{FoM} = \frac{1}{\sqrt{|C_{\theta}|}}, \tag{5.5}$$

where C_θ is the parameter covariance matrix estimated from the posterior distribution. Assuming that a given set of parameters θ are Gaussianly distributed (following the posterior), the FoB is the square root of the corresponding chi-squared evaluated at the parameters best-fit values. This metric indicates how significant the biases on the parameters are (assuming a Gaussian posterior). The FoM is useful to estimate the overall constraints on a set of parameters, accounting for the correlations.

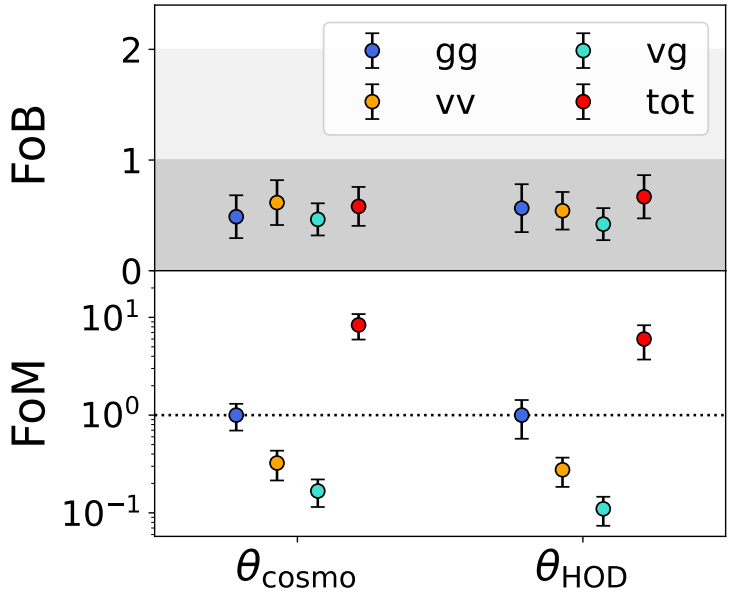


FIGURE 5.9 : Cosmological and HOD parameters recovery for the four inference strategies run on the recovery set. Top panel : FoB normalised by the $\sqrt{\chi^2}$ associated with a p-value of 0.68%, that is 10.4 for θ_{cosmo} and 5.9 for θ_{HOD}. The shaded area correspond to 1 and 2 σ CL. Bottom panel : FoM normalised by the "gg" values : 4.97×10^{14} for θ_{cosmo} and 2.20×10^{6} for θ_{HOD}.

For every realisation of the recovery set, we compute the FoM and FoB for the 9 cosmological parameters' subspace θ_{cosmo}, and the 5 HOD parameters sub-spaces θ_{HOD}. Figure 5.9 gives the average results for the corresponding "gg", "vv", "vg" and "tot" analyses, designed by the same colors as previously. The errorbars correspond to the standard deviations of the FoM and FoB across the 25 realisations. The bottom panel shows the cosmological and HOD mean FoMs, normalised by the value obtained for the "gg" analysis : 4.97×10^{14} for θ_{cosmo} and 2.20×10^{6} for θ_{HOD}. The top panel of figure 5.9 shows the FoB normalised by the $\sqrt{\chi^2}$ values associated with a p-value of 0.68%, corresponding to $\sqrt{\chi^2} = 10.4$ for the 9 dof cosmological space, and $\sqrt{\chi^2} = 5.9$ for the 5 dof HOD space. Hence, the grey shaded area correspond to 1 and 2 σ CL. We find unbiased θ_{cosmo} and θ_{HOD} for every inference strategy, and retrieve the same ordering of the constraining power $\mathrm{FoM_{vg}} < \mathrm{FoM_{vv}} < \mathrm{FoM_{gg}} < \mathrm{FoM_{tot}}$. On mildly to non-linear scales, peculiar

velocity clustering holds additional information not contained in the galaxy clustering alone, both for the cosmological and galaxy-halo bias parameters.

5.4 Inference on noisy measurements

So far we have assumed that every galaxy had its peculiar velocity perfectly measured without any noise. In this section we study how constraints are impacted when using sparse noisy measurements of peculiar velocities.

5.4.1 Constraining power

Following the same inference strategy (with the corresponding data covariance matrix), we fit the clustering of the 25 realistic mocks described in section 5.1.2, reproducing the expected density of tracer, uncertainties on the velocity measurements, and associated weights (Eq 5.1). Additionally, we include AP distortions in the clustering, using the c003 fiducial cosmology in AbacusSummit to convert redshifts into comoving distances. Given that we do not have a theoretical prediction for the 2D correlation functions $\xi^{ab}(s,\mu)$, we model the AP distortions -

TABLE 5.2 : Summary statistics for the fits of the 25 realistic mocks with Planck2018 cosmology and the median HOD. $\langle\Delta_\theta\rangle$ is the bias, the difference between the true and best fitted value for any parameter θ scaled by 10^2. $\langle\sigma_\theta\rangle$ is the average over the 25 realisations of the (symmetrized) one sigma confidence level scaled by 10^2. $\sigma(Z_\theta)$ is the standard deviation of the pull for any parameter θ. Note that f is a derived parameter from the other cosmological ones.

	gg			vv			vg			tot		
p	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$	$\langle\Delta_\theta\rangle$	$\langle\sigma_\theta\rangle$	$\sigma(Z_\theta)$
f	-0.04	2.72	0.55	-2.87	4.77	0.18	-4.09	4.65	0.38	-0.40	2.56	0.61
ω_m	-0.09	0.89	0.60	-0.44	1.12	0.21	-0.47	0.97	0.40	-0.14	0.87	0.49
ω_b	-0.00	0.09	0.54	-0.02	0.11	0.29	-0.02	0.10	0.34	-0.01	0.09	0.54
σ_8	-1.64	4.16	0.44	-0.43	5.97	0.53	-5.20	5.46	0.37	-1.93	3.14	0.64
w_0	-2.62	14.12	0.43	-0.19	16.83	0.17	1.17	15.45	0.46	-0.26	12.48	0.52
w_a	-4.03	31.38	0.60	2.13	37.36	0.30	2.01	36.76	0.31	-3.68	30.80	0.53
h	0.26	3.43	0.51	2.48	4.78	0.22	3.43	4.60	0.39	0.20	2.89	0.53
n_s	0.66	3.07	0.47	-0.64	3.55	0.18	-0.15	3.40	0.38	0.90	2.83	0.50
N_ur	5.43	51.68	0.45	-4.64	61.66	0.26	-1.32	58.10	0.34	13.51	47.56	0.63
α_s	0.44	1.86	0.50	-0.04	2.31	0.24	-0.30	2.10	0.37	0.10	1.83	0.57
α	-31.07	13.82	0.46	-8.78	37.02	0.34	-16.31	34.77	0.30	-33.60	13.85	0.64
κ	-2.76	23.17	0.48	1.54	29.76	0.29	3.96	29.66	0.27	-3.29	22.67	0.58
$\log M_1$	51.93	7.58	0.38	32.74	29.90	0.27	41.58	22.82	0.30	53.39	7.18	0.46
$\log M_\mathrm{cut}$	22.89	20.96	0.47	15.25	32.93	0.22	19.35	33.39	0.28	27.03	21.35	0.57
$\log\sigma$	-25.73	8.32	0.56	-14.40	25.78	0.31	-20.74	23.13	0.31	-24.21	9.50	0.60

using the isotropic and anisotropic dilations parameters (α_V, ϵ) defined in Eq 2.24 - directly on the multipoles, following Xiaoying XU et al. 2013 at first order (neglecting the derivative terms). The parameters are updated at each likelihood evaluation. Note that an additional dependence on the fiducial cosmology is also introduced in the velocity measurement through Eq 2.74. We leave the detailed study of this effect to future work and use the true underlying cosmology (Planck18) to compute velocities from log-distance ratio.

Table 5.2 presents the summary statistics computed from the fits. The corresponding mean posterior distribution and best-fitted parameters histograms are shown is appendix 5.5. We find similar results to the previous section. While the standard deviation of the pull $\sigma(Z_\theta)$ is small compared to one, each of the cosmological parameters are recovered within 1σ. The "vv" and "vg" analyses yield larger systematic shift and errorbars compared to the "gg" results, and the combined probes give the tightest (unbiased) cosmological constraints. We get the following mean reduced chi-squared $\left[\text{gg}, \text{vv}, \text{vg}, \text{tot}\right]_{r\chi^2} = [1.1, 4.3, 3.0, 0.4]$, with the corresponding p-values $\left[\text{gg}, \text{vv}, \text{vg}, \text{tot}\right]_{\text{p-value}} = \left[0.3, 2\times10^{-20}, 6\times10^{-5}, 1-1\times10^{-10}\right]$. While the AP distortion modelling is only approximated, it remains a weak effect and is expected to cause the worse quality fits for the "vv" and "vg" inferences, most likely due to the reduced density and the random scatter in the velocities. As seen in figure 5.5, while the random noise on the velocity measurement is almost averaged out for the largest modes, on the smallest scales it reduces the correlations between separation bins and significantly increase the variance of the clustering. For the HOD parameters which strongly affect those scales, we find systematic biases comparable or larger than the estimated uncertainties.

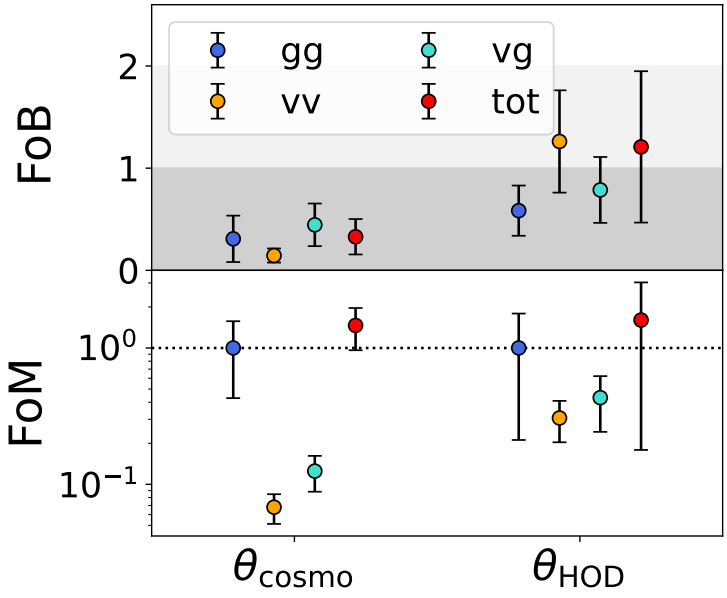


FIGURE 5.10 : Cosmological and HOD parameters recovery for the four inference strategies run on the realistic mocks. Top panel : FoB normalised by the $\sqrt{\chi^2}$ associated with a p-value of 0.68%, that is 10.4 for θ_{cosmo} and 5.9 for θ_{HOD}. The shaded area correspond to 1 and 2 σ CL. Bottom panel : FoM normalised by the "gg" values, 4.85×10^{13} for θ_{cosmo} and 5.59×10^4 for θ_{HOD}.

As done previously, in order to assess the overall systematic bias and constraining power, we compute for each analysis the mean and standard deviation of the FoM and FoB in the

parameter subspaces $\theta_{\rm cosmo}$ and $\theta_{\rm HOD}$. Figure 5.10 presents the results obtained for the four methods, preserving the same colour scheme. First, we notice that fitting only the velocity-autocorrelation or the total joint clustering yields a 1σ level bias on HOD parameters. The FoB of the cross-correlation analysis is increased as well, but remains in average bellow one. Furthermore, we notice that compared to the noise-free high density constraints in figure 5.9, the FoM for the gg analysis - used as a normalisation - dropped by a factor 10 for $\theta_{\rm cosmo}$, and almost 40 for $\theta_{\rm HOD}$, indicating a loss of constraining power. Yet, for every method we obtain unbiased cosmological parameters. Despite the reduced number of tracers and the noisy peculiar velocity measurements, we find that jointly fitting the galaxy and velocity clustering improves the cosmological constraints, even though the improvement is less significant than when using the full velocity sample.

5.4.2 Varying the minimum fitting scale

Since the noise in the peculiar velocity measurements mostly affects the smallest scales, the inference might be optimised by removing some separation bins, thereby further compressing the data vector. We investigate the dependence of the constraining power and systematic bias on the minimum fitted scale. We fit the total clustering information $[\xi_0^{gg}, \xi_2^{gg}, \xi_0^{vv}, \xi_2^{vv}, \xi_1^{vg}]$ for each individual realisations of the the realistic mocks, using different minimum scales $s_{\rm min} = [0.33, 1.13, 3.92, 13.4]$ in h^{-1}Mpc, corresponding to $[30, 21, 14, 7]$ data-points for each multipole.

Figures 5.11 displays the best-fitted values for the cosmological and HOD parameters as a function of $s_{\rm min}$. The grey lines indicate the results of the individual realisations, the blue dots correspond to the average, and the errorbars the mean of the estimated uncertainties. The red dotted lines point the expected values of the parameters. We observe that the true cosmological parameters are consistently retrieved within the error bars. Additionally, decreasing $s_{\rm min}$ tends to reduce the shift of the best-fitted values and the size of the error bars. Conversely, while decreasing $s_{\rm min}$ reduces uncertainties in the HOD parameters, it does not sufficiently mitigate the systematic shift, leading to biased constraints except for the power-law index of the satellite occupation. We measure poor p-values, larger than $1 - 1 \times 10^{-3}$, except for the largest minimum scale $s_{\rm min} = 13.4 h^{-1}$Mpc, where we find a p-value of 0.1. This is in line with the fact that despite the enlarged variance on small scales, the HOD parameters are effectively overfitting the signal.

We further compute for each $s_{\rm min}$ the mean and standard deviation of the FoM and FoB for the cosmological and HOD parameters, and show the results in figure 5.12. For the HOD parameters, the mean FoM increases as we consider smaller scales, but its variance also rises. Simultaneously, the mean FoB grows and exceeds 1 for $s_{\rm min} \gtrsim 4 h^{-1}$Mpc. However, the flexibility of the HOD formalism allows to marginalise over unmodelled noise, preventing bias in the cosmological parameters. For any minimum scale, the cosmological parameters FoB remains below 1. Moreover, as the minimum fitted scale decreases, the cosmological parameters' FoM increases, tightening the constraints.

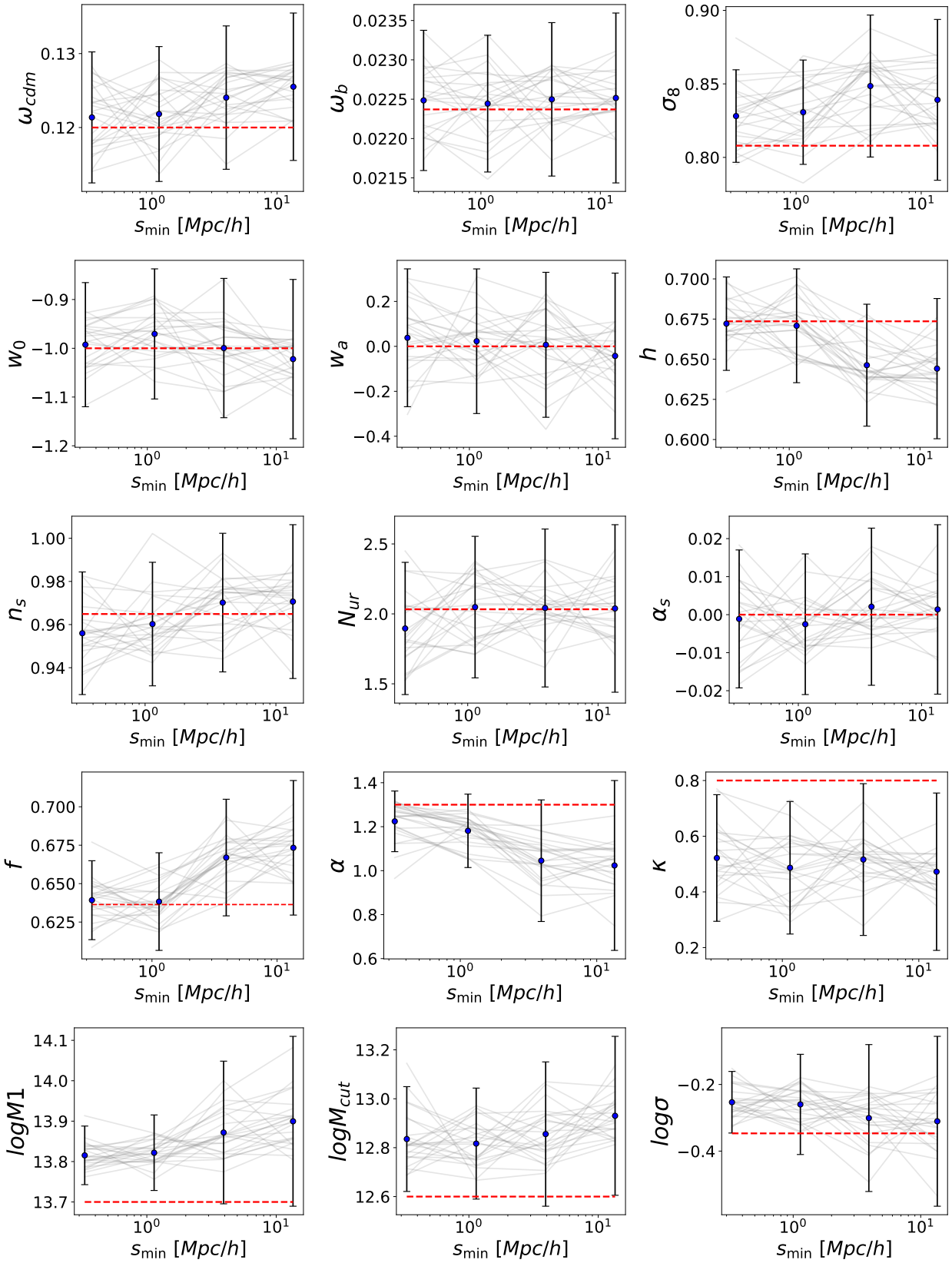


FIGURE 5.11 : Best-fitted parameters obtained for different minimum scales of the realistic mocks. The inferences are run using the total clustering information. The red dotted lines indicates the expected values. The blue dots show the average best-fit of the individual realisations corresponding to the grey lines. The errorbars are the mean of the estimated uncertainties.

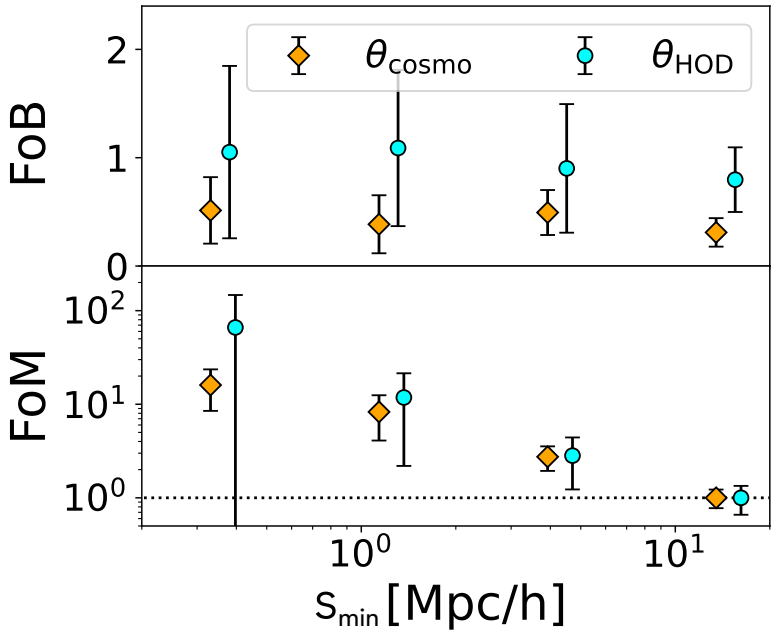


FIGURE 5.12 : Cosmological and HOD parameters recovery using different minimum scales of the realistic mocks to run inference with the total clustering information. Top panel : FoB normalised by the $\sqrt{\chi^2}$ associated with a p-value of 0.68%, that is 10.4 for $\theta_{\rm cosmo}$ and 5.9 for $\theta_{\rm HOD}$. The shaded area correspond to 1 and 2 σ CL. Bottom panel : FoM normalised by the value obtained for the largest $s_{\rm min}$, 5.96×10^{12} for $\theta_{\rm cosmo}$ and 1.32×10^{3} for $\theta_{\rm HOD}$.

It is worth emphasizing that for a real survey, fibre collision systematics could influence this result. Indeed, if not properly corrected, fibre collisions mainly bias the clustering measurement for the smallest separation bins. Additionally, while we found that the HOD parameters were able to absorb additional sources of noise, effectively unmodelled physics, our noisy sample is built from the same underlying formalism. We acknowledge that this empirical relationship does not fully capture the complexity of galaxy formation physics. Therefore, before applying such emulator model to real data, it is crucial to assess its robustness against more physically motivated galaxy-halo relation (such as SHAM or hydrodynamical simulations described in section 2.1.5).

5.5 Conclusion

In this study, we applied our estimator for the velocity two-point statistics (introduced in Chapter 2) to the SBI framework. We trained emulator models for the redshift space galaxy and velocity clustering using the AbacusSummit simulations and HOD formalism. We demonstrated that using the galaxy-galaxy, velocity-velocity, or velocity-galaxy correlation functions, we were able to recover unbiased constraints for the cosmological and HOD parameters. Moreover, combining the velocity and galaxy clustering on mildly to non-linear scales yielded tighter constraints.

We built a set of realistic mocks, describing the typical density and error in the peculiar velocity measurements of type-Ia supernovae and the TF/FP relations. While fitting those mocks results in systematic errors in HOD parameters, the cosmological parameters remain unbiased. Although the HOD model is flexible enough to describe the noisy clustering without biasing the cosmological parameters when marginalising, their physical interpretation must be treated with caution. Even with realistic measurements, we found that the joint analysis of galaxy and peculiar velocity clustering leads to tighter constraints. Although the gain is reduced compared to the case where we have a velocity measurement for every galaxy, this motivates the measurement of larger samples and reduction of the empirical relation scatter with accurate calibration. Additionally, we found that even the smallest scales, going down to 0.33 h^{-1}Mpc, improve the cosmological constraints. Thus, it would be worthwhile to push the effort to correct for fibre collision systematics.

It is essential to ensure that the HOD parameters are properly sampled to describe the clustering of a given tracer. Making use of the MKGP formalism, the emulator parameter space could be extended to more elaborate HOD models, including assembly bias (Xiaoju Xu et al. 2021) or velocity bias extension (Guo et al. 2015) to describe any additional scatter. The emulator could also be extended to different redshift bins, additional multipoles, larger scales, or forward-modelled observational effects. To properly account for wide-angle effects for low redshift surveys, the emulated $\xi(s,\mu)$ (or multipoles) should either be trained on a given footprint or binned along an additional dimension to fully characterize triangles formed by each pair of galaxies and the observer (as described in section 2.2.3). This last option would require a detailed study of the compression/information-loss balance.

It is worth mentioning that any emulator model highly depends on the simulation suite's numerical specificity and the galaxy model employed. In particular, the HOD is an empirical formalism and does not capture the true complexity of galaxy formation. We require a galaxy model as constrained (i.e., physically motivated) as possible while being flexible enough to marginalise over unmodelled physics without biasing the cosmological measurements. Therefore, any SBI emulator should be tested against a simulation with a higher level of complexity before being applied to real data.

Appendix A : Emulator dynamical range

The following figures describe the variations of the multipole emulators, varying one cosmological or HOD parameters at the time over the train set range, while the rest of the parameters are kept fixed to the Planck2018 or median HOD respectively.

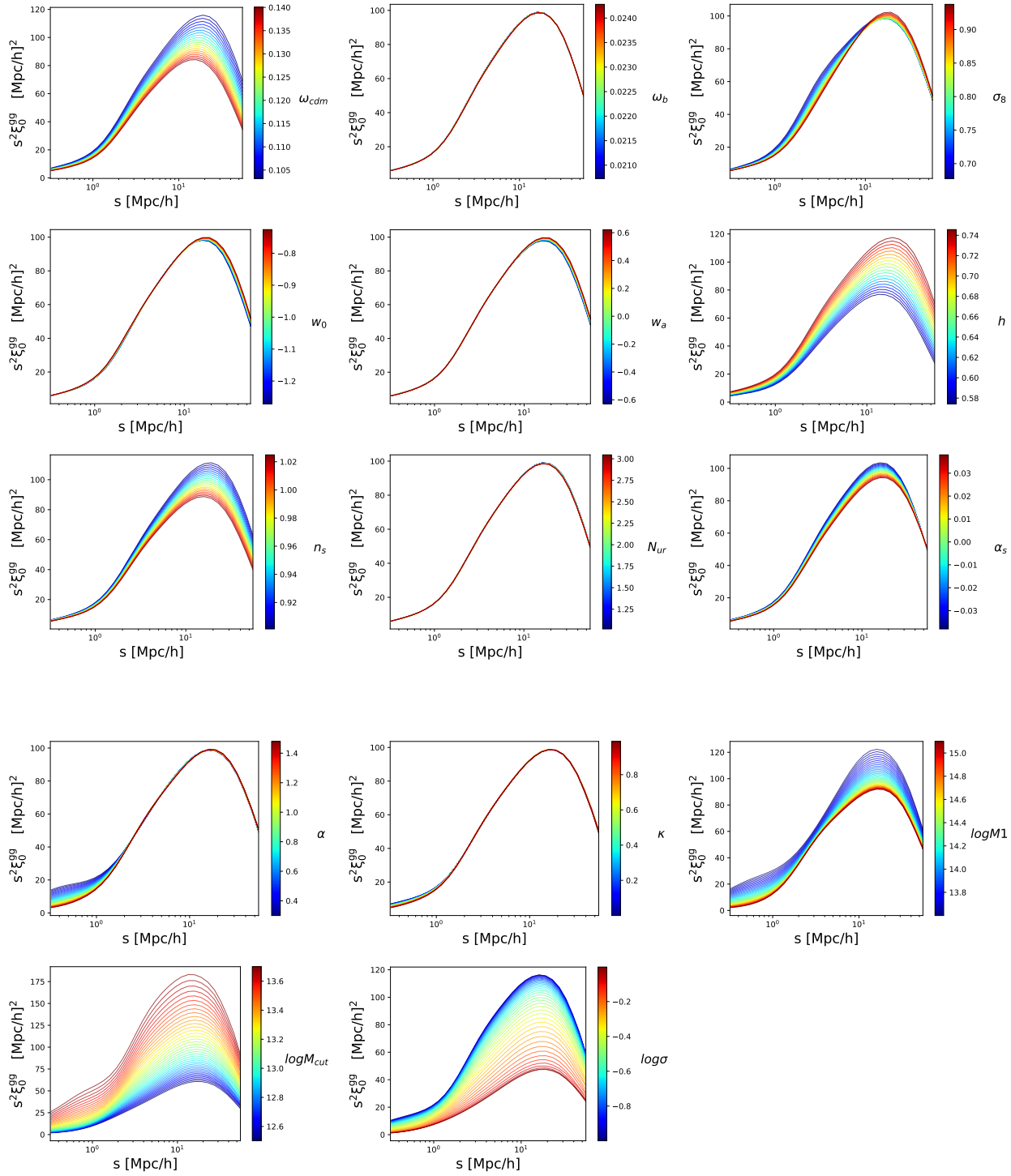


FIGURE 5.13 : Dynamical range over the cosmological and HOD parameter space for the $\xi_0^{gg}(s, \theta_{\rm cosmo}, \theta_{\rm HOD})$ emulator model.

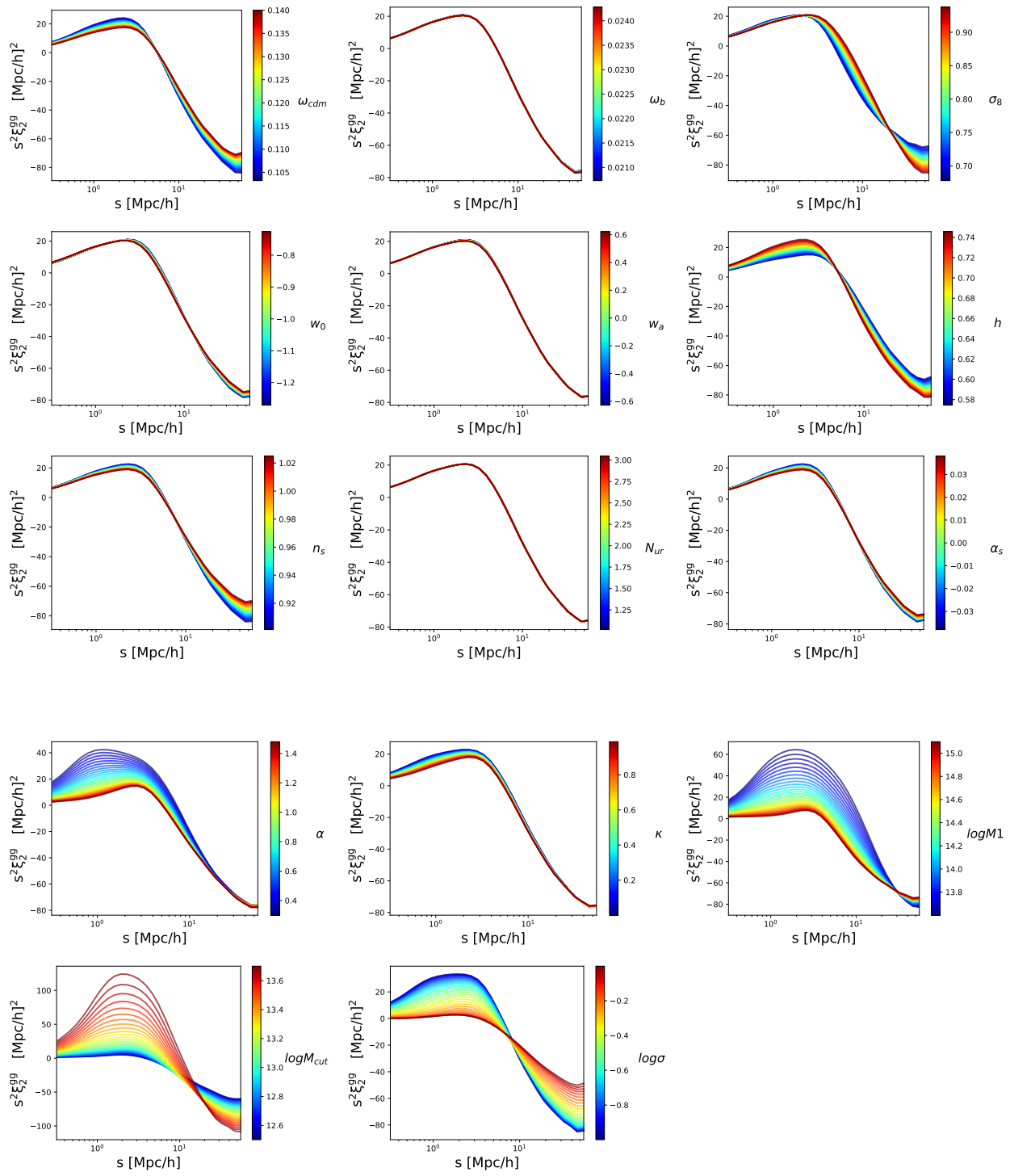


FIGURE 5.14 : Dynamical range over the cosmological and HOD parameter space, within the train set range, for the $\xi_2^{gg}(s, \theta_{\rm cosmo}, \theta_{\rm HOD})$ emulator model.

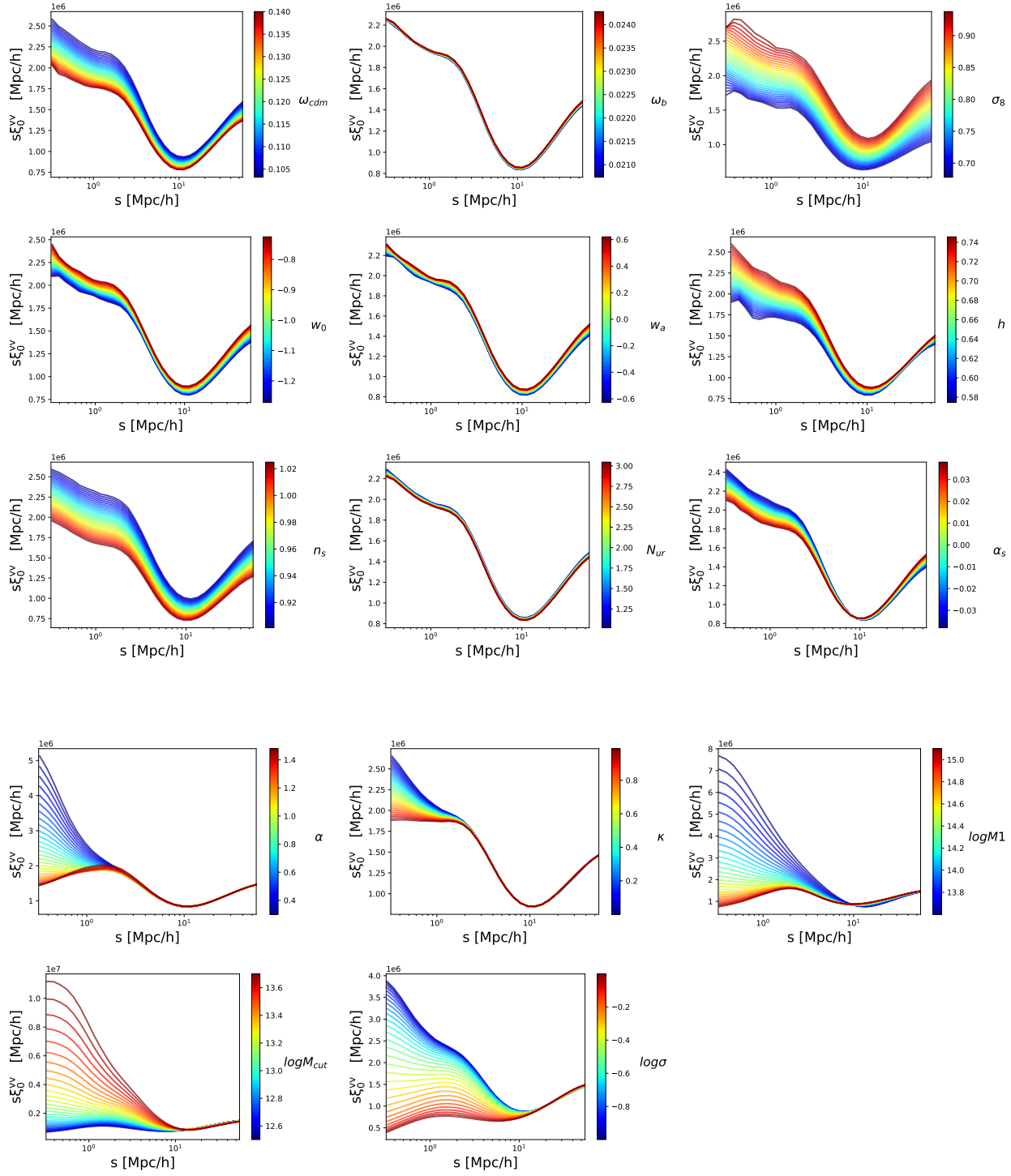


FIGURE 5.15 : Dynamical range over the cosmological and HOD parameter space, within the train set range, for the $\xi_0^{vv}(s, \theta_{\rm cosmo}, \theta_{\rm HOD})$ emulator model.

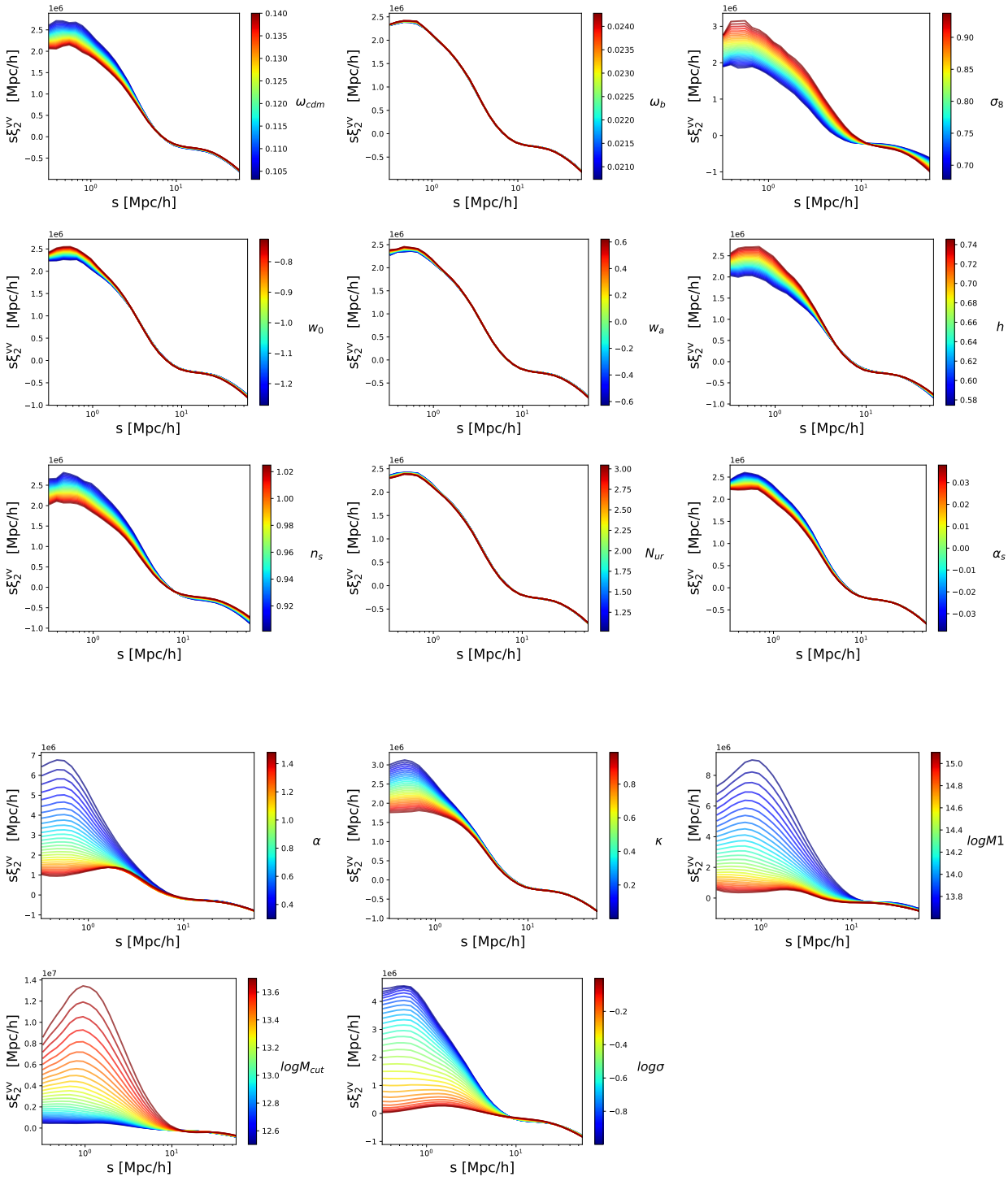


FIGURE 5.16 : Dynamical range over the cosmological and HOD parameter space, within the train set range, for the $\xi_2^{vv}(s, \theta_{\rm cosmo}, \theta_{\rm HOD})$ emulator model.

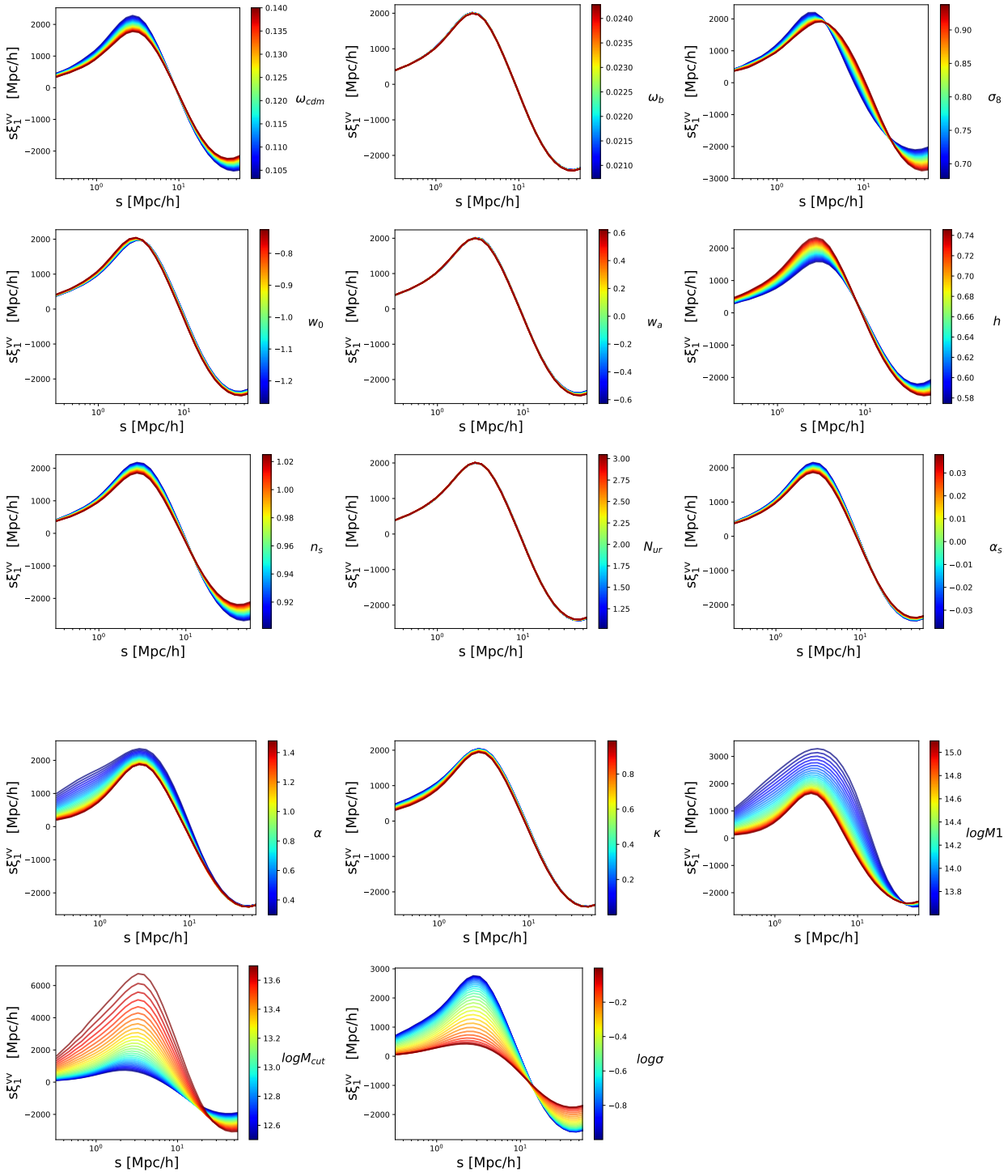


FIGURE 5.17 : Dynamical range over the cosmological and HOD parameter space, within the train set range, for the $\xi_1^{vg}(s, \theta_{\rm cosmo}, \theta_{\rm HOD})$ emulator model.

Appendix B : Covariance template scaling

We run a 3×2 point joint analysis on the recovery set while letting the emulator predicted covariance vary, or fixing it to its expected value, evaluated for the true parameters of the simulation.

Figure 5.18 show the corresponding posterior obtained by averaging over the 25 chains. While varying the covariance increases the constraints on some parameters, we find consistent results with the results between the two methods.

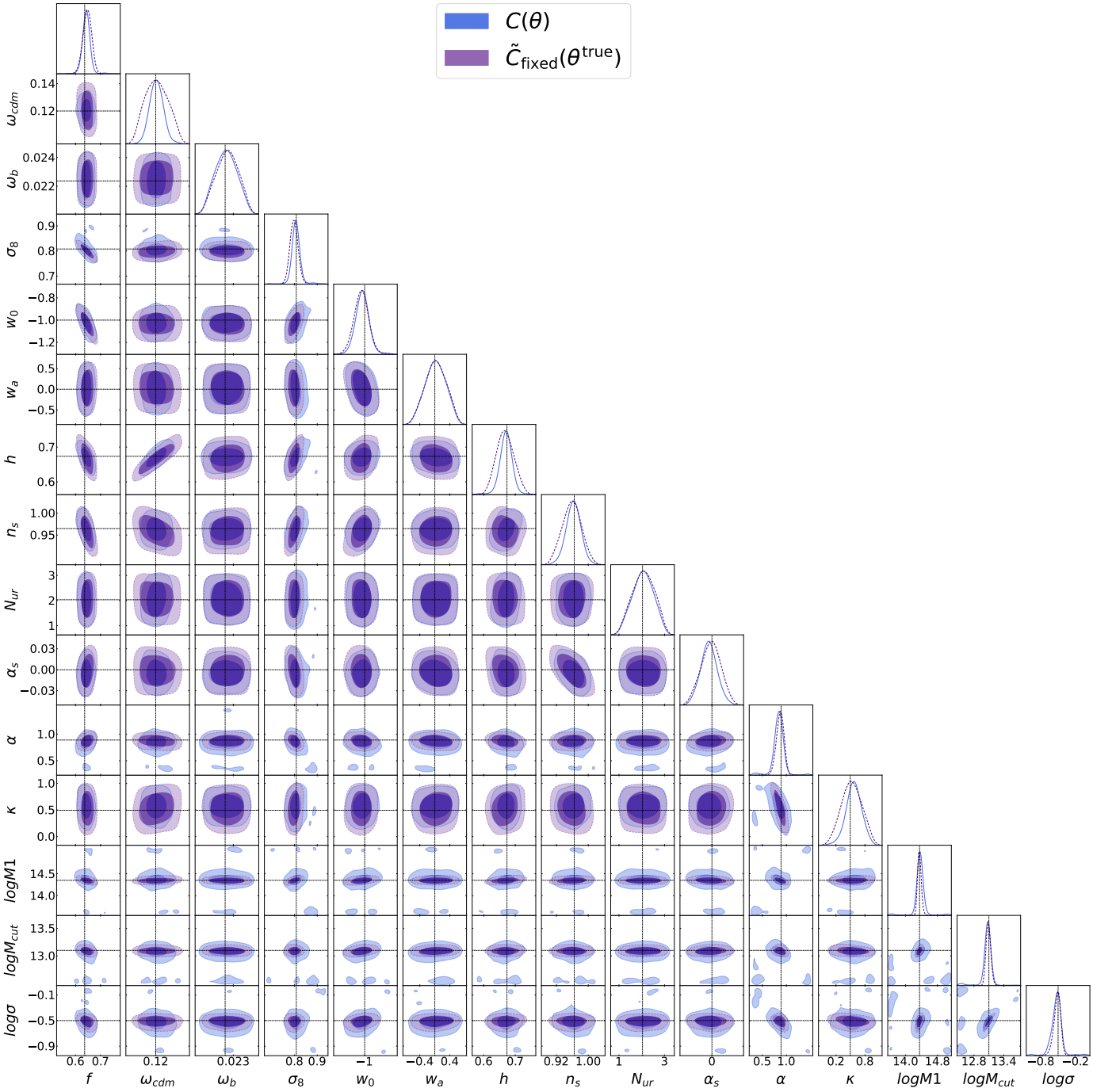


FIGURE 5.18 : Average posterior distribution of the HOD and cosmological parameters, obtained after running joint 3×2 point inference on the recovery set. Blue contours correspond to the results obtained when updating the emulator model covariance for each likelihood estimation. Purple contours are the results obtained by fixed the model covariance at its expectation value. While varying the covariance slightly improve the overall constraints, we find consistent contours between the two methods.

Appendix C : Recovery mocks best-fitted parameters

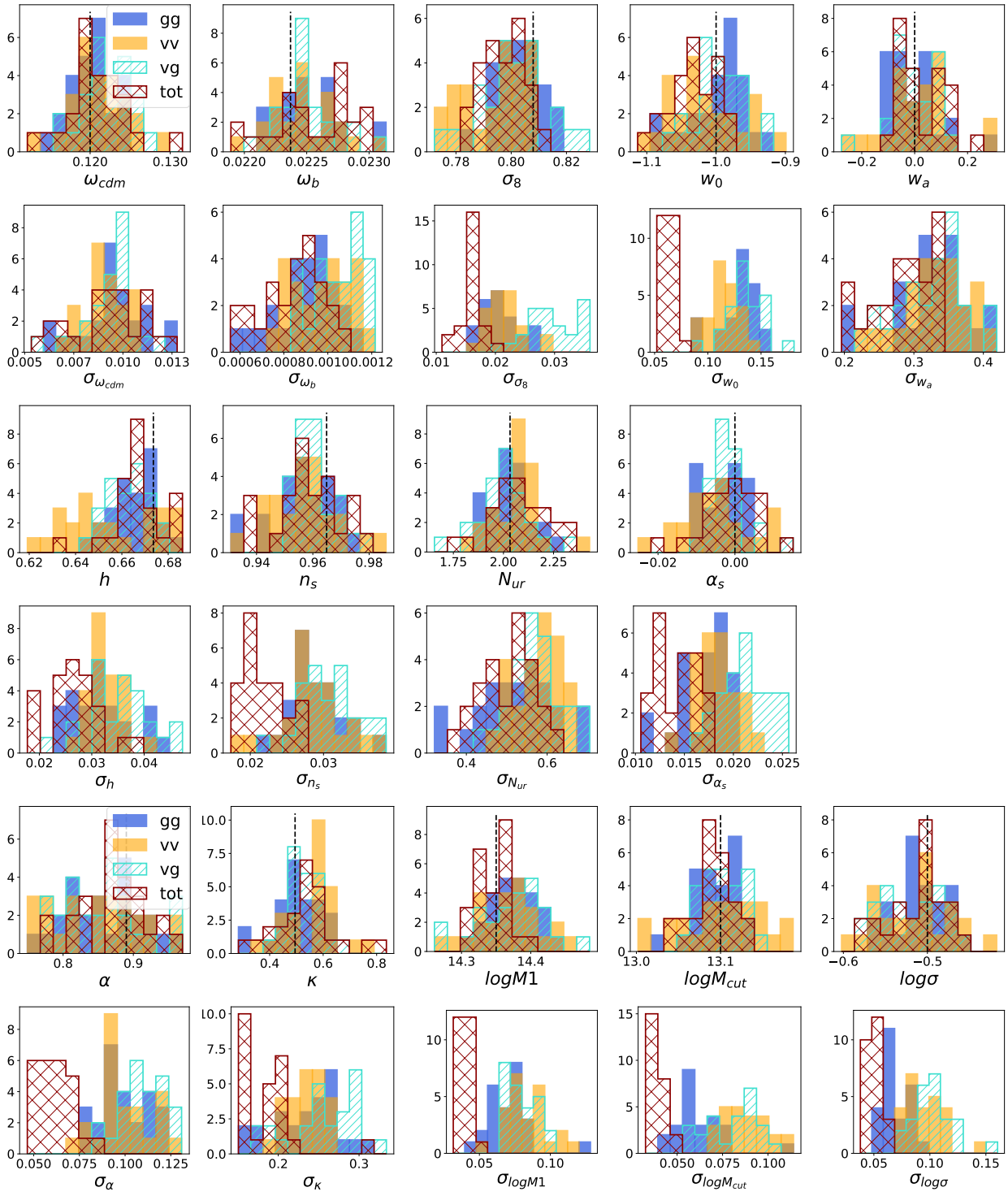


FIGURE 5.19 : Best fitted parameters along with the corresponding uncertainties, measured on the 25 mocks of the recovery set. Results obtained using the data vectors ξ^{gg}, ξ^{vv}, ξ^{vg}, and the joint analysis are respectively shown in blue, orange, turquoise and red. The true parameters values are pointed by a red dotted line.

Appendix D : Realistic mocks best-fitted parameters

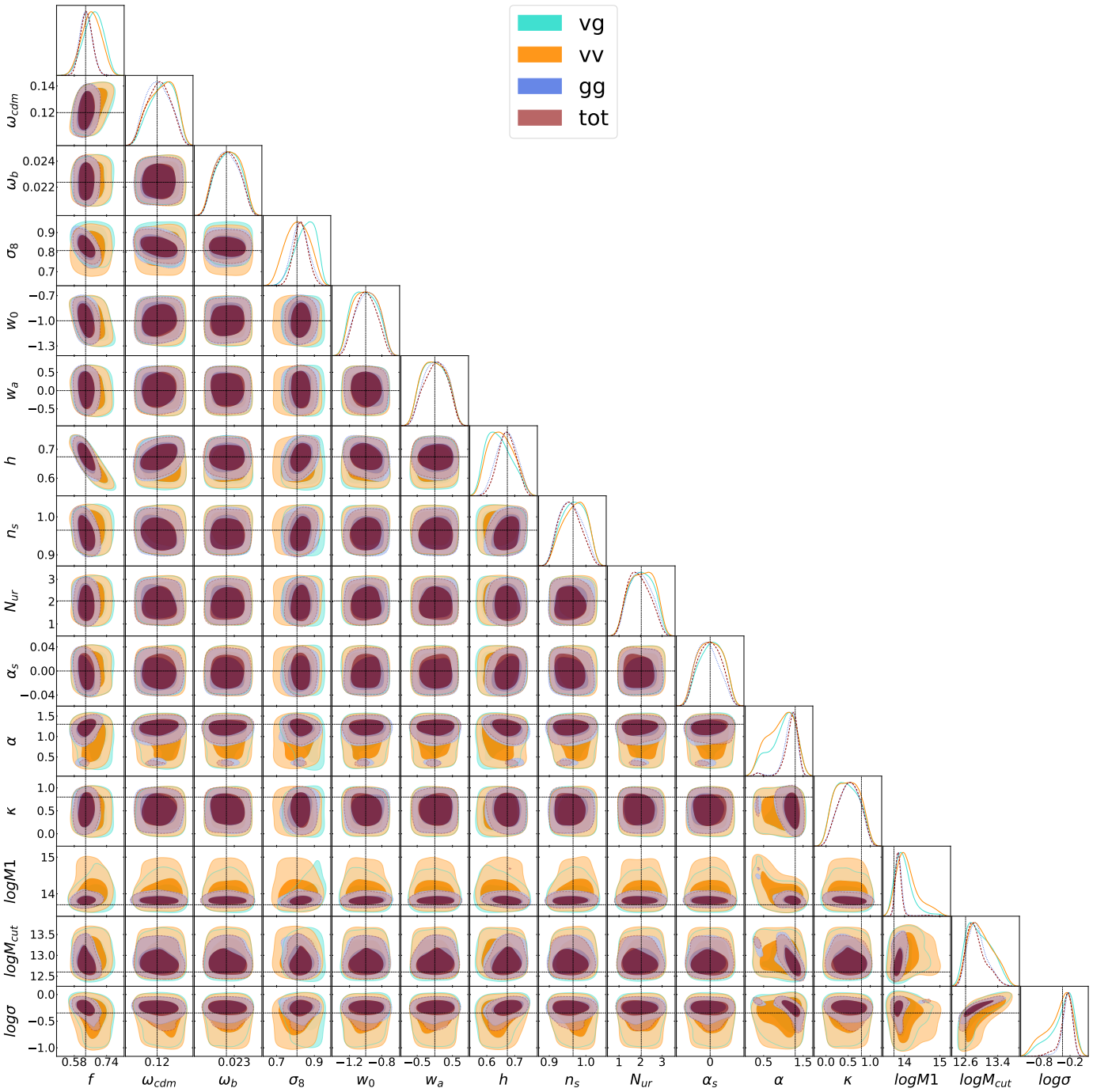


FIGURE 5.20 : Average posterior distribution of the inferred parameters, from the fits of the realistic mocks. The contours correspond to 1 and 2σ confidence levels. The results obtained using the galaxy, velocity, cross, and total data vectors are shown in blue, orange, cyan and red respectively.

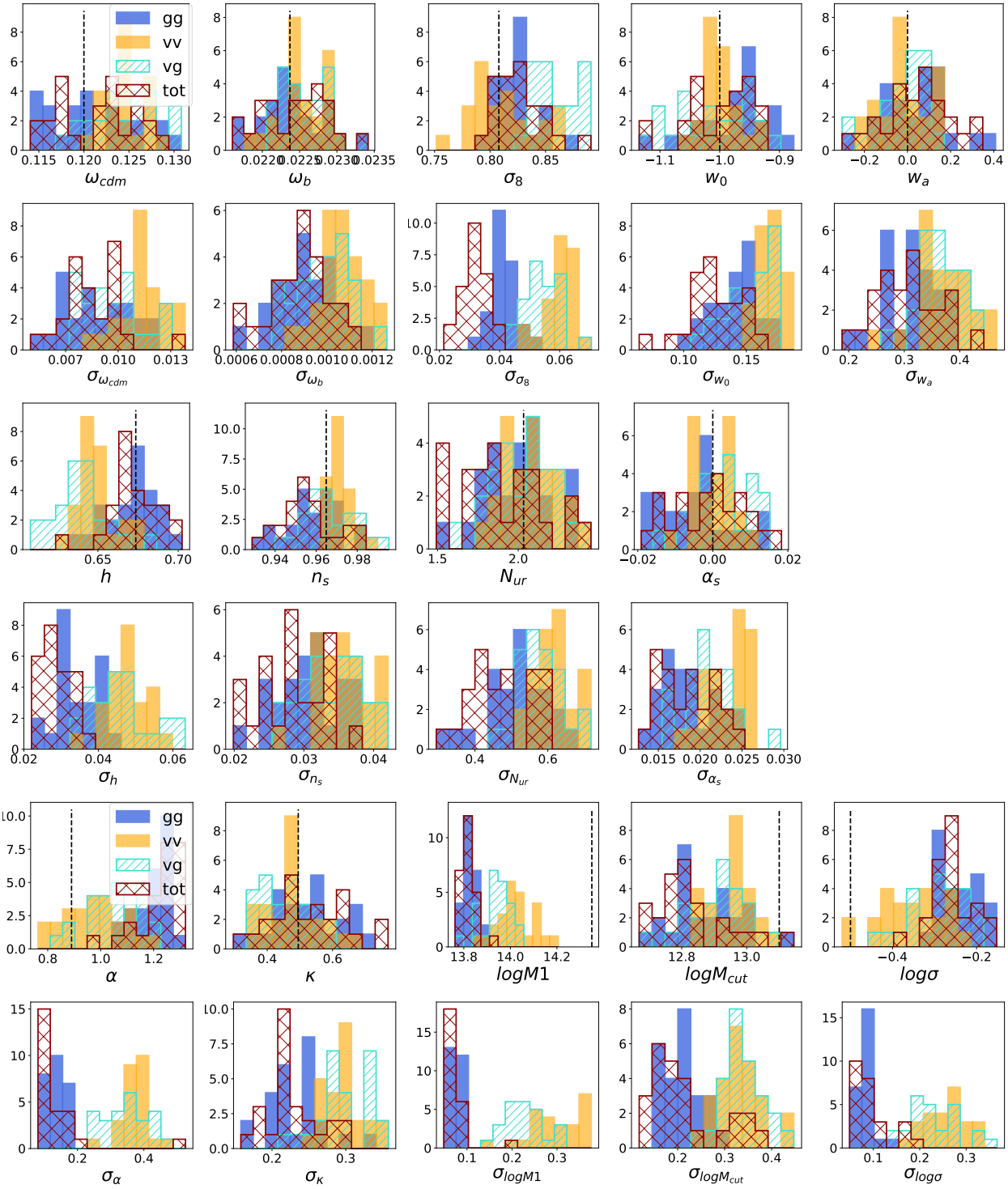


FIGURE 5.21 : Best fitted parameters along with the corresponding uncertainties, measured on the 25 realistic mocks. Results obtained using the data vectors ξ^{gg}, ξ^{vv}, ξ^{vg}, and the joint analysis are respectively shown in blue, orange, turquoise and red. The true parameters values are pointed by a red dotted line.

6 Tully-Fisher and fundamental plane peculiar velocity measurements in the MilleniumTNG hydrodynamical simulation

Summary

6.1 Mock galaxy catalogue . . . 181
6.1.1 The Millennium TNG simulation . . . 181
6.1.2 Sample selection . . . 182
6.2 Peculiar velocity measurement . . . 184
6.2.1 Tully-Fisher and fundamental plane relations . . . 184
6.2.2 Peculiar velocity clustering . . . 186
6.3 Dependency of residuals with environment . . . 188
6.3.1 Properties dependency . . . 189
6.3.2 Extended TF and FP relations . . . 192
6.4 Star formation history . . . 193
6.5 Conclusion . . . 195

Standardisable extragalactic objects play a crucial role in measuring cosmic distances and understanding large-scale structures. In particular, the properties of spiral and elliptical galaxies respectively follow the Tully-Fisher (TF) and Fundamental Plane (FP) relations. These relations have been historically used for many decades to obtain extragalactic distances (SAULDER et al. 2013; CAMPBELL et al. 2014; HONG et al. 2019; HOWLETT et al. 2022; R. Brent TULLY et al. 2023; SAULDER et al. 2023a). The distance measurement of a given galaxy combined with its spectroscopic redshift provides an estimate for its peculiar velocity (assuming some distance-redshift relation). As discussed in the previous chapter, jointly analysing the peculiar velocity and galaxy clustering significantly improves constraints on cosmological parameters, particularly on the growth rate of structure.

The TF and FP relations are empirical and contain significant intrinsic scatter due to complex galaxy formation processes, which propagate directly to peculiar velocity measurements. Recent TF and FP catalogues report an average scatter of slightly more than 20% in distance ratio (to be compared to 7% for type Ia supernovae). Understanding and mitigating this scatter is essential for precise cosmological measurements. In this chapter, we study the intrinsic scatter of these relations using the state-of-the-art hydrodynamical simulations, the MillenniumTNG (PAKMOR et al. 2023). Hydrodynamical simulations are the best theoretical tools available for describing large-scale structures, accounting for the impact of baryonic physics, and providing a physical description of galaxy formation processes. Previous hydrodynamical simulations, such as the Illustris project (VOGELSBERGER et al. 2014), have successfully reproduced the TF and FP relations with a high degree of similarity to real observations (LU et al. 2020; GODDY et al. 2023). Other simulation suites, such as Simba (GLOWACKI et al. 2020) and Eagle (DE GRAAFF et al. 2023), have also been used to study these relations further, aiming to better understand the physical processes involved and reduce the intrinsic scatter. We do not study the intrinsic scatter of standardised type Ia supernovae in the simulation, as the physical models of the key processes are simplified to focus primarily on the formation and evolution of galaxies.

It is commonly assumed that the residuals in the TF and FP relations are due to the coherent peculiar velocities plus some additional uncorrelated Gaussian random noise. In that case, the extra scatter acts as shot noise component in the velocity statistics, and typically reduces the constraining power without biasing the recovered cosmological parameters. However, if the intrinsic scatter is spatially correlated, due to dependencies with the local environment, it might bias the estimated peculiar velocity statistics. Previous studies have detected such correlations in the TF scatter (HUANG et al. 2024), focusing on the impact on weak lensing analysis using the IllustrisTNG simulation. FP correlated residuals have also been identified within the SDSS sample (JOACHIMI et al. 2015)

In this study, we use the MillenniumTNG hydrodynamical simulation of a $(500h^{-1}\mathrm{Mpc})^3$ volume to investigate the intrinsic scatter in the TF and FP relations and their impact on the peculiar velocity clustering. In section 6.1, we present the MillenniumTNG simulation and our definitions of elliptical and spiral galaxies. In section 6.2, we build the TF and FP relations in order to estimate the peculiar velocity of our sample, and measure the corresponding clustering. In section 6.3 we study the residuals distributions and correlations with the large scale environment. Finally, in section 6.4 we explore the galaxy formation history to give a physical interpretation to our results.

This work results from a one-month stay and was made possible thanks to the collaboration with Raúl Angulo at the Donostia International Physics Center.

6.1 Mock galaxy catalogue

In order to study how the intrinsic scatter in the TF and FP relations affects the peculiar velocity measurements, we need to reproduce realistic samples of young spiral star forming galaxies and older quenched elliptical galaxies, respectively.

6.1.1 The Millennium TNG simulation

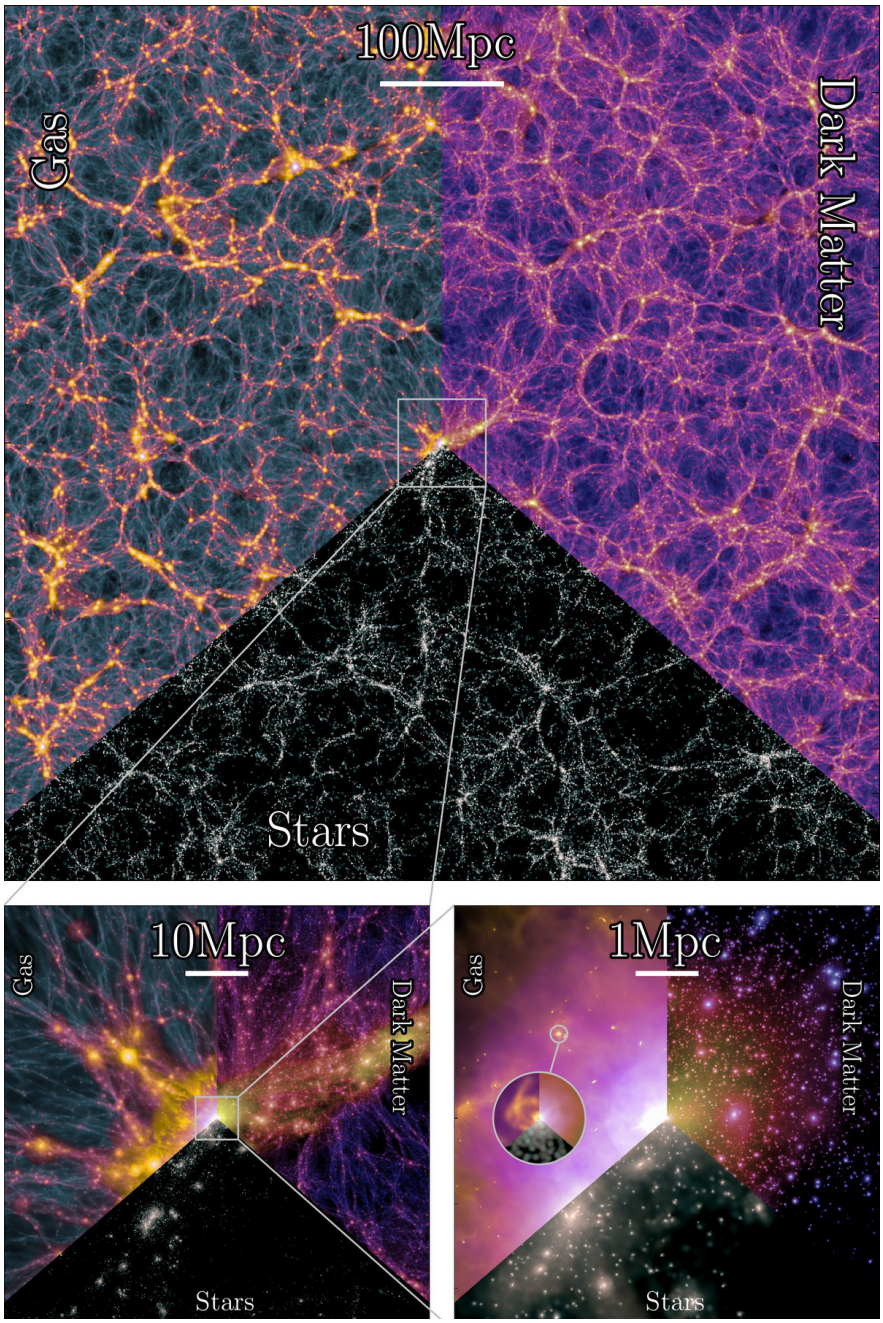


FIGURE 6.1 : Output of the $(500h^{-1}\text{Mpc})^3$ volume of the MileniumTNG hydrodynamical simulation. The different cuts describe the distribution of dark matter, gas, and stars at redshift $z = 0$, projected in a slice of $10h^{-1}$Mpc. Figure taken from PAKMOR et al. 2023.

The MilleniumTNG (MTNG) simulations have been used to study galaxy formation and evolution, and its impact on different probes of the large scale structure (PAKMOR et al. 2023; HERNÁNDEZ-AGUAYO et al. 2023; BARRERA et al. 2023; KANNAN et al. 2023; BOSE et al. 2023; HADZHIYSKA et al. 2023b). This suite of simulations is designed to extend the Millennium (SPRINGEL et al. 2005) and IllustrisTNG (PILLEPICH et al. 2018; MARINACCI et al. 2018; SPRINGEL et al. 2018; NELSON et al. 2018; NAIMAN et al. 2018) projects. It aims to provide an accurate description of the galaxy-halo connection and the impact of baryonic physics on clustering, within large volumes, up to $(500h^{-1}\mathrm{Mpc})^3$. This state-of-the-art hydrodynamical suite also includes a series of dark-matter-only simulations, with the effective volume boosted using variance suppression technique (ANGULO et PONTZEN 2016a). The N-body simulation can be populated with galaxies by applying semi-analytical models calibrated on hydrodynamical results.

We will use the largest hydrodynamical periodic box with 500 h^{-1}Mpc on a side, focusing on the clustering on scales below $\sim 60h^{-1}$Mpc, which are weakly affected by cosmic variance. This simulation evolves 4320^3 particles to describe both the dark matter and baryon fluid, leading to mass resolutions of $1.7\times10^8 M_\odot$ and $3.1\times10^7 M_\odot$ respectively. This simulation, run on the moving-mesh code AREPO (SPRINGEL 2010), accounts for various baryonic effects such as radiative cooling and star formation in the gas, the growth of supermassive black holes, feedback processes from supernovae and black holes. The simulation identifies dark matter halos and sub-halos, keeping track of the merging histories and larger systems' formation over time. Galaxies are identified as agglomerations of stellar particles within subhalos. Since we have access to various galaxy properties, we do not compute any post-processing quantities. The cosmological model is the same as the one used for the IllustrisTNG (PLANCK COLLABORATION et al. 2016).

Figure 6.1 depicts the resulting distributions of dark matter, gas, and stars at redshift $z = 0$ within a thin slice of $10h^{-1}$Mpc. The different panels show the range of scales and structures described, from the full box to an individual galaxy in a radius of 50kpc. Since errors on the peculiar velocity measurements scale with redshift, we are only interested in building the TF and FP relations at redshift $z = 0$, and we do not consider the redshift evolution of those relations

6.1.2 Sample selection

We are interested in the galaxy kinematic properties to select spiral and elliptical galaxies, and establish the TF and FP relations. The accuracy on measured kinematic properties of simulated galaxies is limited by the mass resolution. Therefore, we define our galaxy sample as well-resolved subhalos with a stellar mass larger than $10^9 M_\odot$ and at least 1000 stellar particles. The TF relations holds for late type galaxies (LTG), defined as spiral, star-forming blue galaxies, with the distribution of stars mainly supported by rotation. The FP relation is valid for early type galaxies (ETG), defined as elliptical, quiescent red galaxies, primarily supported by the stellar velocity dispersion. In order to select star-forming galaxies, we use the specific star formation rate (sSFR) defined as

$$\mathrm{sSFR} = \frac{\mathrm{SFR}}{M_*}, \tag{6.1}$$

with SFR the star formation rate in $M_\odot$ Gyr^{-1}, and M_* the total stellar mass in $M_\odot$. We further define, for the stellar population of a given galaxy, the dimensionless rotational to kinetic energy

ratio

$$\kappa_{\rm rot} = \frac{\sum_n \frac{1}{2} m_n \left(\frac{j_n}{r_n}\right)^2}{\sum_n \frac{1}{2} m_n \mathbf{v}_n^2}, \tag{6.2}$$

where we sum over stellar particles n, with mass m_n, angular momentum (in the direction of the minor axis of the galaxy's shape) j_n, distance to the center of the galaxy r_n, and velocity in the center-of-mass frame v_n We define LTG as galaxies with sSFR> 0.04Gyr^{-1} (above the main sequence) and $\kappa_{\rm rot} > 0.5$, and ETG as galaxies with sSFR< 0.04Gyr^{-1} (below the main sequence) and $\kappa_{\rm rot} < 0.5$.

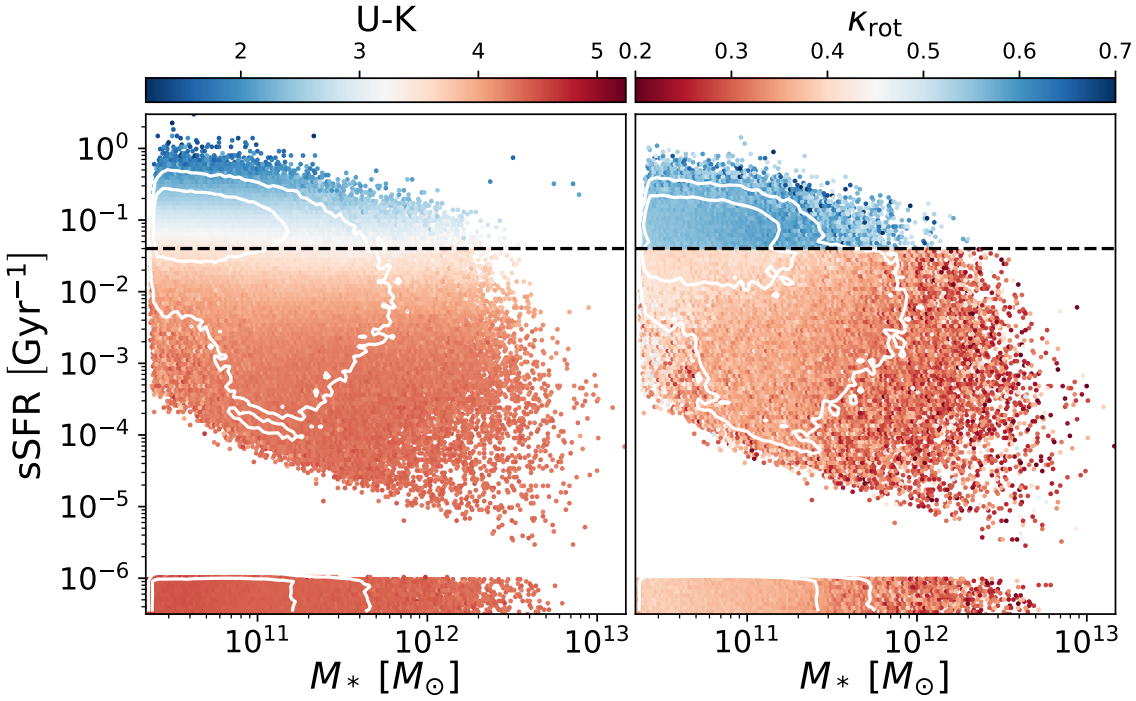


FIGURE 6.2 : Specific star formation rate distribution as a function of stellar mass for the galaxy samples. The left panel shows the total galaxy samples colour coded with the U-K colour. The black dotted line indicates the sSFR cut around 0.04 Gyr^{-1} to applied to define the star forming and quiescent galaxies. The white contours indicate the one and two sigma density distribution. The right panel gives the selected ETG and LTG samples obtained after applying the additional selection cut on the rotational support $\kappa_{\rm rot}$.

Figure 6.2 describes the galaxy sample' sSFR distribution as a function of stellar mass. The left panel corresponds to the total galaxy sample. The black dotted line indicates the separation between star-forming and quiescent galaxy, with the selection cut at sSFR = 0.04Gyr^{-1}. The binned density distribution is indicated by the white contours (68 and 95 %). The colour bar designs the $U-K$ colours, indicating that the star-forming galaxies are bluer than the quiescent ones. The right panel shows the resulting ETG and LTG samples - colour coded with the value of the rotational to kinetic energy ratio - after applying the additional selection cut around $\kappa_{\rm rot} = 0.5$. We end up with 271358 ETG to build the FP relation, and 84787 LTG for the TF relation. Following

the average $n(z)$ distribution forcasted for the upcoming DESI TF and FP sample(SAULDER et al. 2023a), we place the obsrever to an effective redshift of $z_{\rm FP} = 0.08$ of the ETG sample, and $z_{\rm TF} = 0.04$ of the LTG sample.

6.2 Peculiar velocity measurement

We have two galaxy catalogues with known positions, velocity and physical properties. For our galaxy catalogues, we want to study the residuals between the true and estimated velocities. We first build the TF and FP relations to estimate the log distance-ratio. In the following we will present our results using the colour coding : blue for LTG and red for ETG.

6.2.1 Tully-Fisher and fundamental plane relations

As described in section 2.2, LTG follow the Tully-Fisher relation, describing the correlation between their absolute magnitude M and their maximum rotation velocity $V_{\rm max}$ (in km.s^{-1}). While in realistic surveys only the LoS rotation velocities can be measured, in this simulation study we compute $V_{\rm max}$ from the spherically-averaged rotation curve. We adopt the convention $v_{\rm max} = \log V_{\rm max}$, and use the r-band magnitudes M_r. The TF relation is usually calibrated through a linear regression between M_r and $v_{\rm max}$ (Eq 2.65) on a reduced sample of galaxies with known absolute magnitudes. In this study, as we know M_r for every galaxy, we use the full sample to calibrate the TF relation, suppressing undesirable selection effects. Moreover, as the $(M_r, v_{\rm max})$ estimations are noise-free, we extend the model beyond the linear prescription in Eq 2.65, in order to increase the predictive accuracy. We build a non-parametric model $M_r^{\rm TF}(v_{\rm max})$ by interpolating over $v_{\rm max}$ the mean $\langle M_r \rangle_k$ in a given bin $v_{\rm max,\,k}$. For any spiral galaxy in our catalogue, we can now predict its absolute magnitude from its maximum rotation velocity using the TF relation $M_r^{\rm TF}(v_{\rm max})$.

The fundamental plane relation describes the correlation between the physical effective radius R (in h^{-1} kpc), central velocity dispersion σ_V (in km.s^{-1}) and surface brightness I (in $L_\odot$ kpc^{-2}) for ETG. Note that σ_V is computed as the velocity dispersion averaged over the three dimensions. We use the radius containing half of the total mass of the galaxy as an approximation for the 2D effective radius. While the surface brightness can be estimated as a distance independent quantity (Eq 2.69), we directly compute it as

$$I = \frac{L}{\pi R^2}, \tag{6.3}$$

with L the luminosity (in $L_\odot$) of the galaxy. In the following we adopt the conventions $r = \log R$, $s = \log \sigma_V$, and $i = \log I$. We follow the same methodology described previously for the TF relation to calibrate the FP. We use the full sample of galaxy - with known effective radius - to calibrate the relation, performing a 2D interpolation of the mean $\langle r \rangle_k$ in bins of $(s, i)_k$ The physical radius of any elliptical galaxy can then be predicted by measuring s and i.

Figure 6.3 shows the performance of our non-parametric TF and FP models predictions, evaluated for the same samples as the one used for the calibration. The left panel displays the true negative magnitudes' distribution against the ones predicted from the TF relation $M_r^{\rm TF}(v_{\rm max})$. The dashed line shows a one to one relationship and the black contours correspond to one, two and three sigmas levels of the distribution. We see that although the model is unbiased $\langle M_r - M_r^{\rm TF} \rangle \approx 0$, the TF relation is characterised by a scatter around the true magnitudes. The right

panel shows the true physical radius distribution against the FP predictions $r^{\,\mathrm{FP}}(s,i)$. Similarly to the TF, we find an unbiased FP relation $\langle r - r^{\,\mathrm{FP}}\rangle \approx 0$ characterised by a scatter around the true radius.

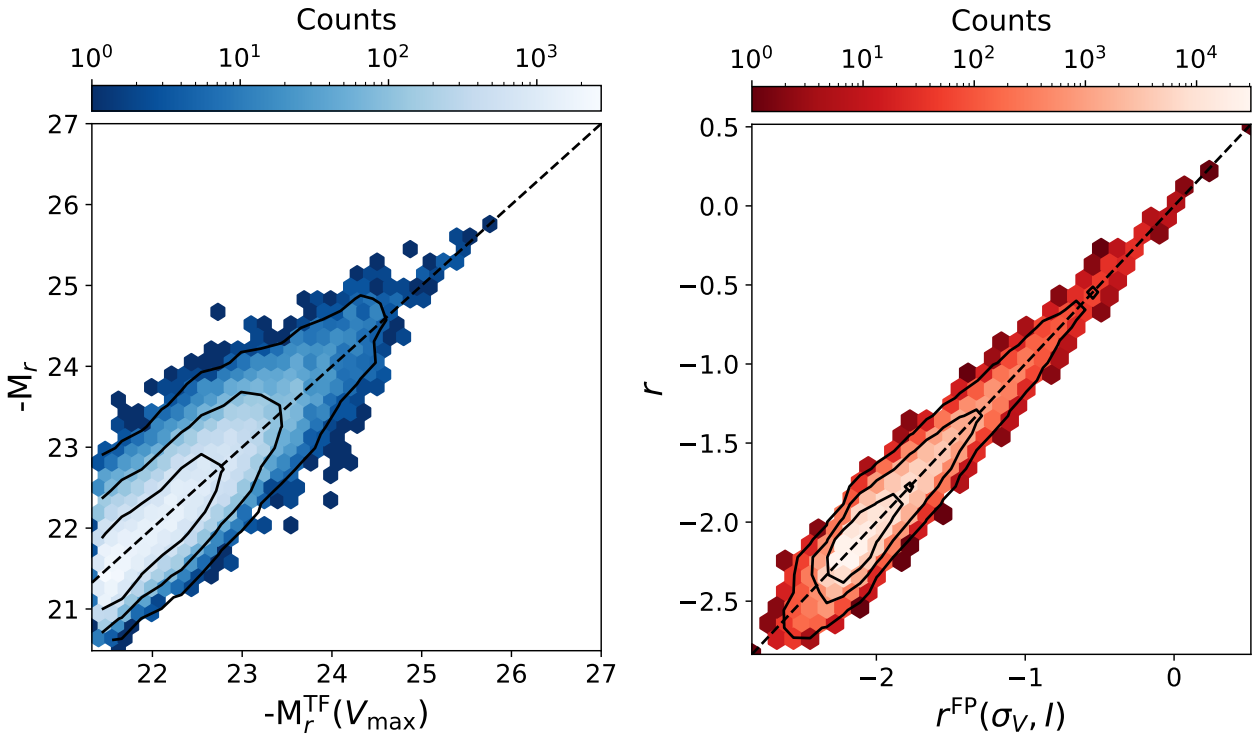


FIGURE 6.3 : TF and FP calibrated predictions. Left panel : negative true r-band absolute magnitudes against the ones predicted using the non-parametric TF relation. Right panel : true logarithmic effective radius against the ones predicted using the non-parametric FP relation. The black dashed lines show a one to one relation. The black contours correspond to regions containing 68, 95, and 99.7% of the distribution.

Now we reproduce the RSD effect on the observed galaxy sample, following Eq 2.30. Under the flat sky approximation the true cosmological redshift z_{cos} for the ETG and LTG is fixed at $z_{\mathrm{FP}} = 0.08$ and $z_{\mathrm{TF}} = 0.04$ respectively. The observed redshift is then

$$z_{\mathrm{obs}} = (1 + z_{\mathrm{cos}})(1 + v_p/c) - 1. \tag{6.4}$$

Following Eqs 2.66 and 2.70, we compute the apparent magnitude m_r and angular radius θ to derive the "observed" magnitudes $M_r(z_{\mathrm{obs}})$ and physical radius $r(z_{\mathrm{obs}})$ respectively. RSD induces a scatter in magnitude and physical radius, vertically shifting the point in figure 6.3 around the black dotted line, on top of the intrinsic scatter. From our TF and FP predictions, we get our log distance-ratio measurements

$$\eta_L^{\mathrm{obs}} = 0.2\left(M_r^{\mathrm{TF}} - M_r(z_{\mathrm{obs}})\right), \quad \text{and} \quad \eta_A^{\mathrm{obs}} = r(z_{\mathrm{obs}}) - r^{\,\mathrm{FP}}. \tag{6.5}$$

Because M_r^{TF} and $r^{\,\mathrm{FP}}$ are only noisy estimates of the true $M_r(z_{\mathrm{cos}})$ and $r(z_{\mathrm{cos}})$, the scatter

propagates in the log-distance ratio measurements such that

$$\eta_L^{\rm obs} = \eta_L + \lambda_{\rm TF}, \quad \text{and} \quad \eta_A^{\rm obs} = \eta_A + \lambda_{\rm FP}, \tag{6.6}$$

where we have defined the log-distance ratio residuals λ. For the TF sample, we find std($\lambda_{\rm TF}$) = 0.062, corresponding to a fractional error on the distance measurement of about 14%. The most recent TF catalogues (HONG et al. 2019; R. Brent TULLY et al. 2023) - considering various observational uncertainties and the relatively small size of the calibration sample - found an average scatter of around 22% in distance ratio. For the FP sample, we find std($\lambda_{\rm FP}$) = 0.075, corresponding to a fractional error on the distance measurement of about 17%, while the SDSS FP catalogue (SAULDER et al. 2013; HOWLETT et al. 2022) describes an average scatter of 23% in distance ratio. We want to study if this lack of information about the galaxy properties, affecting the magnitude and radius, is purely described by stochastic astrophysical processes, or if it correlates with the local environment and therefore affects the clustering measurements.

6.2.2 Peculiar velocity clustering

We measure the galaxy peculiar velocities $v_p^{\rm obs}$ from the log distance-ratios, following Eq 2.77, and define the corresponding velocity residual as

$$\lambda_v \equiv v_p^{\rm obs} - v_p \approx \begin{cases} \lambda_{\rm TF}/\alpha_L(z_{\rm obs}) & \text{for LTG} \\ \lambda_{\rm FP}/\alpha_A(z_{\rm obs}) & \text{for ETG} \end{cases} \tag{6.7}$$

Figure 6.4 shows the velocity distributions measured using the TF and FP samples in blue and red respectively, with the underlying true distribution in black We find a poor SNR, smaller than one in average as the residuals λ_v are large compared to the true velocities. The noise in the velocities is larger in the ETG sample for two reasons. First, the intrinsic scatter is larger for the FP than the TF relation $\lambda_{\rm FP} > \lambda_{\rm TF}$. Second (dominant error), the FP sample is located in average at a larger redshift, increasing the velocity errors for a given scatter.

We measured the velocity auto-correlation and the velocity-galaxy cross-correlation functions (Eq 2.96) for the LTG, ETG, and combined samples, using respectively the velocities estimated from our TF and FP relations, and the true velocities of the simulations. The clustering is averaged over three LoS - rotating the box - in order to reduce cosmic variance (SMITH et al. 2021). Figure 6.5 displays the resulting TF and FP clustering on the left (in blue) and right (in red) sides respectively. The top, middle, and bottom panels show the measurements for ξ_0^{vv}, ξ_1^{vg}, and ξ_0^{vg}. The clustering using the combined samples with the true velocities is shown in black. The shaded area corresponds to the standard deviation - 30 realisations for each LoS - of the clustering measured after applying a Gaussian random noise on top of the true velocities, with a scatter corresponding to std($\lambda_{\rm TF}$) = 0.062 and std($\lambda_{\rm FP}$) = 0.075. For ξ_0^{vv} of the TF and FP samples, we find that on the largest scales the scatter averages out, leading to unbiased estimate of the TPCF. However, on small scales the realistic measurements significantly deviate from the expected clustering, exceeding 1σ of the random noise at $s \sim 1$ h^{-1}Mpc for the TF sample, and $s \sim 10$ h^{-1}Mpc for the FP sample. For the cross-correlation dipole ξ_1^{vg}, the deviation from the excepted clustering is quite consistent with random noise, barely reaching 1σ even for the smallest scales. The cross-correlation monopole ξ_0^{vg} of the true sample is consistent with zero. However, we find a strong signal in the TF and FP samples, exceeding 3 and 5 σ respectively.

Those results provide valuable information on the residuals properties. According to Eq 6.7,

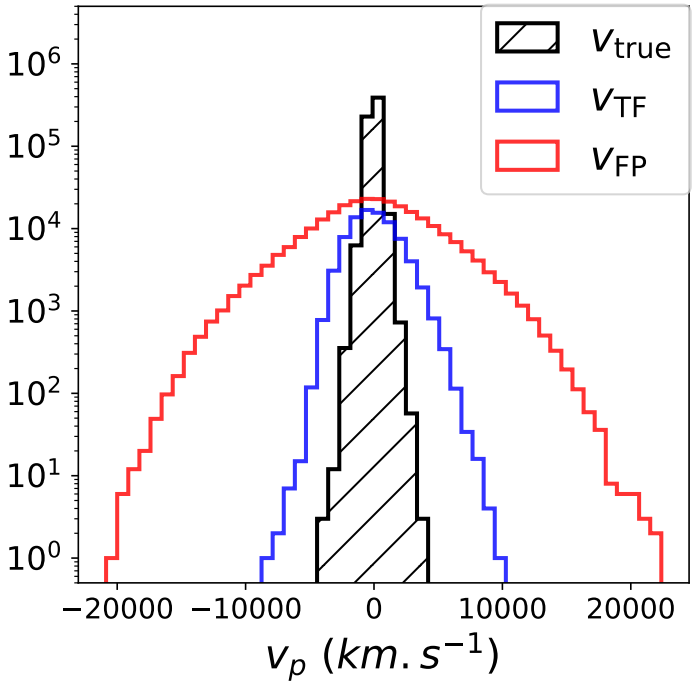


FIGURE 6.4 : Radial velocity distribution of the galaxy samples. The blue and red histograms correspond to the measured velocities for the LTG and ETG samples respectively. The true distribution of the combined samples is shown in black.

the observed velocity two-point functions can be written

$$\begin{aligned} \xi^{v^{\rm obs}v^{\rm obs}} &= \xi^{vv} + \xi^{\lambda_v\lambda_v} + \xi^{v\lambda_v} + \xi^{\lambda_v v}, \\ \xi^{v^{\rm obs}g} &= \xi^{vg} + \xi^{\lambda_v g}, \end{aligned} \tag{6.8}$$

where we dropped the explicit spatial dependencies for conciseness. Compared to the expected result, the observed ξ_1^{vg} indicates that the dipole of the residual-galaxy TPCF vanishes $\xi_1^{\lambda_v g} \approx 0$. The correlated signal lies in the monopole, since $\xi_0^{vg} \approx 0$, we get that $\xi_0^{\lambda_v g} \neq 0$. This demonstrates that the residuals are not properly described by random Gaussian noise, but actually correlate with the local environment. In principle, $\xi_0^{\lambda_v g}$ should also be observed on real data, without knowing the true velocities since $\xi_0^{v^{\rm obs}g} \approx \xi_0^{\lambda_v g}$. While this multipole is not used in cosmological inferences, it could be quite convenient to quantify the amount of correlated residuals. Finally, in the monopole $\xi_0^{v^{\rm obs}v^{\rm obs}}$ we find smaller but non-vanishing signals, for scales typically below 1 and 10 h^{-1}Mpc for TF and FP samples. While we do not show the results here, we also measured the cross correlation $\xi^{v\lambda_v}$ and found $\xi_0^{v\lambda_v} \approx 0$, indicating that all the signal of the monopole is contained in the residuals' autocorrelation $\xi_0^{\lambda_v\lambda_v} \neq 0$. Overall the correlated signal from the residuals λ_v is stronger and affects larger scales in the FP sample compared to the TF sample. These considerations would be compatible with a residual field written as a functional of the matter (or galaxy) field

$$\lambda_v(\boldsymbol{x}) = \mathscr{F}\left[\delta_m(\boldsymbol{x})\right], \tag{6.9}$$

or as a perturbative bias expansion, as it has been proposed in JOACHIMI et al. 2015. We study

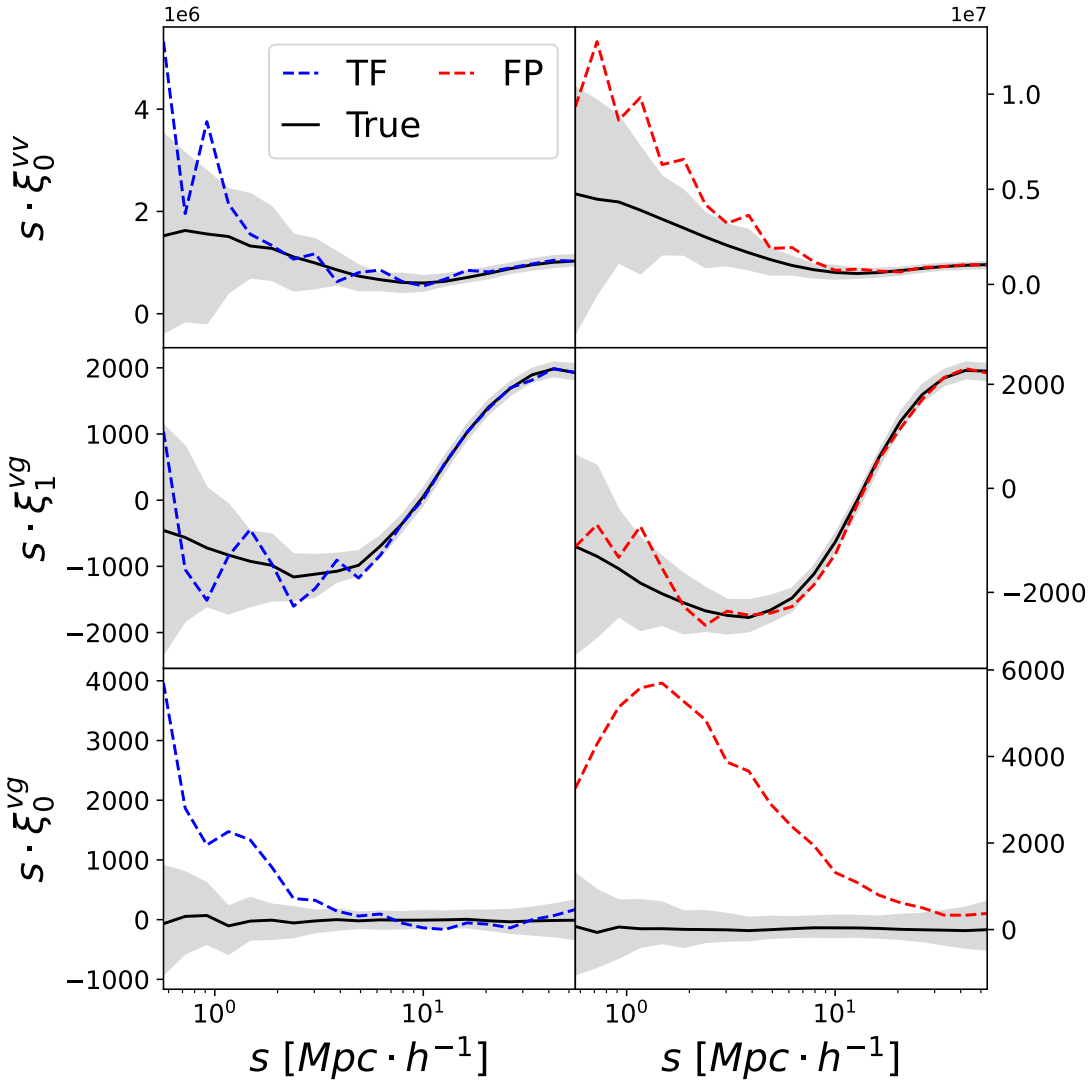


FIGURE 6.5 : Velocity clustering statistics measured using the TF (blue), FP(red) and true(black) samples. The clustering measurements are averaged over three LoS by rotating the box. The shaded area correspond to 1σ deviation from the expected clustering. They are obtained by applying to the true velocities some random Gaussian noise (30 realisations for each LoS), with standard deviations of std(λ_{TF}) = 0.062 and std(λ_{FP}) = 0.075 in the left and right panels respectively.

in the next section how the different residuals depend on the galaxy properties and their local environments.

6.3 Dependency of residuals with environment

The TF and FP physical description is rooted in the dynamics of gravity and galactic rotation of LTG, and the serial merger with smaller galaxies undergone by ETG. Both relations rely on the Virial theorem (VT) to predict the mass or size of a galaxy, given some structural parameters.

The brightness or effective radius can be derived assuming a power-law for the luminosity-mass function. The free parameters used for calibration describe deviation from VT predictions arising from energy dissipation during merger, mass loss or different feedback.

One hypothesis to explain the correlated scatter in TF and FP relations seen in the previous section is as follows : The structural properties (matter and velocity profiles, stellar age, etc) of the galaxies we observe today depends on different physical processes undergone through time, and are therefore correlated with their local environments. Assuming full knowledge of a galaxy's structural properties, any luminosity-related quantity — such as absolute magnitude or effective radius — should be predictable, up to some stochastic variations originating from astrophysical processes. However, the standard TF and FP relations only rely on the information provided by $V_{\rm max}$ and σ_V (along with the intensity), which does not completely describe the structural properties, leading to residuals correlated with small scales cosmological fluctuations. We test this hypothesis by studying the residuals in different subsamples.

6.3.1 Properties dependency

We use the complete sample of galaxies - before applying the sSFR and $\kappa_{\rm rot}$ cuts - to define our environmental indicator as the galaxy over density. We first assign each galaxy in a 256^3 mesh, smooth the galaxy number density $n_g(\boldsymbol{x})$ using a spherical top hat kernel with a characteristic length-scale of about $5h^{-1}$Mpc, and compute the resulting over density $\delta_g(\boldsymbol{x}) = n_g(\boldsymbol{x})/\bar{n} - 1$. Note that the following results are roughly unchanged for choices of length-scales between 2 and $15h^{-1}$Mpc. We further decompose the TF and FP samples into central and satellite galaxies, as they form and live in different environments that may affect the TF and FP relations residuals.

Figure 6.6 gives the mean TF and FP residuals as functions of $1+\delta_g$. The residuals for central galaxies are depicted with circles, while those for satellites are shown with diamonds, with errorbars corresponding to the standard deviation divided by the square root of the number counts in each bin. The TF and FP results are displayed in the top and bottom panels, respectively. The histograms represent the TF and FP over-density distribution, and the shaded areas indicate the standard deviation of the residuals $\mathrm{std}(\lambda_{\rm TF}) = 0.062$ and $\mathrm{std}(\lambda_{\rm FP}) = 0.075$. Although it only partially explains the scatter amplitude, significant correlation between residuals and the galaxy environments. Beside the dependence on the local over-density, particularly for FP, centrals and satellites occupy different regions of the TF and FP relations, so they have slightly different average residuals. For LTG, satellites lie above the TF relation, resulting in positive residuals, and therefore an underestimation of the absolute brightness and peculiar velocity (one can write $\lambda_{\rm TF} = 0.2(M_r^{\rm TF} - M_r^{\rm true})$). Conversely, for central galaxies, the brightness (and velocity) is overestimated. For ETG in the bottom panel of figure 6.6, we find a clear trend in the residuals which increase with density. On average, satellite galaxies have negative residuals and lie below the FP relation, resulting in an overestimation of the physical radius and velocity (one can write $\lambda_{\rm FP} = (r_{\rm true} - r_{\rm FP})$). In the highest density environments, the average residual for the ETG satellites tends towards zero. For central galaxies, the radius is overestimated in under-dense regions and underestimated in over-dense regions. Because central and satellite galaxies have different properties distributions, a small portion of the scatter amplitude is described by splitting the samples. However, it does not remove the correlation with the local environment.

We further investigate the dependencies of the residuals on several galaxy properties. It has been established that using the total baryonic mass instead of $V_{\rm max}$ to describe the structural properties of LTG can reduce the intrinsic scatter of the relation (Baryonic TF, McGaugh et al.

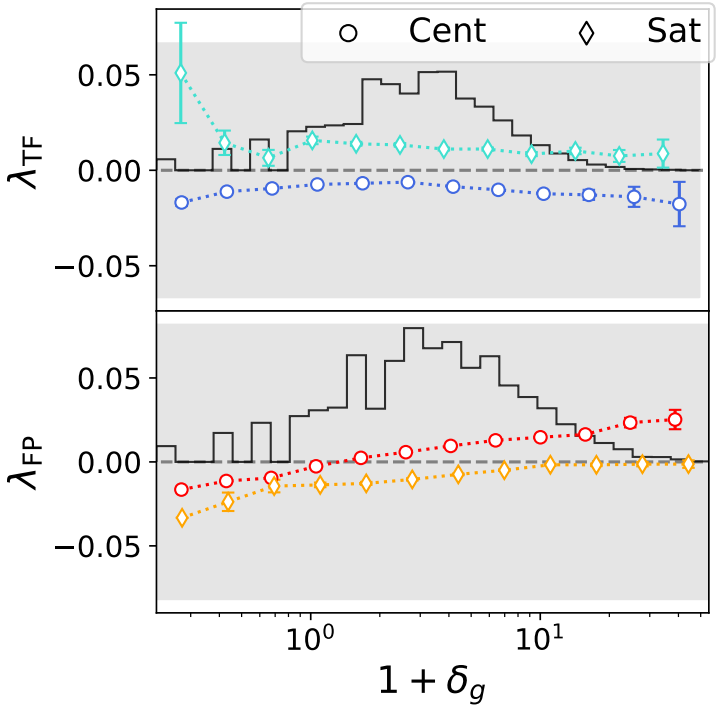


FIGURE 6.6 : Environmental dependence of the residuals in log-distance-ratio. The top panel shows the results obtained for the TF residuals $\lambda_{\rm TF}$. The bottom panel shows the results obtained for the FP residuals $\lambda_{\rm FP}$. The samples are decomposed into central and satellite galaxies, corresponding to circles and diamonds. The shaded areas indicate the standard deviation of samples residuals std($\lambda_{\rm TF}$) = 0.062 and std($\lambda_{\rm FP}$) = 0.075.

2000). We split the TF central and satellite samples in three $v_{\rm max}$ bins with equal percentiles, and study the residuals' correlation with density for different bins of stellar mass or SFR. We proceed analogously for the FP centrals and satellites, dividing the samples in bins of s, and studying the residuals-density correlation for different bins of stellar mass and sSFR.

Figure 6.7 shows the corresponding results for the LTG in blue and ETG in red. We follow the same convention as before, using circles and diamonds for the centrals and satellites residuals scatter plots, with the shaded area indicating the total samples standard deviations std($\lambda_{\rm TF}$) and std($\lambda_{\rm FP}$). In every sub-figure, three panels display from bottom to top the three different bins of increasing values of $V_{\rm max}$ (for TF) and σ_V (for FP).

Let us first start by analysing TF results. We retrieve the expected correlations between $V_{\rm max}$ and M_* or SFR, where the higher mass and SFR bins are included in the larger $V_{\rm max}$ bins. The main part of the residual amplitude can be explained by the unaccounted dependence of the magnitude on stellar properties. Indeed, we find that for galaxies with the larger (smaller) stellar mass, the inferred magnitudes are underestimated (overestimated), and similarly for SFR. Furthermore, the central and satellite residual distributions are more consistent with each other when grouped by stellar mass or SFR values. The residuals grouped by such values do not show strong dependency with δ_g. The remaining small correlations could come from the M_* and SFR distributions within the considered bins. In that case, the TF residuals dependence on the local environment might indicate that, for a given $V_{\rm max}$, the stellar properties of LTG (and thus their brightness) still depend on small-scale cosmological fluctuations. Therefore, a generalised TF

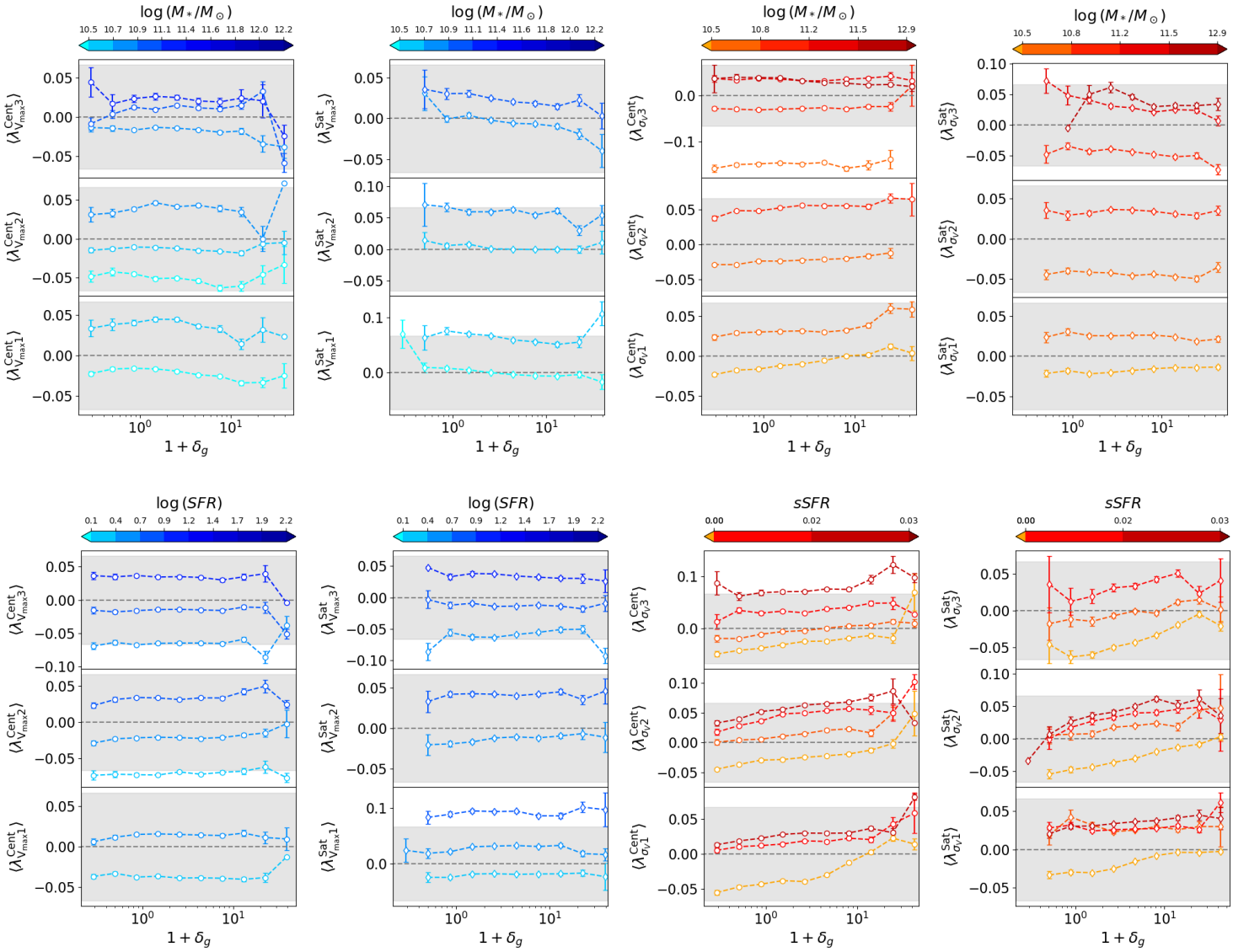


FIGURE 6.7 : Environmental dependence of the average residuals in log-distance-ratio in several bins of $V_{\rm max}$ and σ_V, for different galaxy properties. The results for the TF and FP relations are shown in blue and red respectively. The circles and diamonds scatter plots correspond to the central and satellite galaxies respectively. The shaded areas indicate the standard deviation of samples residuals std($\lambda_{\rm TF}$) = 0.062 and std($\lambda_{\rm FP}$) = 0.075. In every sub-figure, the bins of $V_{\rm max}$ and σ_V are stacked by increasing order.

relation $M_r^{\rm TF}(V_{\rm max}, M_*, SFR)$ should reduce the correlated signal in the residuals seen in figure 6.5. For the FP residuals on the right side of figure 6.7, we find similar results. Note that for FP we group galaxies by their specific star formation rate rather than just the SFR to reduce the skewness of the distribution due to the large number of quenched galaxies (SFR=0). The residual amplitude and the discrepancy between satellite and central galaxies are mainly explained by the stellar mass and sSFR distributions within bins on σ_V. However, some correlations with the environment remain on some bins, especially when grouping by sSFR values. Again, assuming for a given galaxy the internal properties of interest depends on the local environment even for a

fixed value of σ_V, the remaining correlations could simply arise from the large width of the sSFR bins.

6.3.2 Extended TF and FP relations

In this section we study an extension of the parameter space for the TF and FP relations, including additional properties in order to better capture the galaxies' structural properties relevant to predict brightness and effective radius. As previously, we use the full samples of LTG and ETG to calibrate non-parametric models, interpolating the mean magnitude or radius in bins of various properties. We build the extended models $M_r^{\mathrm{TF}}\left(V_{\mathrm{max}}, M_*, M_{\mathrm{gas}}, \mathrm{SFR}\right)$ and $r^{\mathrm{FP}}\left(\sigma_V, I, M_*, M_{\mathrm{gas}}, \mathrm{sSFR}\right)$, including additional information on the stellar mass M_*, the gas mass M_{gas}, and the (s)SFR. We use the logarithm of every quantity except the sSFR. We remind that for real data, the TF and FP relations should be calibrated using distance-independent properties to be consistent for diverse samples without introducing systematic biases. Thus, those extended models are unlikely to be applied to real observations.

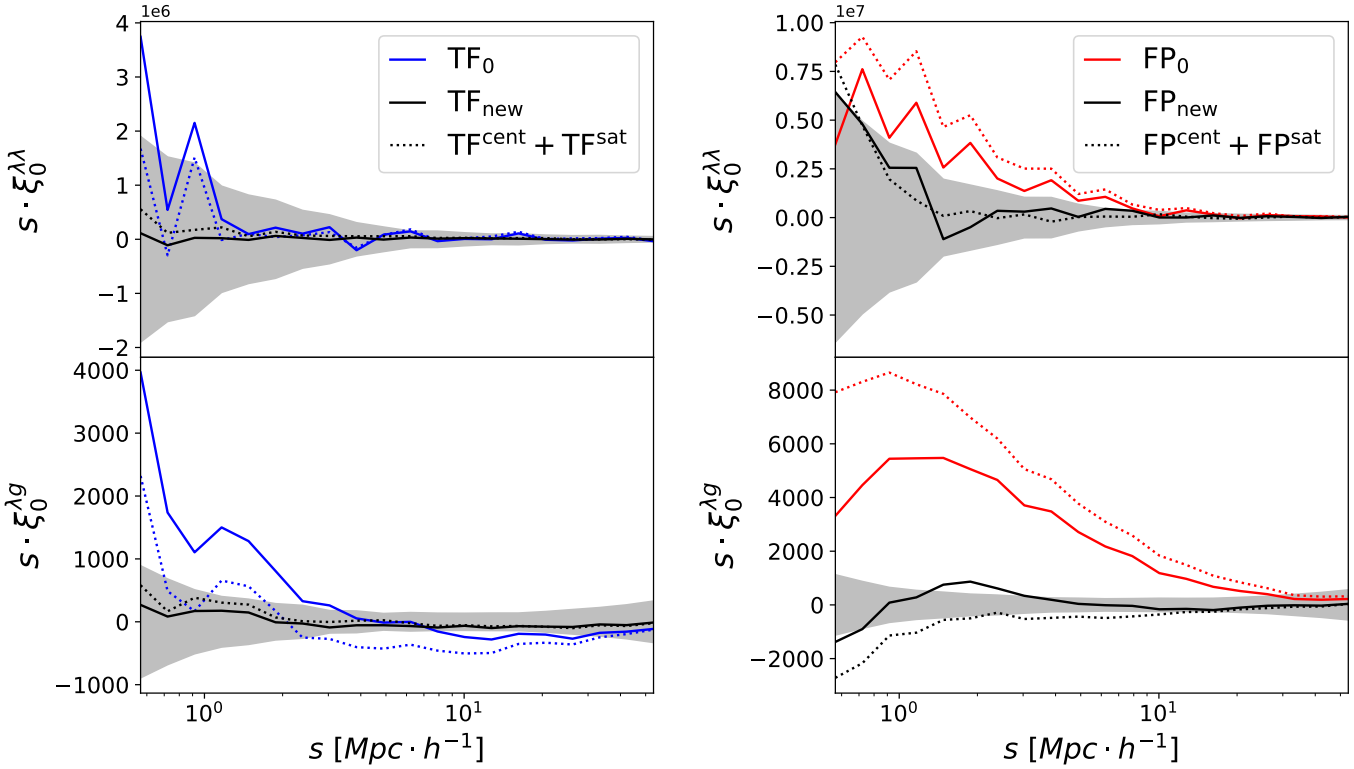


FIGURE 6.8 : Clustering of the residuals λ obtained with the previously used ("0") and extended ("new") TF and FP relations. The LTG and ETG clustering are shown in the left and right panels respectively. The dotted lines correspond to the clustering obtained by calibrating the relations on central and satellite galaxies separately. The grey shaded areas indicate the standard deviation for the clustering of 30 realisations of random Gaussian noise for each LoS corresponding to the scatter of the standard TF and FP : $\mathrm{std}(\lambda_{\mathrm{TF}}) = 0.062$ and $\mathrm{std}(\lambda_{\mathrm{FP}}) = 0.075$.

Figure 6.8 compares the clustering of the residuals λ (in long-distance ratio) obtained using

	Standard	Extended
$\mathrm{std}(\lambda_{\mathrm{TF}})$	0.062	0.022
$\mathrm{std}(\lambda_{\mathrm{TF}}^{\mathrm{cent}})$	0.062	0.026
$\mathrm{std}(\lambda_{\mathrm{TF}}^{\mathrm{sat}})$	0.070	0.029
$\mathrm{std}(\lambda_{\mathrm{FP}})$	0.075	0.055
$\mathrm{std}(\lambda_{\mathrm{FP}}^{\mathrm{cent}})$	0.081	0.052
$\mathrm{std}(\lambda_{\mathrm{FP}}^{\mathrm{sat}})$	0.072	0.047

TABLE 6.1 : Standard deviation of the TF and FP residuals. We show the results obtained when calibrating the relations separately for central and satellite galaxies, using the standard and extended models.

our standard and extended TF and FP relations. The dotted lines correspond to the clustering obtained by applying the relations on central and satellite galaxies separately. The LTG and ETG clustering are shown in the left and right panels respectively. The grey shaded areas correspond to 1σ scatter for the clustering of random Gaussian residuals with $\mathrm{std}(\lambda_{\mathrm{TF}}) = 0.062$ and $\mathrm{std}(\lambda_{\mathrm{FP}}) = 0.075$. Once again the TPCF are averaged over three LoS by rotating the box. For every calibrated relation, the corresponding standard deviation of the residuals is given in table 6.1. The extended TF and FP models, accounting for additional galaxy properties, significantly reduce the scatter amplitude and suppress the correlation signal, especially for the TF residuals.

Building separate relations for centrals and satellites leads to overall larger scatter in the residuals, except for the extended FP. Yet, for the standard TF model, the correlated signal is reduced on the smallest scales. For the extended TF model, the clustering remains almost unaffected. For the FP, splitting the relation for central and satellite galaxies surprisingly enhances the amount of correlation in the residuals, both for the standard and extended models. Due to their distinct evolutionary paths, central and satellite galaxies should exhibit significantly different structural properties (at a fixed V_{max} or (σ, I)). However, it appears that these discrepancies do not correlate with cosmological fluctuations on the considered scales. In order to validate our hypothesis, we now study how the environment affects the evolutionary history of the different types of galaxies considered so far.

6.4 Star formation history

The MillenniumTNG simulation provides subhalo merger trees, enabling us to trace the evolution of galaxies back in time. We examine how the amounts of gas, dark matter, and stars evolve over time for various types of galaxies. Specifically, we select the top 10% and bottom 10% of ETG and LTG residing in the highest and lowest density environments, respectively. Our goal is to understand how the large-scale structure environment affects galaxy properties beyond the standard quantities of the TF and FP relations. To achieve this, we only keep LTG with a v_{max} within $[2.35, 2.38]$, and ETG with s within $[2.12, 2.19]$, corresponding to the 40 to 60th percentile range of their distribution. We further separate our samples into satellites and centrals, resulting in four subsamples for both ETG and LTG. We trace back the evolution of 100 galaxies in each subsample.

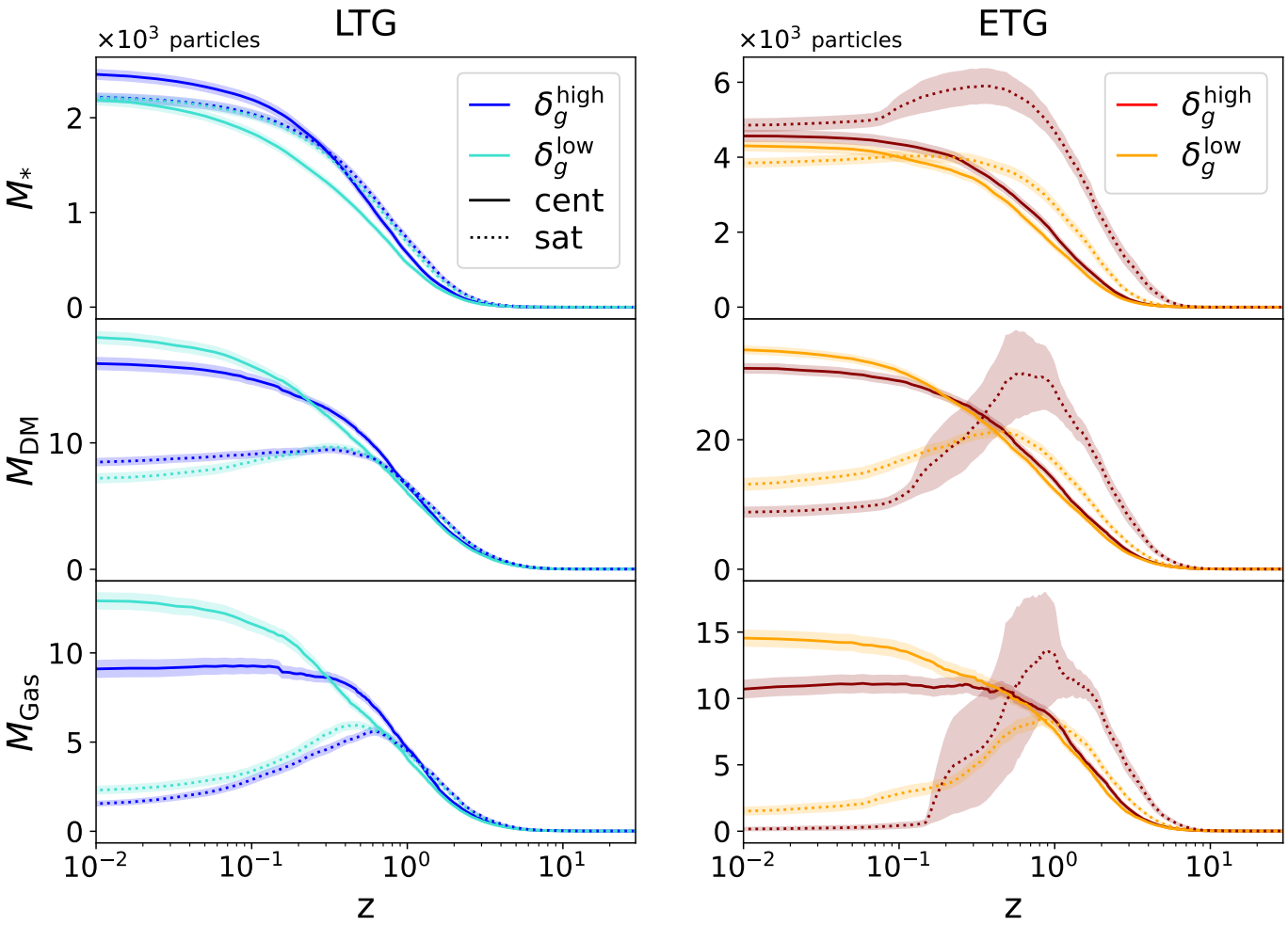


FIGURE 6.9 : Redshift evolution of gas, dark matter, and stars (in number of particles) within several types of galaxies. Galaxies are selected within narrow bands of $v_{\rm max}$ and s, for LTG(left) and ETG(right) respectively. The lines and shaded areas correspond to the median and standard deviation over 100 galaxies selected in each subsamples. Darker (lighter) colours give the evolution of galaxies living in high (low) density environments. Solid and dotted lines indicate the results for central and satellite galaxies respectively. The top, middle, and bottom panels show the evolution of stellar, dark matter, and gas masses.

Figure 6.9 shows the median evolution of stellar, gas, and dark matter masses, for LTG (left) and ETG (right) in the narrow bands of $v_{\rm max}$ and s. The shaded areas correspond to the standard deviation over the 100 galaxies in each subsample. Solid and dotted lines indicate the results for central and satellite galaxies respectively. Darker (lighter) colours give the evolution of galaxies living in high (low) density environments. These growth histories emphasize that ETG are statistically more massive and quiescent than LTG, as they have lost most of their gas, thereby stopping their star formation process. Interestingly, for all quantities, there is a clear distinction between galaxies living in high and low-density environments.

For both ETG and LTG, central galaxies living in dense regions formed and started accreting mass earlier. As they formed more stars, a larger amount of their gas reservoir is consumed compared to centrals living in lower density environments. Due to hierarchical clustering, large halos form by accreting smaller subhalos that become satellite galaxies. Both ETG and LTG satellites tend to form earlier than central galaxies, but their star formation is quenched more rapidly.

In general, the evolution of satellites is more complicated than centrals due to processes that cause galaxies to lose their gas and undergo structural changes (WETZEL et M. WHITE 2010; WATSON et al. 2012; CHAVES-MONTERO et al. 2016), especially in dense environments. Both baryonic and dark matter can be stripped away by tidal forces exerted by clusters or close encounters with massive galaxies, which can also distort the galaxy morphology and trigger bursts of star formation. Additionally, ram pressure exerted by the intracluster medium filled with hot, dense gas can strip away gas from the galaxies. It is important to note that the realism of these processes in hydrodynamical simulations is uncertain, as they are still not fully understood. We find that the evolution of satellites in different density bins is quite similar for LTG. Although for a given $v_{\rm max}$ the amount of gas and dark matter vary slightly, the final stellar mass is nearly independent of the local density fluctuations.

The evolution of satellites differs for ETG. Satellite galaxies start forming stars earlier and at a faster rate in denser regions, reaching larger M_* than central galaxies. While initially larger, the amounts of dark matter and gas decrease more significantly for satellite galaxies residing in high-density environments, causing them to stop forming stars earlier than those in low-density regions. As they typically undergo more intense processes such as mergers, the infall of material towards large gravitational potentials is stronger, and satellites in high-density environments can lose a significant amount of stars. In the end, for a fixed velocity dispersion, satellite galaxies in high-density environments will have smaller $M_{\rm Gas}$ and $M_{\rm DM}$ and larger M_*.

Overall, at fixed s or $v_{\rm max}$, the evolution of the mass components for different types of galaxies is correlated with the local environment. These components, particularly M_*, are crucial for deriving the absolute magnitude and effective radius of a galaxy. The standard structural parameters used in the TF and FP relations capture only at zeroth order the star formation history (SFH) and its large-scale correlations.

6.5 Conclusion

In this chapter, we studied the intrinsic scatter of the TF and FP relations and the systematic errors induced in the measured peculiar velocities and their clustering. We reproduced these relations using early- and late-type galaxies within the MillenniumTNG hydrodynamical simulation. We found that the TF and FP residuals correlate on small scales with the matter density field. Additionally, we observed opposite trends between centrals and satellites. On average, for LTG, the brightness and peculiar velocity are underestimated for satellites and overestimated for centrals. The contrary was found for ETG, where the size and velocity are overestimated for satellites and underestimated for centrals. We observed that for fixed bins of maximum rotational and dispersion velocities, the distribution of star formation rate and stellar mass could account for a portion of the residual scatter, without completely removing the correlation with the local environment, especially for the FP residuals. To suppress these correlations, we developed extended TF and FP relations that include gas and stellar masses along with the SFR to describe the galaxies' magnitudes and effective radii. These new relations reduced the amount of correlated

signal in the residuals, especially for LTG. However, calibrating separate relations for the central and satellite samples proved to be less effective.

Those results indicate as misdescription of galaxy structural properties. Tracing back the galaxies evolutionary history, we observed that even for fixed $V_{\rm max}$ or σ_V, some internal properties (mainly SFH) affecting the brightness or size depend on the local density fluctuations. The commonly TF and FP structural quantities ($V_{\rm max}, \sigma_V, I$) are not sufficient to describe the galaxies properties, and do not fully capture their correlations with the local environments. This detection has major implications for measurements relying on TF and FP distance indicators for the purpose of peculiar velocity studies. If not properly accounted for, the systematic errors induced in the non-linear velocity clustering could bias cosmological analysis, and the smallest scales should not been used without a thorough understanding of the underlying galaxy samples and physical processes.

Since ETG and LTG are only relative standard objects, the calibration of the TF and FP relations should be revisited with additional structural properties, in order to remove the residual correlation with the local cosmological fluctuations. The challenge is to measure structural or kinematic indicators (concentration, velocity profile, etc) in a distance independent way. Ultimately, if the relations cannot be calibrated to account for these residuals correlations, they could be analytically modelled as an agnostic bias expansion of the density field.

For future work, one should study different definition of environments, as different galaxy formation mechanisms dominate different physical scales, and test the impact of various sample selection criteria. The redshift evolution of the TF and FP relations might also alternate the correlate signal that we observed in this work. It would also be necessary to study the impact of observational effects of those correlated residuals.

Conclusion and prospects

Throughout this thesis, I have studied how to extract cosmological information from the LSS by analysing the clustering of galaxies and their peculiar velocities, measured in spectroscopic and photometric surveys. To address the challenges of modern cosmology, I aimed to extract as much cosmological information as possible from the large amount of available data. I compared different methods of combining standard probes and studied the optimal ways to estimate observables. Additionally, I developed machine learning tools, making use of state-of-the-art theoretical descriptions of the LSS provided by cosmological simulations to model and study the physical processes occurring on the smallest scales.

In chapter 3, I focused on BAO measurements and how to define consensus results from clustering analyses in Fourier space using the power spectrum, and in configuration space using the correlation function. I developed an analysis to jointly fit the power spectrum and correlation function multipoles while accounting for their strong correlations, as measured from the same galaxy catalogue. I compared our results with an alternative method (GA, (SÁNCHEZ et al. 2017)) used for the eBOSS analysis (GIL-MARÍN et al. 2021 ; J. E. BAUTISTA et al. 2021), which directly combines the parameters' correlated posteriors. I tested and validated our framework on the BAO analysis of the eBOSS Luminous Red Galaxy (LRG) sample, fitting for the BAO parameters $(\alpha_\parallel, \alpha_\perp)$. I found that our joint analysis performed better for non-Gaussian posteriors. Overall, I obtained consistent results between the two methods. I found that the GA method requires fewer mocks to properly build the covariance matrix and gives slightly tighter constraints but with larger systematic errors on the best-fitted values. This study is important as it highlights the need to balance between data compression and loss of information in any multiprobe analysis.

In an effort to accurately model non-linear clustering beyond perturbation theories and improve cosmological constraints, I developed a new machine learning framework in chapter 4. This improved Gaussian Process algorithm extends the model's parameter space to higher dimensions. I tested the formalism using the ABACUSSUMMIT N-body simulation suite and the HOD framework, emulating the monopole of the redshift-space galaxy correlation function over a continuous range of scales, in addition to the cosmological and HOD parameters. This extended model allows to consider the well-known correlations between scales and include of the AP distortion effect in the analysis. By performing cosmological inference with the trained model, I successfully retrieved unbiased parameters. It is crucial to account for every source of error in the model prediction when using an emulator model, as neglecting these errors can lead to biased results. I demonstrated that using the predicted model uncertainty (covariance) instead of estimating it from a test set yields more accurate and precise constraints. Therefore, the multiscale model performed better compared to the standard model that emulates each scale independently. With this new formalism, the emulator model could be extended to include more galaxy bias parameters or different redshift bins. Additionally, one could parameterise and forward model systematic effects such as fibre collision, which is the main observational limitation for measuring unbiased clustering on small scales. Reproducing selection effects also raises the issue of cosmic variance in the model, especially at low redshift where the volume is limited. Future studies could

employ different approaches (ANGULO et PONTZEN 2016b; HADZHIYSKA et al. 2023a) to suppress cosmic variance in the emulated statistics. This SBI pipeline can, of course, be adapted to any alternative statistics of the galaxy distribution.

In chapter 5, I applied the SBI framework to emulate the non-linear scales of my new estimators describing the joint clustering of galaxy and peculiar velocity. I demonstrated that running inference on the galaxy-galaxy, velocity-velocity, or velocity-galaxy correlation functions consistently retrieves unbiased parameters. Moreover, I showed that after reproducing the tracers' expected density (SAULDER et al. 2023a) and velocity errors (BORUAH et al. 2020), combining the galaxy and peculiar velocity information on non-linear scales yields tighter constraints on cosmological parameters, especially on the growth rate. Furthermore, including up to the smallest scale $0.33h^{-1}$Mpc in my analysis could reduce the error bars. However, emulator predictions depend on the simulation suite specificities. Consequently, its robustness must be tested on higher-fidelity simulations, with physically motivated description before fitting real data. It would be relevant to extend the galaxy-matter connection beyond the vanilla HOD formalism by including assembly or peculiar velocity biases, using more physically motivated approaches such as SHAM (BEHROOZI et al. 2010) and SAM (S. D. M. WHITE et FRENK 1991), or employing model-independent approaches like hybrid bias expansion (PELLEJERO IBAÑEZ et al. 2023). The challenge is to have a galaxy-matter connection flexible enough to be marginalised over unmodelled physics without biasing cosmological parameters (as reality is necessarily more complex than our best models), while still extracting as much information as possible. One could consider extending the 3×2 points analysis to include more multipoles, and conduct a hybrid analysis using emulators and linear theory to respectively model the non-linear and linear scales. The largest scales of the low redshift clustering would require proper accounting for the wide-angle effect, which can be treated analytically (YOO et SELJAK 2015). For future work, I believe it would also be important to assess and model the impact of the fiducial cosmology when estimating peculiar velocity.

In chapter 6, I tested the standard assumptions about the theoretical systematic errors in the peculiar velocities measured from standardisable objects. Using the state-of-the-art hydrodynamical simulation MilleniumTNG (covering a $(500\ h^{-1}\text{Mpc})^3$ volume) I reproduced the FP and TF relations, commonly used as distance indicators for ETG and LTG respectively, standardising their effective radius and absolute magnitude. The intrinsic scatter of these relations propagates to the measured distances, hence peculiar velocities. It has been assumed to be caused by astrophysical processes, uncorrelated with the LSS, and typically modelled with a Gaussian random distribution, mostly affecting the clustering on small scales without biasing the derived cosmological parameters. However, I found that the residuals between the true and estimated velocities were correlated with the local environment on scales typically below $10\ h^{-1}$Mpc. Such correlations systematically bias the velocity clustering, which could lead to biased cosmological parameters when including the small scales in the analysis. I interpreted this result as indicating that the quantities used to calibrate the TF and FP relations are insufficient to fully capture the galaxies' structural properties and their correlation with the matter field fluctuations. I observed that including stellar and gas masses along with the SFR information allowed for a reduction in the scatter amplitude and the amount of correlated signal. Tracing back the star formation history confirmed that additional structural indicators were necessary to fully capture the galaxies' size and brightness dependence on the local cosmological fluctuations. Using the non-linear scales' information is a challenging task, as they are affected by several observational and theoretical systematic effects that need to be properly treated. For TF and FP measurements, the smallest

scales of the velocity clustering should either be removed from the analysis, modelled with additional agnostic bias expansion, or their uncertainties increased. One should, of course, push further our understanding of TF and FP relations to include additional (distance-independent) environmental properties and suppress the correlated signal in the residuals. This work could also be extended to study the intrinsic scatter of the standardised SnIa, though supernovae events rely on sub-grid physics which is difficult to properly model in hydrodynamical simulations. Of course, it is important to assess if this kind of signal will be observed in the upcoming data (DESI SAULDER et al. 2023a, ZTF BELLM et al. 2019, LSST SCIENCE COLLABORATION et al. 2009) as it was already the case in previous measurements (JOACHIMI et al. 2015).

In the next few years, DESI and Euclid will observe the redshift of tens of millions of galaxies, making up the largest map of the Universe ever made. The ZTF and upcoming Rubin-LSST science observations will produce the largest sample of SnIa, providing a substantial amount of distance and peculiar velocity measurements. With such a vast amounts of data, it is crucial to continually develop and refine our statistical tools, theoretical models, and treatment of systematic effects. Ongoing advancements in these areas are essential to ensure we can fully leverage the available information. The cross-analysis of galaxy clustering, peculiar velocity, and additional overlapping probes (such as weak lensing) will yield precise growth rate measurements and allow for rigorous testing of cosmological models, ultimately pushing deeper our understanding of Dark Energy, gravity, and the laws of our Universe.

As I write the final lines of my thesis, it is quite difficult for me to know whether these three years of research will have any impact on our understanding of the Universe (I kind of doubt it). But if I have convinced you of my ability to contribute to our scientific community, then it was worth it.

Bibliographie

[1] Caitlin ADAMS et Chris BLAKE. « Improving Constraints on the Growth Rate of Structure by Modelling the Density–Velocity Cross-Correlation in the 6dF Galaxy Survey ». In : *Monthly Notices of the Royal Astronomical Society* 471.1 (oct. 2017), p. 839-856. ISSN : 0035-8711. DOI : 10.1093/mnras/stx1529 (cf. p. 13, 54, 149).

[2] Nishant AGARWAL, Vincent DESJACQUES, Donghui JEONG et Fabian SCHMIDT. « Information content in the redshift-space galaxy power spectrum and bispectrum ». In : *jcap* 2021.3, 021 (mars 2021), p. 021. DOI : 10.1088/1475-7516/2021/03/021. arXiv : 2007.04340 [astro-ph.CO] (cf. p. 149).

[3] Shadab ALAM et al. « The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey : cosmological analysis of the DR12 galaxy sample ». In : *MNRAS* 470.3 (sept. 2017), p. 2617-2652. DOI : 10.1093/mnras/stx721. arXiv : 1607.03155 [astro-ph.CO] (cf. p. 10, 55, 87).

[4] C. ALCOCK et B. PACZYNSKI. « An evolution free test for non-zero cosmological constant ». In : *nat* 281 (oct. 1979), p. 358. DOI : 10.1038/281358a0 (cf. p. 11, 47, 137, 153).

[5] Raul E. ANGULO et Oliver HAHN. « Large-scale dark matter simulations ». In : *Living Reviews in Computational Astrophysics* 8.1, 1 (déc. 2022), p. 1. DOI : 10.1007/s41115-021-00013-z. arXiv : 2112.05165 [astro-ph.CO] (cf. p. 62).

[6] Raul E. ANGULO et Andrew PONTZEN. « Cosmological N-Body Simulations with Suppressed Variance ». In : *Monthly Notices of the Royal Astronomical Society : Letters* 462.1 (oct. 2016), p. L1-L5. ISSN : 1745-3925, 1745-3933. DOI : 10.1093/mnrasl/slw098 (cf. p. 182).

[7] Raul E. ANGULO et Andrew PONTZEN. « Cosmological N-body simulations with suppressed variance ». In : *mnras* 462.1 (oct. 2016), p. L1-L5. DOI : 10.1093/mnrasl/slw098. arXiv : 1603.05253 [astro-ph.CO] (cf. p. 18, 63, 111, 198).

[8] Monica BARRERA et al. « The MillenniumTNG Project : semi-analytic galaxy formation models on the past lightcone ». In : *mnras* 525.4 (nov. 2023), p. 6312-6335. DOI : 10.1093/mnras/stad2688. arXiv : 2210.10419 [astro-ph.CO] (cf. p. 182).

[9] Julian E. BAUTISTA et al. « The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey : Baryon Acoustic Oscillations at Redshift of 0.72 with the DR14 Luminous Red Galaxy Sample ». In : *The Astrophysical Journal* 863.1, 110 (août 2018), p. 110. DOI : 10.3847/1538-4357/aacea5. arXiv : 1712.08064 [astro-ph.CO] (cf. p. 10, 55).

[10] Julian E. BAUTISTA et al. « The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : measurement of the BAO and growth rate of structure of the luminous red galaxy sample from the anisotropic correlation function between redshifts 0.6 and 1 ». In : *MNRAS* 500.1 (jan. 2021), p. 736-762. DOI : 10.1093/mnras/staa2800. arXiv : 2007.08993 [astro-ph.CO] (cf. p. 10, 55, 87, 88, 90, 92, 97, 100, 104, 197).

[11] Peter S. BEHROOZI, Charlie CONROY et Risa H. WECHSLER. « A Comprehensive Analysis of Uncertainties Affecting the Stellar Mass-Halo Mass Relation for 0 < z < 4 ». In : *apj* 717.1 (juill. 2010), p. 379-403. DOI : 10.1088/0004-637X/717/1/379. arXiv : 1001.0015 [astro-ph.CO] (cf. p. 18, 64, 198).

[12] Eric C. BELLM et al. « The Zwicky Transient Facility : System Overview, Performance, and First Results ». In : *pasp* 131.995 (jan. 2019), p. 018002. DOI : 10.1088/1538-3873/aaecbe. arXiv : 1902.01932 [astro-ph.IM] (cf. p. 14, 18, 68, 152, 199).

[13] Andreas A. BERLIND et David H. WEINBERG. « The Halo Occupation Distribution : Toward an Empirical Determination of the Relation between Galaxies and Mass ». In : *apj* 575.2 (août 2002), p. 587-616. DOI : 10.1086/341469. arXiv : astro-ph/0109001 [astro-ph] (cf. p. 63).

[14] F. BERNARDEAU, S. COLOMBI, E. GAZTANAGA et R. SCOCCIMARRO. « Large-Scale Structure of the Universe and Cosmological Perturbation Theory ». en. In : *Physics Reports* 367.1-3 (sept. 2002). arXiv :astro-ph/0112551, p. 1-248. ISSN : 03701573. DOI : 10.1016/S0370-1573(02)00135-7 (cf. p. 25, 61).

[15] Florian BEUTLER et al. « The 6dF Galaxy Survey : Baryon Acoustic Oscillations and the Local Hubble Constant ». en. In : *Monthly Notices of the Royal Astronomical Society* 416.4 (oct. 2011). arXiv :1106.3366 [astro-ph], p. 3017-3032. ISSN : 00358711. DOI : 10.1111/j.1365-2966.2011.19250.x (cf. p. 87).

[16] Florian BEUTLER et al. « The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey : baryon acoustic oscillations in the Fourier space ». In : *mnras* 464.3 (jan. 2017), p. 3409-3430. DOI : 10.1093/mnras/stw2373. arXiv : 1607.03149 [astro-ph.CO] (cf. p. 58).

[17] Davide BIANCHI, Héctor GIL-MARÍN, Rossana RUGGERI et Will J. PERCIVAL. « Measuring line-of-sight-dependent Fourier-space clustering using FFTs ». In : *mnras* 453.1 (oct. 2015), p. L11-L15. DOI : 10.1093/mnrasl/slv090. arXiv : 1505.05341 [astro-ph.CO] (cf. p. 76).

[18] Davide BIANCHI, Héctor GIL-MARÍN, Rossana RUGGERI et Will J. PERCIVAL. « Measuring line-of-sight-dependent Fourier-space clustering using FFTs ». In : *MNRAS* 453.1 (oct. 2015), p. L11-L15. DOI : 10.1093/mnrasl/slv090. arXiv : 1505.05341 [astro-ph.CO] (cf. p. 89).

[19] Michael R. BLANTON et al. « Sloan Digital Sky Survey IV : Mapping the Milky Way, Nearby Galaxies, and the Distant Universe ». In : *The Astrophysical Journal* 154.1, 28 (juill. 2017), p. 28. DOI : 10.3847/1538-3881/aa7567. arXiv : 1703.00052 [astro-ph.GA] (cf. p. 45, 87).

[20] Supranta S. BORUAH, Michael J. HUDSON et Guilhem LAVAUX. « Cosmic flows in the nearby Universe : new peculiar velocities from SNe and cosmological constraints ». en. In : *Monthly Notices of the Royal Astronomical Society* 498.2 (sept. 2020). arXiv :1912.09383 [astro-ph], p. 2703-2718. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/staa2485 (cf. p. 14, 68, 153, 198).

[21] Sownak BOSE et al. « The MillenniumTNG Project : the large-scale clustering of galaxies ». In : *mnras* 524.2 (sept. 2023), p. 2579-2593. DOI : 10.1093/mnras/stad1097. arXiv : 2210.10065 [astro-ph.CO] (cf. p. 182).

[22] Samuel BRIEDEN, Héctor GIL-MARÍN et Licia VERDE. « ShapeFit : extracting the power spectrum shape information in galaxy surveys beyond BAO and RSD ». In : *jcap* 2021.12, 054 (déc. 2021), p. 054. DOI : 10.1088/1475-7516/2021/12/054. arXiv : 2106.07641 [astro-ph.CO] (cf. p. 59).

[23] A. BURDEN, W. J. PERCIVAL et C. HOWLETT. « Reconstruction in Fourier space ». en. In : *Monthly Notices of the Royal Astronomical Society* 453.1 (oct. 2015), p. 456-468. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stv1581 (cf. p. 10).

[24] Lachlan A. CAMPBELL et al. « The 6dF Galaxy Survey : Fundamental Plane data ». In : *Monthly Notices of the Royal Astronomical Society* 443 (sept. 2014), p. 1231-1251. ISSN : 0035-8711. DOI : 10.1093/mnras/stu1198 (cf. p. 152, 180).

[25] Jordan CARLSON, Beth REID et Martin WHITE. « Convolution Lagrangian perturbation theory for biased tracers ». en. In : *Monthly Notices of the Royal Astronomical Society* 429.2 (fév. 2013), p. 1674-1685. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/sts457 (cf. p. 61).

[26] John Joseph M. CARRASCO, Simon FOREMAN, Daniel GREEN et Leonardo SENATORE. « The Effective Field Theory of Large Scale Structures at Two Loops ». en. In : *Journal of Cosmology and Astroparticle Physics* 2014.07 (juill. 2014). arXiv :1310.0464 [astro-ph, physics :gr-qc, physics :hep-ph, physics :hep-th], p. 057-057. ISSN : 1475-7516. DOI : 10.1088/1475-7516/2014/07/057 (cf. p. 61).

[27] J. CARRON. « On the assumption of Gaussianity for cosmological two-point statistics and parameter dependent covariance matrices ». In : *aap* 551, A88 (mars 2013), A88. DOI : 10.1051/0004-6361/201220538. arXiv : 1204.4724 [astro-ph.CO] (cf. p. 79).

[28] Paul CARTER, Florian BEUTLER, Will J. PERCIVAL, Joseph DEROSE, Risa H. WECHSLER et Cheng ZHAO. « The Impact of the Fiducial Cosmology Assumption on BAO Cosmological Parameter Inference ». en. In : *Monthly Notices of the Royal Astronomical Society* 494.2 (mai 2020). arXiv : 1906.03035, p. 2076-2089. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/staa761 (cf. p. 55).

[29] Michael J. CHAPMAN et al. « The Completed SDSS-IV Extended Baryon Oscillation Spectroscopic Survey : Measurement of the Growth Rate of Structure from the Small-Scale Clustering of the Luminous Red Galaxy Sample ». In : *Monthly Notices of the Royal Astronomical Society* 516 (oct. 2022), p. 617-635. ISSN : 0035-8711. DOI : 10.1093/mnras/stac1923 (cf. p. 111).

[30] Edmond CHAUSSIDON et al. « Target Selection and Validation of DESI Quasars ». In : *apj* 944.1, 107 (fév. 2023), p. 107. DOI : 10.3847/1538-4357/acb3c2. arXiv : 2208.08511 [astro-ph.CO] (cf. p. 66).

[31] Jonás CHAVES-MONTERO et al. « Subhalo abundance matching and assembly bias in the EAGLE simulation ». In : *mnras* 460.3 (août 2016), p. 3100-3118. DOI : 10.1093/mnras/stw1225. arXiv : 1507.01948 [astro-ph.GA] (cf. p. 195).

[32] M. CHEVALLIER et D. POLARSKI. « Accelerating Universes with Scaling Dark Matter ». en. In : *International Journal of Modern Physics D* 10.02 (avr. 2001). arXiv :gr-qc/0009008, p. 213-223. ISSN : 0218-2718, 1793-6594. DOI : 10.1142/S0218271801000822 (cf. p. 31).

[33] N. CHISARI et al. « Intrinsic alignments of galaxies in the Horizon-AGN cosmological hydrodynamical simulation ». In : *mnras* 454.3 (déc. 2015), p. 2736-2753. DOI : 10.1093/mnras/stv2154. arXiv : 1507.07843 [astro-ph.CO] (cf. p. 69).

[34] Chia-Hsun CHUANG, Francisco-Shu KITAURA, Francisco PRADA, Cheng ZHAO et Gustavo YEPES. « EZmocks : extending the Zel'dovich approximation to generate mock galaxy catalogues with accurate clustering statistics ». In : *mnras* 446.3 (jan. 2015), p. 2621-2628. DOI : 10.1093/mnras/stu2301. arXiv : 1409.1124 [astro-ph.CO] (cf. p. 63).

[35] Chia-Hsun CHUANG, Francisco-Shu KITAURA, Francisco PRADA, Cheng ZHAO et Gustavo YEPES. « EZmocks : extending the Zel'dovich approximation to generate mock galaxy catalogues with accurate clustering statistics ». In : *MNRAS* 446.3 (jan. 2015), p. 2621-2628. DOI : 10.1093/mnras/stu2301. arXiv : 1409.1124 [astro-ph.CO] (cf. p. 88).

[36] Douglas CLOWE et al. « A Direct Empirical Proof of the Existence of Dark Matter ». In : *apjl* 648.2 (sept. 2006), p. L109-L113. DOI : 10.1086/508162. arXiv : astro-ph/0608407 [astro-ph] (cf. p. 30).

[37] DESI COLLABORATION et al. *The DESI Experiment Part I : Science,Targeting, and Survey Design*. en. arXiv :1611.00036 [astro-ph]. Déc. 2016 (cf. p. 45, 111).

[38] DESI COLLABORATION et al. *DESI 2024 VI : Cosmological Constraints from the Measurements of Baryon Acoustic Oscillations*. en. arXiv :2404.03002 [astro-ph]. Avr. 2024 (cf. p. 50, 87).

[39] Euclid COLLABORATION et al. « Euclid Preparation. I. The Euclid Wide Survey ». In : *Astronomy and Astrophysics* 662 (juin 2022), A112. ISSN : 0004-6361. DOI : 10.1051/0004-6361/202141938 (cf. p. 53, 111, 149).

[40] Planck COLLABORATION et al. « Planck 2018 results. VI. Cosmological parameters ». In : *Astronomy & Astrophysics* 641 (sept. 2020). arXiv :1807.06209 [astro-ph], A6. ISSN : 0004-6361, 1432-0746. DOI : 10.1051/0004-6361/201833910 (cf. p. 113).

[41] Matthew COLLESS et al. « The 2dF Galaxy Redshift Survey : Final Data Release ». In : *arXiv e-prints* (juin 2003), astro-ph/0306581 (cf. p. 45).

[42] Ralph D'AGOSTINO et E S PEARSON. « Tests for departure from normality. Empirical results for the distributions of b% and V*i ». en. In : (1973), p. 10 (cf. p. 100).

[43] Ralph B D'AGOSTINO. « An Omnibus Test of Normality for Moderate and Large Size Samples ». en. In : (1971), p. 9 (cf. p. 100).

[44] DARK ENERGY SURVEY COLLABORATION et al. « The Dark Energy Survey : more than dark energy - an overview ». In : *MNRAS* 460.2 (août 2016), p. 1270-1299. DOI : 10.1093/mnras/stw641. arXiv : 1601.00329 [astro-ph.CO] (cf. p. 87).

[45] M. DAVIS, G. EFSTATHIOU, C. S. FRENK et S. D. M. WHITE. « The evolution of large-scale structure in a universe dominated by cold dark matter ». In : *apj* 292 (mai 1985), p. 371-394. DOI : 10.1086/163168 (cf. p. 62).

[46] M. DAVIS, G. EFSTATHIOU, C. S. FRENK et S. D. M. WHITE. « The evolution of large-scale structure in a universe dominated by cold dark matter ». In : *apj* 292 (mai 1985), p. 371-394. DOI : 10.1086/163168 (cf. p. 63).

[47] Kyle S. DAWSON et al. « The Baryon Oscillation Spectroscopic Survey of SDSS-III ». In : *The Astrophysical Journal* 145.1, 10 (jan. 2013), p. 10. DOI : 10.1088/0004-6256/145/1/10. arXiv : 1208.0022 [astro-ph.CO] (cf. p. 45, 87).

[48] Kyle S. DAWSON et al. « The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey : Overview and Early Data ». In : *The Astrophysical Journal* 151.2, 44 (fév. 2016), p. 44. DOI : 10.3847/0004-6256/151/2/44. arXiv : 1508.04473 [astro-ph.CO] (cf. p. 45, 87, 88).

[49] Anna DE GRAAFF et al. « A common origin for the fundamental plane of quiescent and star-forming galaxies in the EAGLE simulations ». In : *mnras* 518.4 (fév. 2023), p. 5376-5402. DOI : 10.1093/mnras/stac3277. arXiv : 2207.13491 [astro-ph.GA] (cf. p. 180).

[50] Arnaud DE MATTIA et Vanina RUHLMANN-KLEIDER. « Integral constraints in spectroscopic surveys ». In : *jcap* 2019.8, 036 (août 2019), p. 036. DOI : 10.1088/1475-7516/2019/08/036. arXiv : 1904.08851 [astro-ph.CO] (cf. p. 77).

[51] Arnaud DE MATTIA et al. « The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : measurement of the BAO and growth rate of structure of the emission line galaxy sample from the anisotropic power spectrum between redshift 0.6 and 1.1 ». In : *MNRAS* 501.4 (mars 2021), p. 5616-5645. DOI : 10.1093/mnras/staa3891. arXiv : 2007.09008 [astro-ph.CO] (cf. p. 87).

[52] Joseph DEROSE et al. « The Aemulus Project. I. Numerical Simulations for Precision Cosmology ». en. In : *The Astrophysical Journal* 875.1 (avr. 2019), p. 69. ISSN : 1538-4357. DOI : 10.3847/1538-4357/ab1085 (cf. p. 111).

[53] DES COLLABORATION et al. « Dark Energy Survey Year 3 Results : A 2.7% measurement of Baryon Acoustic Oscillation distance scale at redshift 0.835 ». In : *arXiv e-prints*, arXiv :2107.04646 (juill. 2021), arXiv :2107.04646. arXiv : 2107.04646 [astro-ph.CO] (cf. p. 87).

[54] DESI COLLABORATION et al. « Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument ». In : *The Astronomical Journal* 164 (nov. 2022), p. 207. ISSN : 0004-6256. DOI : 10.3847/1538-3881/ac882b (cf. p. 53, 111, 149).

[55] DESI COLLABORATION et al. « DESI 2024 III : Baryon Acoustic Oscillations from Galaxies and Quasars ». In : *arXiv e-prints*, arXiv :2404.03000 (avr. 2024), arXiv :2404.03000. DOI : 10.48550/arXiv.2404.03000. arXiv : 2404.03000 [astro-ph.CO] (cf. p. 87).

[56] DESI COLLABORATION et al. « DESI 2024 IV : Baryon Acoustic Oscillations from the Lyman Alpha Forest ». In : *arXiv e-prints*, arXiv :2404.03001 (avr. 2024), arXiv :2404.03001. DOI : 10.48550/arXiv.2404.03001. arXiv : 2404.03001 [astro-ph.CO] (cf. p. 87).

[57] Vincent DESJACQUES, Donghui JEONG et Fabian SCHMIDT. « Large-Scale Galaxy Bias ». en. In : *Physics Reports* 733 (fév. 2018). arXiv : 1611.09787, p. 1-193. ISSN : 03701573. DOI : 10.1016/j.physrep.2017.12.002 (cf. p. 46).

[58] S. DJORGOVSKI et Marc DAVIS. « Fundamental Properties of Elliptical Galaxies ». In : *apj* 313 (fév. 1987), p. 59. DOI : 10.1086/164948 (cf. p. 69).

[59] Scott DODELSON et Michael D. SCHNEIDER. « The effect of covariance estimator error on cosmological parameter constraints ». In : *PRD* 88.6, 063537 (sept. 2013), p. 063537. DOI : 10.1103/PhysRevD.88.063537. arXiv : 1304.2593 [astro-ph.CO] (cf. p. 25, 42, 83).

[60] Michael J. DRINKWATER et al. « The WiggleZ Dark Energy Survey : Survey Design and First Data Release ». In : *Monthly Notices of the Royal Astronomical Society* 401.3 (jan. 2010), p. 1429. ISSN : 0035-8711. DOI : 10.1111/j.1365-2966.2009.15754.x (cf. p. 45).

[61] Hélion DU MAS DES BOURBOUX et al. « The Completed SDSS-IV Extended Baryon Oscillation Spectroscopic Survey : Baryon Acoustic Oscillations with Lyα Forests ». In : *The Astrophysical Journal* 901.2, 153 (oct. 2020), p. 153. DOI : 10.3847/1538-4357/abb085. arXiv : 2007.08995 [astro-ph.CO] (cf. p. 87).

[62] Tyann DUMERCHAT et Julian BAUTISTA. « Galaxy clustering multi-scale emulation ». In : *aap* 684, A57 (avr. 2024), A57. DOI : 10.1051/0004-6361/202348640. arXiv : 2311.10425 [astro-ph.CO] (cf. p. 111).

[63] Tyann DUMERCHAT et Julian E. BAUTISTA. « Baryon Acoustic Oscillations from a Joint Analysis of the Large-Scale Clustering in Fourier and Configuration Space ». In : *Astronomy and Astrophysics* 667 (nov. 2022), A80. ISSN : 0004-6361. DOI : 10.1051/0004-6361/202244100 (cf. p. 88).

[64] Gia DVALI, Gregory GABADADZE et Massimo PORRATI. « 4D Gravity on a Brane in 5D Minkowski Space ». en. In : *Physics Letters B* 485.1-3 (juill. 2000). arXiv :hep-th/0005016, p. 208-214. ISSN : 03702693. DOI : 10.1016/S0370-2693(00)00669-9 (cf. p. 53).

[65] Bradley EFRON. *The Jackknife, the Bootstrap and other resampling plans*. 1982 (cf. p. 84).

[66] Daniel J. EISENSTEIN, Hee-Jong SEO, Edwin SIRKO et David N. SPERGEL. « Improving Cosmological Distance Measurements by Reconstruction of the Baryon Acoustic Peak ». In : *apj* 664.2 (août 2007), p. 675-679. DOI : 10.1086/518712. arXiv : astro-ph/0604362 [astro-ph] (cf. p. 55).

[67] Daniel J. EISENSTEIN et al. « SDSS-III : Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems ». In : *The Astrophysical Journal* 142.3, 72 (sept. 2011), p. 72. DOI : 10.1088/0004-6256/142/3/72. arXiv : 1101.1529 [astro-ph.IM] (cf. p. 45, 87).

[68] U. FEINDT et al. « Measuring cosmic bulk flows with Type Ia supernovae from the Nearby Supernova Factory ». In : *aap* 560, A90 (déc. 2013), A90. DOI : 10.1051/0004-6361/201321880. arXiv : 1310.4184 [astro-ph.CO] (cf. p. 77).

[69] Hume A. FELDMAN, Nick KAISER et John A. PEACOCK. « Power-Spectrum Analysis of Three-dimensional Redshift Surveys ». In : *The Astrophysical Journal* 426 (mai 1994), p. 23. ISSN : ['0004-637X']. DOI : 10.1086/174036 (cf. p. 71, 76).

[70] R. GANNOUJI, B. MORAES et D. POLARSKI. « The growth of matter perturbations in f(R) models ». en. In : *Journal of Cosmology and Astroparticle Physics* 2009.02 (fév. 2009). arXiv :0809.3374 [astro-ph, physics :gr-qc], p. 034-034. ISSN : 1475-7516. DOI : 10.1088/1475-7516/2009/02/034 (cf. p. 53).

[71] Lehman H. GARRISON, Daniel J. EISENSTEIN, Douglas FERRER, Nina A. MAKSIMOVA et Philip A. PINTO. « The ABACUS cosmological N-body code ». In : *Monthly Notices of the Royal Astronomical Society* 508 (nov. 2021). ADS Bibcode : 2021MNRAS.508..575G, p. 575-596. ISSN : 0035-8711. DOI : 10.1093/mnras/stab2482 (cf. p. 112).

[72] Héctor GIL-MARÍN. « How to optimally combine pre-reconstruction full shape and post-reconstruction BAO signals ». In : *arXiv e-prints*, arXiv :2203.05581 (mars 2022), arXiv :2203.05581. arXiv : 2203.05581 [astro-ph.CO] (cf. p. 105).

[73] Héctor GIL-MARÍN et al. « The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey : BAO measurement from the LOS-dependent power spectrum of DR12 BOSS galaxies ». In : *MNRAS* 460.4 (août 2016), p. 4210-4219. DOI : 10.1093/mnras/stw1264. arXiv : 1509.06373 [astro-ph.CO] (cf. p. 10, 55).

[74] Héctor GIL-MARÍN et al. « The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample : structure growth rate measurement from the anisotropic quasar power spectrum in the redshift range 0.8 < z < 2.2 ». In : *MNRAS* 477.2 (juin 2018), p. 1604-1638. DOI : 10.1093/mnras/sty453. arXiv : 1801.02689 [astro-ph.CO] (cf. p. 10, 55).

[75] Héctor GIL-MARÍN et al. « The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : measurement of the BAO and growth rate of structure of the luminous red galaxy sample from the anisotropic power spectrum between redshifts 0.6 and 1.0 ». In : *MNRAS* 498.2 (oct. 2021), p. 2492-2531. DOI : 10.1093/mnras/staa2455. arXiv : 2007.08994 [astro-ph.CO] (cf. p. 10, 55, 87, 88, 90, 92, 97, 100, 104, 197).

[76] M. GLOWACKI, E. ELSON et R. DAVÉ. « The baryonic Tully-Fisher relation in the SIMBA simulation ». In : *mnras* 498.3 (nov. 2020), p. 3687-3702. DOI : 10.1093/mnras/staa2616. arXiv : 2003.03402 [astro-ph.GA] (cf. p. 180).

[77] Julian S. GODDY, David V. STARK, Karen L. MASTERS, Kevin BUNDY, Niv DRORY et David R. LAW. « A comparison of the baryonic Tully-Fisher relation in MaNGA and IllustrisTNG ». In : *mnras* 520.3 (avr. 2023), p. 3895-3908. DOI : 10.1093/mnras/stad298. arXiv : 2302.05029 [astro-ph.GA] (cf. p. 180).

[78] Krzysztof GORSKI. « On the Pattern of Perturbations of the Hubble Flow ». In : *The Astrophysical Journal Letters* 332 (sept. 1988), p. L7-L11. ISSN : 0004-637X. DOI : 10.1086/185255 (cf. p. 73).

[79] Jan Niklas GRIEB, Ariel G. SÁNCHEZ, Salvador SALAZAR-ALBORNOZ et Claudio DALLA-VECCHIA. « Gaussian covariance matrices for anisotropic galaxy clustering measurements ». en. In : *Monthly Notices of the Royal Astronomical Society* 457.2 (avr. 2016), p. 1577-1592. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stw065 (cf. p. 83, 109).

[80] James E. GUNN et al. « The 2.5 m Telescope of the Sloan Digital Sky Survey ». In : *The Astrophysical Journal* 131.4 (avr. 2006), p. 2332-2359. DOI : 10.1086/500975. arXiv : astro-ph/0602326 [astro-ph] (cf. p. 88).

[81] Hong GUO, Idit ZEHAVI et Zheng ZHENG. « A New Method to Correct for Fiber Collisions in Galaxy Two-point Statistics ». In : *apj* 756.2, 127 (sept. 2012), p. 127. DOI : 10.1088/0004-637X/756/2/127. arXiv : 1111.6598 [astro-ph.CO] (cf. p. 72).

[82] Hong GUO et al. « Velocity bias from the small-scale clustering of SDSS-III BOSS galaxies ». In : *mnras* 446.1 (jan. 2015), p. 578-594. DOI : 10.1093/mnras/stu2120. arXiv : 1407.4811 [astro-ph.CO] (cf. p. 169).

[83] L. GUZZO et al. « The VIMOS Public Extragalactic Redshift Survey (VIPERS). An Unprecedented View of Galaxies and Large-Scale Structure at 0.5 < z < 1.2 ». In : *Astronomy and Astrophysics* 566 (juin 2014), A108. ISSN : ['0004-6361']. DOI : 10.1051/0004-6361/201321489 (cf. p. 45).

[84] Boryana HADZHIYSKA, Daniel EISENSTEIN, Sownak BOSE, Lehman H. GARRISON et Nina MAKSIMOVA. « \textsc{CompaSO} : A new halo finder for competitive assignment to spherical overdensities ». en. In : *Monthly Notices of the Royal Astronomical Society* 509.1 (nov. 2021). arXiv :2110.11408 [astro-ph], p. 501-521. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stab2980 (cf. p. 113).

[85] Boryana HADZHIYSKA et al. « Mitigating the noise of DESI mocks using analytic control variates ». In : *The Open Journal of Astrophysics* 6, 38 (oct. 2023), p. 38. DOI : 10.21105/astro.2308.12343. arXiv : 2308.12343 [astro-ph.CO] (cf. p. 18, 63, 111, 198).

[86] Boryana HADZHIYSKA et al. « The MillenniumTNG Project : refining the one-halo model of red and blue galaxies at different redshifts ». In : *mnras* 524.2 (sept. 2023), p. 2524-2538. DOI : 10.1093/mnras/stad279. arXiv : 2210.10068 [astro-ph.CO] (cf. p. 182).

[87] ChangHoon HAHN, Florian BEUTLER, Manodeep SINHA, Andreas BERLIND, Shirley HO et David W. HOGG. « Likelihood non-Gaussianity in large-scale structure analyses ». In : *mnras* 485.2 (mai 2019), p. 2956-2969. DOI : 10.1093/mnras/stz558. arXiv : 1803.06348 [astro-ph.CO] (cf. p. 79).

[88] ChangHoon HAHN, Roman SCOCCIMARRO, Michael R. BLANTON, Jeremy L. TINKER et Sergio A. RODRÍGUEZ-TORRES. « The Effect of Fiber Collisions on the Galaxy Power Spectrum Multipoles ». In : *mnras* 467.2 (mai 2017), p. 1940-1956. DOI : 10.1093/mnras/stx185. arXiv : 1609.01714 [astro-ph.CO] (cf. p. 72).

[89] A. J. S. HAMILTON. « Uncorrelated modes of the non-linear power spectrum ». In : *MNRAS* 312.2 (fév. 2000), p. 257-284. DOI : 10.1046/j.1365-8711.2000.03071.x. arXiv : astro-ph/9905191 [astro-ph] (cf. p. 57).

[90] J. HARTLAP, P. SIMON et P. SCHNEIDER. « Why your model parameter confidences might be too optimistic. Unbiased estimation of the inverse covariance matrix ». In : *AAP* 464.1 (mars 2007), p. 399-404. DOI : 10.1051/0004-6361:20066170. arXiv : astro-ph/0608064 [astro-ph] (cf. p. 83).

[91] Michael G. HAUSER et Eli DWEK. « The Cosmic Infrared Background : Measurements and Implications ». In : *araa* 39 (jan. 2001), p. 249-307. DOI : 10.1146/annurev.astro.39.1.249. arXiv : astro-ph/0105539 [astro-ph] (cf. p. 30).

[92] César HERNÁNDEZ-AGUAYO et al. « The MillenniumTNG Project : high-precision predictions for matter clustering and halo statistics ». In : *mnras* 524.2 (sept. 2023), p. 2556-2578. DOI : 10.1093/mnras/stad1657. arXiv : 2210.10059 [astro-ph.CO] (cf. p. 182).

[93] Ryley HILL, Kiyoshi W. MASUI et Douglas SCOTT. « The Spectrum of the Universe ». In : *Applied Spectroscopy* 72.5 (mai 2018), p. 663-688. DOI : 10.1177/0003702818767133. arXiv : 1802.03694 [astro-ph.CO] (cf. p. 30).

[94] G. HINSHAW et al. « Five-Year Wilkinson Microwave Anisotropy Probe Observations : Data Processing, Sky Maps, and Basic Results ». In : *apjs* 180.2 (fév. 2009), p. 225-245. DOI : 10.1088/0067-0049/180/2/225. arXiv : 0803.0732 [astro-ph] (cf. p. 77).

[95] Tao HONG et al. « 2MTF - VII. 2MASS Tully-Fisher survey final data release : distances for 2,062 nearby spiral galaxies ». en. In : *Monthly Notices of the Royal Astronomical Society* 487.2 (août 2019). arXiv :1905.08530 [astro-ph], p. 2061-2069. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stz1413 (cf. p. 71, 152, 180, 186).

[96] Jiamin HOU et al. « The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : BAO and RSD measurements from anisotropic clustering analysis of the quasar sample in configuration space between redshift 0.8 and 2.2 ». In : *MNRAS* 500.1 (jan. 2021), p. 1201-1221. DOI : 10.1093/mnras/staa3234. arXiv : 2007.08998 [astro-ph.CO] (cf. p. 87).

[97] Jiamin HOU et al. *Cosmological Probes of Structure Growth and Tests of Gravity*. Juin 2023 (cf. p. 45).

[98] Cullan HOWLETT. « The redshift-space momentum power spectrum - I. Optimal estimation from peculiar velocity surveys ». In : *mnras* 487.4 (août 2019), p. 5209-5234. DOI : 10.1093/mnras/stz1403. arXiv : 1906.02875 [astro-ph.CO] (cf. p. 76).

[99] Cullan HOWLETT, Aaron S. G. ROBOTHAM, Claudia D. P. LAGOS et Alex G. KIM. « Measuring the growth rate of structure with Type IA Supernovae from LSST ». en. In : *The Astrophysical Journal* 847.2 (oct. 2017). arXiv :1708.08236 [astro-ph], p. 128. ISSN : 0004-637X, 1538-4357. DOI : 10.3847/1538-4357/aa88c8 (cf. p. 53).

[100] Cullan HOWLETT, Ashley J. ROSS, Lado SAMUSHIA, Will J. PERCIVAL et Marc MANERA. « The clustering of the SDSS main galaxy sample - II. Mock galaxy catalogues and a measurement of the growth of structure from redshift space distortions at z = 0.15 ». In : *MNRAS* 449.1 (mai 2015), p. 848-866. DOI : 10.1093/mnras/stu2693. arXiv : 1409.3238 [astro-ph.CO] (cf. p. 87).

[101] Cullan HOWLETT et al. « The Sloan Digital Sky Survey Peculiar Velocity Catalogue ». en. In : *Monthly Notices of the Royal Astronomical Society* 515.1 (juill. 2022). arXiv :2201.03112 [astro-ph], p. 953-976. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stac1681 (cf. p. 71, 152, 180, 186).

[102] Yu-Hsiu HUANG, Elisabeth KRAUSE, Jiachuan XU, Tim EIFLER, S. PRANJAL R. et Eric HUFF. « Astrophysical systematics in Kinematic Lensing : quantifying an Intrinsic Alignment analog ». In : *arXiv e-prints*, arXiv :2404.00197 (mars 2024), arXiv :2404.00197. DOI : 10.48550/arXiv.2404.00197. arXiv : 2404.00197 [astro-ph.CO] (cf. p. 180).

[103] Dragan HUTERER. « Growth of Cosmic Structure ». In : *Astronomy and Astrophysics Review* 31 (déc. 2023), p. 2. ISSN : 0935-4956. DOI : 10.1007/s00159-023-00147-4 (cf. p. 45).

[104] Benjamin JOACHIMI. « Non-linear shrinkage estimation of large-scale structure covariance ». In : *mnras* 466.1 (mars 2017), p. L83-L87. DOI : 10.1093/mnrasl/slw240. arXiv : 1612.00752 [astro-ph.IM] (cf. p. 84).

[105] Benjamin JOACHIMI, Sukhdeep SINGH et Rachel MANDELBAUM. « Detection of spatial correlations of Fundamental Plane residuals, and cosmological implications ». In : *mnras* 454.1 (nov. 2015), p. 478-488. DOI : 10.1093/mnras/stv1962. arXiv : 1504.02662 [astro-ph.CO] (cf. p. 18, 180, 187, 199).

[106] D. Heath JONES et al. « The 6dF Galaxy Survey : Samples, Observational Techniques and the First Data Release ». In : *Monthly Notices of the Royal Astronomical Society* 355 (déc. 2004), p. 747-763. ISSN : 0035-8711. DOI : 10.1111/j.1365-2966.2004.08353.x (cf. p. 45).

[107] Nick KAISER. « Clustering in Real Space and in Redshift Space ». In : *Monthly Notices of the Royal Astronomical Society* 227 (juill. 1987), p. 1-21. ISSN : 0035-8711. DOI : 10.1093/mnras/227.1.1 (cf. p. 52).

[108] Rahul KANNAN et al. « The MillenniumTNG project : the galaxy population at z ≥ 8 ». In : *mnras* 524.2 (sept. 2023), p. 2594-2605. DOI : 10.1093/mnras/stac3743. arXiv : 2210.10066 [astro-ph.GA] (cf. p. 182).

[109] Eyal A. KAZIN et al. « The WiggleZ Dark Energy Survey : Improved Distance Measurements to z = 1 with Reconstruction of the Baryonic Acoustic Feature ». en. In : *Monthly Notices of the Royal Astronomical Society* 441.4 (juill. 2014). arXiv :1401.0358 [astro-ph], p. 3524-3542. ISSN : 1365-2966, 0035-8711. DOI : 10.1093/mnras/stu778 (cf. p. 87).

[110] David KIRKBY et al. « Fitting methods for baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9 ». In : *JCAP* 2013.3, 024 (mars 2013), p. 024. DOI : 10.1088/1475-7516/2013/03/024. arXiv : 1301.3456 [astro-ph.CO] (cf. p. 10, 55, 56, 89).

[111] Yosuke KOBAYASHI, Takahiro NISHIMICHI, Masahiro TAKADA et Hironao MIYATAKE. « Full-Shape Cosmology Analysis of the SDSS-III BOSS Galaxy Power Spectrum Using an Emulator-Based Halo Model : A 5% Determination of Σ8 ». In : *Physical Review D* 105 (avr. 2022), p. 083517. ISSN : 1550-79980556-2821. DOI : 10.1103/PhysRevD.105.083517 (cf. p. 111).

[112] Yosuke KOBAYASHI, Takahiro NISHIMICHI, Masahiro TAKADA, Ryuichi TAKAHASHI et Ken OSATO. « Accurate Emulator for the Redshift-Space Power Spectrum of Dark Matter Halos and Its Application to Galaxy Power Spectrum ». In : *Physical Review D* 102.6 (sept. 2020), p. 063504. ISSN : 2470-0010, 2470-0029. DOI : 10.1103/PhysRevD.102.063504 (cf. p. 111).

[113] Jun KODA et al. « Are peculiar velocity surveys competitive as a cosmological probe? » In : *mnras* 445.4 (déc. 2014), p. 4267-4286. DOI : 10.1093/mnras/stu1610. arXiv : 1312.1022 [astro-ph.CO] (cf. p. 58).

[114] Eiichiro KOMATSU et David N. SPERGEL. « Acoustic Signatures in the Primary Microwave Background Bispectrum ». en. In : *Physical Review D* 63.6 (fév. 2001). arXiv :astro-ph/0005036, p. 063002. ISSN : 0556-2821, 1089-4918. DOI : 10.1103/PhysRevD.63.063002 (cf. p. 61).

[115] Andrey V. KRAVTSOV et al. « The Dark Side of the Halo Occupation Distribution ». In : *apj* 609.1 (juill. 2004), p. 35-49. DOI : 10.1086/420959. arXiv : astro-ph/0308519 [astro-ph] (cf. p. 64).

[116] Juliana KWAN et al. « Galaxy Clustering in the Mira-Titan Universe I : Emulators for the redshift space galaxy correlation function and galaxy-galaxy lensing ». en. In : *The Astrophysical Journal* 952.1 (juill. 2023). arXiv :2302.12379 [astro-ph], p. 80. ISSN : 0004-637X, 1538-4357. DOI : 10.3847/1538-4357/acd92f (cf. p. 111, 146).

[117] Stephen D. LANDY et Alexander S. SZALAY. « Bias and Variance of Angular Correlation Functions ». In : *The Astrophysical Journal* 412 (juill. 1993), p. 64. DOI : 10.1086/172900 (cf. p. 73, 89).

[118] Amiaux LAUREIJS et ARDUINI. « Euclid Definition Study Report ». In : (oct. 2011). arXiv :1110.3193 [astro-ph]. DOI : 10.48550/arXiv.1110.3193 (cf. p. 45, 53, 149).

[119] R. LAUREIJS et al. « Euclid Definition Study Report ». In : *arXiv :1110.3193 [astro-ph]* (oct. 2011) (cf. p. 111).

[120] Olivier LEDOIT et Michael WOLF. « Nonlinear shrinkage estimation of large-dimensional covariance matrices ». In : *arXiv e-prints*, arXiv :1207.5322 (juill. 2012), arXiv :1207.5322. DOI : 10.48550/arXiv.1207.5322. arXiv : 1207.5322 [math.ST] (cf. p. 84).

[121] Julien LESGOURGUES et Sergio PASTOR. « Massive neutrinos and cosmology ». en. In : *Physics Reports* 429.6 (juill. 2006). arXiv :astro-ph/0603494, p. 307-379. ISSN : 03701573. DOI : 10.1016/j.physrep.2006.04.001 (cf. p. 25).

[122] Antony LEWIS, Anthony CHALLINOR et Anthony LASENBY. « Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models ». In : *The Astrophysical Journal* 538.2 (août 2000), p. 473-476. DOI : 10.1086/309179. arXiv : astro-ph/9911177 [astro-ph] (cf. p. 10, 55, 89).

[123] Eric V. LINDER et Robert N. CAHN. « Parameterized Beyond-Einstein Growth ». en. In : *Astroparticle Physics* 28.4-5 (déc. 2007). arXiv :astro-ph/0701317, p. 481-488. ISSN : 09276505. DOI : 10.1016/j.astropartphys.2007.09.003 (cf. p. 53).

[124] C. H. LINEWEAVER, L. TENORIO, G. F. SMOOT, P. KEEGSTRA, A. J. BANDAY et P. LUBIN. « The Dipole Observed in the COBE DMR 4 Year Data ». In : *apj* 470 (oct. 1996), p. 38. DOI : 10.1086/177846. arXiv : astro-ph/9601151 [astro-ph] (cf. p. 77).

[125] LSST SCIENCE COLLABORATION et al. « LSST Science Book, Version 2.0 ». In : *arXiv e-prints*, arXiv :0912.0201 (déc. 2009), arXiv :0912.0201. DOI : 10.48550/arXiv.0912.0201. arXiv : 0912.0201 [astro-ph.IM] (cf. p. 18, 68, 199).

[126] Shengdong LU et al. « Redshift evolution of the Fundamental Plane relation in the IllustrisTNG simulation ». In : *mnras* 492.4 (mars 2020), p. 5930-5939. DOI : 10.1093/mnras/staa173. arXiv : 1906.00927 [astro-ph.GA] (cf. p. 180).

[127] Nina A. MAKSIMOVA, Lehman H. GARRISON, Daniel J. EISENSTEIN, Boryana HADZHIYSKA, Sownak BOSE et Thomas P. SATTERTHWAITE. « AbacusSummit : A Massive Set of High-Accuracy, High-Resolution N-Body Simulations ». en. In : *Monthly Notices of the Royal Astronomical Society* 508.3 (oct. 2021). arXiv :2110.11398 [astro-ph], p. 4017-4037. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stab2484 (cf. p. 12, 62, 111, 112, 114).

[128] Federico MARINACCI et al. « First results from the IllustrisTNG simulations : radio haloes and magnetic fields ». In : *mnras* 480.4 (nov. 2018), p. 5113-5139. DOI : 10.1093/mnras/sty2206. arXiv : 1707.03396 [astro-ph.CO] (cf. p. 182).

[129] S. S. MCGAUGH, J. M. SCHOMBERT, G. D. BOTHUN et W. J. G. DE BLOK. « The Baryonic Tully-Fisher Relation ». In : *apjl* 533.2 (avr. 2000), p. L99-L102. DOI : 10.1086/312628. arXiv : astro-ph/0003001 [astro-ph] (cf. p. 189).

[130] Nareg MIRZATUNY et Elena PIERPAOLI. « An Accurate Fitting Function For Scale-dependent Growth Rate in Hu-Sawicki $f(R)$ Gravity ». en. In : *Journal of Cosmology and Astroparticle Physics* 2019.09 (sept. 2019). arXiv :1906.07683 [astro-ph], p. 066-066. ISSN : 1475-7516. DOI : 10.1088/1475-7516/2019/09/066 (cf. p. 53).

[131] Faizan G. MOHAMMAD et Will J. PERCIVAL. « Creating jackknife and bootstrap estimates of the covariance matrix for the two-point correlation function ». In : *mnras* 514.1 (juill. 2022), p. 1289-1301. DOI : 10.1093/mnras/stac1458. arXiv : 2109.07071 [astro-ph.CO] (cf. p. 84).

[132] Jill P. NAIMAN et al. « First results from the IllustrisTNG simulations : a tale of two elements - chemical evolution of magnesium and europium ». In : *mnras* 477.1 (juin 2018), p. 1206-1224. DOI : 10.1093/mnras/sty618. arXiv : 1707.03401 [astro-ph.GA] (cf. p. 182).

[133] Julio F. NAVARRO, Carlos S. FRENK et Simon D. M. WHITE. « The Structure of Cold Dark Matter Halos ». In : *apj* 462 (mai 1996), p. 563. DOI : 10.1086/177173. arXiv : astro-ph/9508025 [astro-ph] (cf. p. 46).

[134] Julio F. NAVARRO, Carlos S. FRENK et Simon D. M. WHITE. « A Universal Density Profile from Hierarchical Clustering ». en. In : *The Astrophysical Journal* 490.2 (déc. 1997). arXiv :astro-ph/9611107, p. 493-508. ISSN : 0004-637X, 1538-4357. DOI : 10.1086/304888 (cf. p. 115).

[135] Dylan NELSON et al. « First results from the IllustrisTNG simulations : the galaxy colour bimodality ». In : *mnras* 475.1 (mars 2018), p. 624-647. DOI : 10.1093/mnras/stx3040. arXiv : 1707.03395 [astro-ph.GA] (cf. p. 182).

[136] Richard NEVEUX et al. « The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : BAO and RSD measurements from the anisotropic power spectrum of the quasar sample between redshift 0.8 and 2.2 ». In : *MNRAS* 499.1 (nov. 2020), p. 210-229. DOI : 10.1093/mnras/staa2780. arXiv : 2007.08999 [astro-ph.CO] (cf. p. 87).

[137] P. NORBERG, C. M. BAUGH, E. GAZTAÑAGA et D. J. CROTON. « Statistical analysis of galaxy surveys - I. Robust error estimation for two-point clustering statistics ». In : *mnras* 396.1 (juin 2009), p. 19-38. DOI : 10.1111/j.1365-2966.2009.14389.x. arXiv : 0810.1885 [astro-ph] (cf. p. 84).

[138] Enrique PAILLAS et al. « Constraining $\nu\Lambda$CDM with density-split clustering ». In : *mnras* 522.1 (juin 2023), p. 606-625. DOI : 10.1093/mnras/stad1017. arXiv : 2209.04310 [astro-ph.CO] (cf. p. 149).

[139] Enrico PAJER et Matias ZALDARRIAGA. « On the Renormalization of the Effective Field Theory of Large Scale Structures ». en. In : *Journal of Cosmology and Astroparticle Physics* 2013.08 (août 2013). arXiv :1301.7182 [astro-ph, physics :gr-qc, physics :hep-ph, physics :hep-th], p. 037-037. ISSN : 1475-7516. DOI : 10.1088/1475-7516/2013/08/037 (cf. p. 61).

[140] Rüdiger PAKMOR et al. « The MillenniumTNG Project : the hydrodynamical full physics simulation and a first look at its galaxy clusters ». In : *mnras* 524.2 (sept. 2023), p. 2539-2555. DOI : 10.1093/mnras/stac3620. arXiv : 2210.10060 [astro-ph.CO] (cf. p. 15, 51, 52, 180-182).

[141] Hauser M. G. PEEBLES P. J. E. en. In : *Astrophysical Journal Supplement* 1 (nov. 1974) (cf. p. 75).

[142] Marcos PELLEJERO IBAÑEZ, Jens STÜCKER, Raul E. ANGULO, Matteo ZENNARO, Sergio CONTRERAS et Giovanni ARICÒ. « Modelling galaxy clustering in redshift space with a Lagrangian bias formalism and N-body simulations ». In : *mnras* 514.3 (août 2022), p. 3993-4007. DOI : 10.1093/mnras/stac1602. arXiv : 2109.08699 [astro-ph.CO] (cf. p. 63).

[143] Marcos PELLEJERO IBAÑEZ et al. « The bacco simulation project : bacco hybrid Lagrangian bias expansion model in redshift space ». In : *mnras* 520.3 (avr. 2023), p. 3725-3741. DOI : 10.1093/mnras/stad368. arXiv : 2207.06437 [astro-ph.CO] (cf. p. 18, 198).

[144] Marco PELOSO, Lorenzo SORBO et Gianmassimo TASINATO. « A falsely fat curvaton with an observable running of the spectral tilt ». en. In : *Journal of Cosmology and Astroparticle Physics* 2014.06 (juin 2014). arXiv :1401.7136 [astro-ph, physics :gr-qc, physics :hep-ph, physics :hep-th], p. 040-040. ISSN : 1475-7516. DOI : 10.1088/1475-7516/2014/06/040 (cf. p. 35).

[145] Will J. PERCIVAL et Davide BIANCHI. « Using angular pair upweighting to improve 3D clustering measurements ». In : *mnras* 472.1 (nov. 2017), p. L40-L44. DOI : 10.1093/mnrasl/slx135. arXiv : 1703.02071 [astro-ph.CO] (cf. p. 72).

[146] Will J. PERCIVAL, Oliver FRIEDRICH, Elena SELLENTIN et Alan HEAVENS. « Matching Bayesian and frequentist coverage probabilities when using an approximate data covariance matrix ». In : *mnras* 510.3 (mars 2022), p. 3207-3221. DOI : 10.1093/mnras/stab3540. arXiv : 2108.10402 [astro-ph.IM] (cf. p. 160).

[147] Will J. PERCIVAL et al. « The clustering of Galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey : including covariance matrix errors ». In : *MNRAS* 439.3 (avr. 2014), p. 2531-2541. DOI : 10.1093/mnras/stu112. arXiv : 1312.4841 [astro-ph.CO] (cf. p. 83, 94).

[148] Oliver H. E. PHILCOX, Daniel J. EISENSTEIN, Ross O'CONNELL et Alexander WIEGAND. « RASCALC : a jackknife approach to estimating single- and multitracer galaxy covariance matrices ». In : *mnras* 491.3 (jan. 2020), p. 3290-3317. DOI : 10.1093/mnras/stz3218. arXiv : 1904.11070 [astro-ph.CO] (cf. p. 84).

[149] Annalisa PILLEPICH et al. « First results from the IllustrisTNG simulations : the stellar mass content of groups and clusters of galaxies ». In : *mnras* 475.1 (mars 2018), p. 648-675. DOI : 10.1093/mnras/stx3112. arXiv : 1707.03406 [astro-ph.GA] (cf. p. 182).

[150] PLANCK COLLABORATION et al. « Planck 2015 results. XIII. Cosmological parameters ». In : *aap* 594, A13 (sept. 2016), A13. DOI : 10.1051/0004-6361/201525830. arXiv : 1502.01589 [astro-ph.CO] (cf. p. 182).

[151] PLANCK COLLABORATION et al. « Planck 2018 Results. VI. Cosmological Parameters ». In : *Astronomy and Astrophysics* 641 (sept. 2020), A6. ISSN : 0004-6361. DOI : 10.1051/0004-6361/201833910 (cf. p. 12, 30, 32, 35, 39, 50, 53).

[152] F. PRADA et al. *The DESI One-Percent Survey : Modelling the clustering and halo occupation of all four DESI tracers with Uchuu*. en. arXiv :2306.06315 [astro-ph]. Juin 2023 (cf. p. 66, 67, 115).

[153] William H. PRESS et Paul SCHECHTER. « Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation ». In : *apj* 187 (fév. 1974), p. 425-438. DOI : 10.1086/152650 (cf. p. 63).

[154] Fei QIN, Cullan HOWLETT et Lister STAVELEY-SMITH. « The redshift-space momentum power spectrum - II. Measuring the growth rate from the combined 2MTF and 6dFGSv surveys ». In : *Monthly Notices of the Royal Astronomical Society* 487 (août 2019), p. 5235-5247. ISSN : 0035-8711. DOI : 10.1093/mnras/stz1576 (cf. p. 72).

[155] Anand RAICHOOR et al. « The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : large-scale structure catalogues and measurement of the isotropic BAO between redshift 0.6 and 1.1 for the Emission Line Galaxy Sample ». In : *MNRAS* 500.3 (jan. 2021), p. 3254-3274. DOI : 10.1093/mnras/staa3336. arXiv : 2007.09007 [astro-ph.CO] (cf. p. 87).

[156] Carl Edward RASMUSSEN et Christopher K. I. WILLIAMS. *Gaussian processes for machine learning*. en. Adaptive computation and machine learning. OCLC : ocm61285753. Cambridge, Mass : MIT Press, 2006. ISBN : 978-0-262-18253-9 (cf. p. 118).

[157] Beth A. REID et Martin WHITE. « Towards an accurate model of the redshift-space clustering of haloes in the quasi-linear regime : Redshift-space distortions ». en. In : *Monthly Notices of the Royal Astronomical Society* 417.3 (nov. 2011), p. 1913-1927. ISSN : 00358711. DOI : 10.1111/j.1365-2966.2011.19379.x (cf. p. 61).

[158] Ashley J. ROSS, Will J. PERCIVAL et Marc MANERA. « The information content of anisotropic Baryon Acoustic Oscillation scale measurements ». In : *MNRAS* 451.2 (août 2015), p. 1331-1340. DOI : 10.1093/mnras/stv966. arXiv : 1501.05571 [astro-ph.CO] (cf. p. 92).

[159] Ashley J. ROSS, Lado SAMUSHIA, Cullan HOWLETT, Will J. PERCIVAL, Angela BURDEN et Marc MANERA. « The clustering of the SDSS DR7 main Galaxy sample - I. A 4 per cent distance measure at z = 0.15 ». In : *MNRAS* 449.1 (mai 2015), p. 835-847. DOI : 10.1093/mnras/stv154. arXiv : 1409.3242 [astro-ph.CO] (cf. p. 87).

[160] Ashley J. ROSS et al. « The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey : observational systematics and baryon acoustic oscillations in the correlation function ». In : *MNRAS* 464.1 (jan. 2017), p. 1168-1191. DOI : 10.1093/mnras/stw2372. arXiv : 1607.03145 [astro-ph.CO] (cf. p. 10, 55).

[161] Ashley J. ROSS et al. « The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : Large-scale structure catalogues for cosmological analysis ». In : *MNRAS* 498.2 (oct. 2020), p. 2354-2371. DOI : 10.1093/mnras/staa2416. arXiv : 2007.09000 [astro-ph.CO] (cf. p. 88).

[162] Yunus SAATCI. « Scalable Inference for Structured Gaussian Process Models ». Thèse de doct. University of Cambridge, 2011 (cf. p. 118, 125).

[163] Ariel G. SÁNCHEZ et al. « The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey : combining correlated Gaussian posterior distributions ». In : *MNRAS* 464.2 (jan. 2017), p. 1493-1501. DOI : 10.1093/mnras/stw2495. arXiv : 1607.03146 [astro-ph.CO] (cf. p. 10, 87, 88, 91, 92, 197).

[164] Christoph SAULDER, Steffen MIESKE, Werner W. ZEILINGER et Igor V. CHILINGARIAN. « Calibrating the fundamental plane with SDSS DR8 data ». en. In : *Astronomy & Astrophysics* 557 (sept. 2013). arXiv :1306.0285 [astro-ph], A21. ISSN : 0004-6361, 1432-0746. DOI : 10.1051/0004-6361/201321466 (cf. p. 71, 180, 186).

[165] Christoph SAULDER et al. *Target Selection for the DESI Peculiar Velocity Survey*. Fév. 2023. DOI : 10.48550/arXiv.2302.13760 (cf. p. 14, 18, 152, 180, 184, 198, 199).

[166] Christoph SAULDER et al. « Target selection for the DESI Peculiar Velocity Survey ». In : *mnras* 525.1 (oct. 2023), p. 1106-1125. DOI : 10.1093/mnras/stad2200. arXiv : 2302.13760 [astro-ph.CO] (cf. p. 68, 69).

[167] Román SCOCCIMARRO. « Fast estimators for redshift-space clustering ». In : *prd* 92.8, 083532 (oct. 2015), p. 083532. DOI : 10.1103/PhysRevD.92.083532. arXiv : 1506.02729 [astro-ph.CO] (cf. p. 76).

[168] Román SCOCCIMARRO. « Fast estimators for redshift-space clustering ». In : *PRD* 92.8, 083532 (oct. 2015), p. 083532. DOI : 10.1103/PhysRevD.92.083532. arXiv : 1506.02729 [astro-ph.CO] (cf. p. 89).

[169] Ryan SCRANTON et al. « Analysis of Systematic Effects and Statistical Uncertainties in Angular Clustering of Galaxies from Early Sloan Digital Sky Survey Data ». In : *apj* 579.1 (nov. 2002), p. 48-75. DOI : 10.1086/342786. arXiv : astro-ph/0107416 [astro-ph] (cf. p. 77).

[170] Uros SELJAK. « Analytic model for galaxy and dark matter clustering ». en. In : *Monthly Notices of the Royal Astronomical Society* 318.1 (oct. 2000). arXiv :astro-ph/0001493, p. 203-213. ISSN : 0035-8711, 1365-2966. DOI : 10.1046/j.1365-8711.2000.03715.x (cf. p. 46).

[171] Hee-Jong SEO, Florian BEUTLER, Ashley J. ROSS et Shun SAITO. « Modeling the reconstructed BAO in Fourier space ». In : *MNRAS* 460.3 (août 2016), p. 2453-2471. DOI : 10.1093/mnras/stw1138. arXiv : 1511.00663 [astro-ph.CO] (cf. p. 10, 55).

[172] Stephen A. SMEE et al. « The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey ». In : *The Astrophysical Journal* 146.2, 32 (août 2013), p. 32. DOI : 10.1088/0004-6256/146/2/32. arXiv : 1208.2233 [astro-ph.IM] (cf. p. 88).

[173] Alex SMITH, Arnaud DE MATTIA, Etienne BURTIN, Chia-Hsun CHUANG et Cheng ZHAO. « Reducing the variance of redshift space distortion measurements from mock galaxy catalogues with different lines of sight ». In : *mnras* 500.1 (jan. 2021), p. 259-271. DOI : 10.1093/mnras/staa3244. arXiv : 2007.11417 [astro-ph.CO] (cf. p. 186).

[174] G. F. SMOOT et al. « Structure in the COBE Differential Microwave Radiometer First-Year Maps ». In : *apjl* 396 (sept. 1992), p. L1. DOI : 10.1086/186504 (cf. p. 30).

[175] Volker SPRINGEL. « E pur si muove : Galilean-invariant cosmological hydrodynamical simulations on a moving mesh ». In : *mnras* 401.2 (jan. 2010), p. 791-851. DOI : 10.1111/j.1365-2966.2009.15715.x. arXiv : 0901.4107 [astro-ph.CO] (cf. p. 182).

[176] Volker SPRINGEL et al. « Simulations of the formation, evolution and clustering of galaxies and quasars ». In : *nat* 435.7042 (juin 2005), p. 629-636. DOI : 10.1038/nature03597. arXiv : astro-ph/0504097 [astro-ph] (cf. p. 182).

[177] Volker SPRINGEL et al. « First results from the IllustrisTNG simulations : matter and galaxy clustering ». In : *mnras* 475.1 (mars 2018), p. 676-698. DOI : 10.1093/mnras/stx3304. arXiv : 1707.03397 [astro-ph.GA] (cf. p. 182).

[178] Julia STADLER, Fabian SCHMIDT et Martin REINECKE. « Cosmology inference at the field level from biased tracers in redshift-space ». In : *jcap* 2023.10, 069 (oct. 2023), p. 069. DOI : 10.1088/1475-7516/2023/10/069. arXiv : 2303.09876 [astro-ph.CO] (cf. p. 149).

[179] Michael A STRAUSS. « The density and peculiar velocity elds of nearby galaxies ». en. In : (1995) (cf. p. 77).

[180] Masahiro TAKADA et Wayne HU. « Power spectrum super-sample covariance ». In : *prd* 87.12, 123504 (juin 2013), p. 123504. DOI : 10.1103/PhysRevD.87.123504. arXiv : 1302.6994 [astro-ph.CO] (cf. p. 84).

[181] Amélie TAMONE et al. « The completed SDSS-IV extended baryon oscillation spectroscopic survey : growth rate of structure measurement from anisotropic clustering analysis in configuration space between redshift 0.6 and 1.1 for the emission-line galaxy sample ». In : *MNRAS* 499.4 (déc. 2020), p. 5527-5546. DOI : 10.1093/mnras/staa3050. arXiv : 2007.09009 [astro-ph.CO] (cf. p. 87).

[182] Atsushi TARUYA, Takahiro NISHIMICHI et Shun SAITO. « Baryon Acoustic Oscillations in 2D : Modeling Redshift-space Power Spectrum from Perturbation Theory ». en. In : *Physical Review D* 82.6 (sept. 2010). arXiv :1006.0699 [astro-ph, physics :gr-qc], p. 063522. ISSN : 1550-7998, 1550-2368. DOI : 10.1103/PhysRevD.82.063522 (cf. p. 61).

[183] Andy TAYLOR, Benjamin JOACHIMI et Thomas KITCHING. « Putting the precision in precision cosmology : How accurate should your data covariance matrix be ? » In : *MNRAS* 432.3 (juill. 2013), p. 1928-1946. DOI : 10.1093/mnras/stt270. arXiv : 1212.4359 [astro-ph.CO] (cf. p. 83, 93, 94, 97).

[184] Max TEGMARK, Andy N. TAYLOR et Alan F. HEAVENS. « Karhunen-Loeve Eigenvalue Problems in Cosmology : How Should We Tackle Large Data Sets ? » en. In : *The Astrophysical Journal* 480.1 (mai 1997), p. 22-35. ISSN : 0004-637X, 1538-4357. DOI : 10.1086/303939 (cf. p. 82).

[185] Roberto TROTTA. « Bayesian Methods in Cosmology ». In : *arXiv e-prints*, arXiv :1701.01467 (jan. 2017), arXiv :1701.01467. DOI : 10.48550/arXiv.1701.01467. arXiv : 1701.01467 [astro-ph.CO] (cf. p. 78).

[186] R. B. TULLY et J. R. FISHER. « A new method of determining distances to galaxies. » In : *aap* 54 (fév. 1977), p. 661-673 (cf. p. 69).

[187] R. Brent TULLY et al. « Cosmicflows-4 ». en. In : *The Astrophysical Journal* 944.1 (fév. 2023). arXiv :2209.11238 [astro-ph], p. 94. ISSN : 0004-637X, 1538-4357. DOI : 10.3847/1538-4357/ac94d8 (cf. p. 71, 180, 186).

[188] Ryan J. TURNER, Chris BLAKE et Rossana RUGGERI. « A Local Measurement of the Growth Rate from Peculiar Velocities and Galaxy Clustering Correlations in the 6dF Galaxy Survey ». In : *Monthly Notices of the Royal Astronomical Society* 518 (jan. 2023), p. 2436-2452. ISSN : 0035-8711. DOI : 10.1093/mnras/stac3256 (cf. p. 13, 54, 149).

[189] Georgios VALOGIANNIS et Cora DVORKIN. « Going beyond the galaxy power spectrum : An analysis of BOSS data with wavelet scattering transforms ». In : *prd* 106.10, 103509 (nov. 2022), p. 103509. DOI : 10.1103/PhysRevD.106.103509. arXiv : 2204.13717 [astro-ph.CO] (cf. p. 149).

[190] Francisco VILLAESCUSA-NAVARRO et al. « The imprint of neutrinos on clustering in redshift-space ». en. In : *The Astrophysical Journal* 861.1 (juill. 2018). arXiv :1708.01154 [astro-ph], p. 53. ISSN : 1538-4357. DOI : 10.3847/1538-4357/aac6bf (cf. p. 43).

[191] Francisco VILLAESCUSA-NAVARRO et al. « The Quijote simulations ». In : *The Astrophysical Journal Supplement Series* 250.1 (août 2020). arXiv :1909.05273 [astro-ph], p. 2. ISSN : 1538-4365. DOI : 10.3847/1538-4365/ab9d82 (cf. p. 111).

[192] Mark VOGELSBERGER, Federico MARINACCI, Paul TORREY et Ewald PUCHWEIN. « Cosmological simulations of galaxy formation ». In : *Nature Reviews Physics* 2.1 (jan. 2020), p. 42-66. DOI : 10.1038/s42254-019-0127-2. arXiv : 1909.07976 [astro-ph.GA] (cf. p. 64, 65).

[193] Mark VOGELSBERGER et al. « Introducing the Illustris Project : simulating the coevolution of dark and visible matter in the Universe ». In : *mnras* 444.2 (oct. 2014), p. 1518-1547. DOI : 10.1093/mnras/stu1536. arXiv : 1405.2921 [astro-ph.CO] (cf. p. 180).

[194] Lile WANG, Beth REID et Martin WHITE. « An analytic model for redshift-space distortions ». en. In : *Monthly Notices of the Royal Astronomical Society* 437.1 (jan. 2014), p. 588-599. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stt1916 (cf. p. 61).

[195] Yuting WANG et Gong-Bo ZHAO. « A brief review on cosmological analysis of galaxy surveys with multiple tracers ». In : *Research in Astronomy and Astrophysics* 20.10, 158 (oct. 2020), p. 158. DOI : 10.1088/1674-4527/20/10/158. arXiv : 2009.03862 [astro-ph.CO] (cf. p. 149).

[196] Douglas F. WATSON, Andreas A. BERLIND et Andrew R. ZENTNER. « Constraining Satellite Galaxy Stellar Mass Loss and Predicting Intrahalo Light. I. Framework and Results at Low Redshift ». In : *apj* 754.2, 90 (août 2012), p. 90. DOI : 10.1088/0004-637X/754/2/90. arXiv : 1201.2407 [astro-ph.CO] (cf. p. 195).

[197] J Beau W WEBBER. « A bi-symmetric log transformation for wide-range data ». In : *Measurement Science and Technology* 24.2 (déc. 2012), p. 027001. DOI : 10.1088/0957-0233/24/2/027001 (cf. p. 156).

[198] David H. WEINBERG, Michael J. MORTONSON, Daniel J. EISENSTEIN, Christopher HIRATA, Adam G. RIESS et Eduardo ROZO. « Observational Probes of Cosmic Acceleration ». In : *Physics Reports*. Observational Probes of Cosmic Acceleration 530.2 (sept. 2013), p. 87-255. ISSN : 0370-1573. DOI : 10.1016/j.physrep.2013.05.001 (cf. p. 45).

[199] Andrew R. WETZEL et Martin WHITE. « What determines satellite galaxy disruption? » In : *mnras* 403.2 (avr. 2010), p. 1072-1088. DOI : 10.1111/j.1365-2966.2009.16191.x. arXiv : 0907.0702 [astro-ph.CO] (cf. p. 195).

[200] Simon D. M. WHITE et Carlos S. FRENK. « Galaxy Formation through Hierarchical Clustering ». In : *apj* 379 (sept. 1991), p. 52. DOI : 10.1086/170483 (cf. p. 18, 64, 198).

[201] M. J. WILSON, J. A. PEACOCK, A. N. TAYLOR et S. DE LA TORRE. « Rapid modelling of the redshift-space power spectrum multipoles for a masked density field ». In : *mnras* 464.3 (jan. 2017), p. 3121-3130. DOI : 10.1093/mnras/stw2576. arXiv : 1511.07799 [astro-ph.CO] (cf. p. 77).

[202] Xiaoju XU, Idit ZEHAVI et Sergio CONTRERAS. « Dissecting and modelling galaxy assembly bias ». In : *mnras* 502.3 (avr. 2021), p. 3242-3263. DOI : 10.1093/mnras/stab100. arXiv : 2007.05545 [astro-ph.GA] (cf. p. 169).

[203] Xiaoying XU, Antonio J. CUESTA, Nikhil PADMANABHAN, Daniel J. EISENSTEIN et Cameron K. MCBRIDE. « Measuring DA and H at z=0.35 from the SDSS DR7 LRGs using baryon acoustic oscillations ». en. In : *Monthly Notices of the Royal Astronomical Society* 431.3 (mai 2013), p. 2834-2860. ISSN : 1365-2966, 0035-8711. DOI : 10.1093/mnras/stt379 (cf. p. 165).

[204] Kazuhiro YAMAMOTO, Masashi NAKAMICHI, Akinari KAMINO, Bruce A. BASSETT et Hiroaki NISHIOKA. « A Measurement of the Quadrupole Power Spectrum in the Clustering of the 2dF QSO Survey ». In : *pasj* 58 (fév. 2006), p. 93-102. DOI : 10.1093/pasj/58.1.93. arXiv : astro-ph/0505115 [astro-ph] (cf. p. 76).

[205] Kazuhiro YAMAMOTO, Masashi NAKAMICHI, Akinari KAMINO, Bruce A. BASSETT et Hiroaki NISHIOKA. « A Measurement of the Quadrupole Power Spectrum in the Clustering of the 2dF QSO Survey ». In : *PASJ* 58 (fév. 2006), p. 93-102. DOI : 10.1093/pasj/58.1.93. arXiv : astro-ph/0505115 [astro-ph] (cf. p. 89).

[206] Victoria YANKELEVICH et Cristiano PORCIANI. « Cosmological information in the redshift-space bispectrum ». In : *mnras* 483.2 (fév. 2019), p. 2078-2099. DOI : 10.1093/mnras/sty3143. arXiv : 1807.07076 [astro-ph.CO] (cf. p. 149).

[207] Jaiyul YOO et Uros SELJAK. « Wide Angle Effects in Future Galaxy Surveys ». en. In : *Monthly Notices of the Royal Astronomical Society* 447.2 (fév. 2015). arXiv :1308.1093 [astro-ph], p. 1789-1805. ISSN : 1365-2966, 0035-8711. DOI : 10.1093/mnras/stu2491 (cf. p. 78, 198).

[208] Sihan YUAN, Lehman H. GARRISON, Daniel J. EISENSTEIN et Risa H. WECHSLER. « Stringent σ_8 constraints from small-scale galaxy clustering using a hybrid MCMC+emulator framework ». en. In : *Monthly Notices of the Royal Astronomical Society* 515.1 (juill. 2022). arXiv :2203.11963 [astro-ph], p. 871-896. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stac1830 (cf. p. 111, 137, 146).

[209] Sihan YUAN, Lehman H. GARRISON, Boryana HADZHIYSKA, Sownak BOSE et Daniel J. EISENSTEIN. « AbacusHOD : A highly efficient extended multi-tracer HOD framework and its application to BOSS and eBOSS data ». en. In : *Monthly Notices of the Royal Astronomical Society* 510.3 (jan. 2022). arXiv :2110.11412 [astro-ph], p. 3301-3320. ISSN : 0035-8711, 1365-2966. DOI : 10.1093/mnras/stab3355 (cf. p. 115).

[210] Sihan YUAN, Alvaro ZAMORA et Tom ABEL. « 2D k-th nearest neighbour statistics : a highly informative probe of galaxy clustering ». In : *mnras* 522.3 (juill. 2023), p. 3935-3947. DOI : 10.1093/mnras/stad1275. arXiv : 2304.10565 [astro-ph.CO] (cf. p. 149).

[211] Ya. B. ZEL'DOVICH. « Gravitational instability : An approximate theory for large density perturbations. » In : *aap* 5 (mars 1970), p. 84-89 (cf. p. 55).

[212] Zhongxu ZHAI, Michael BLANTON, Anže SLOSAR et Jeremy TINKER. « An Evaluation of Cosmological Models from the Expansion and Growth of Structure Measurements ». In : *The Astrophysical Journal* 850.2 (déc. 2017), p. 183. ISSN : 0004-637X. DOI : 10.3847/1538-4357/aa9888 (cf. p. 45).

[213] Zhongxu ZHAI et al. « The Aemulus Project V : Cosmological constraint from small-scale clustering of BOSS galaxies ». en. In : *arXiv :2203.08999 [astro-ph]* (mars 2022). arXiv : 2203.08999 (cf. p. 111, 137, 146).

[214] Cheng ZHAO et al. « The completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey : 1000 multi-tracer mock catalogues with redshift evolution and systematics for galaxies and quasars of the final data release ». In : *MNRAS* 503.1 (mai 2021), p. 1149-1173. DOI : 10.1093/mnras/stab510. arXiv : 2007.08997 [astro-ph.CO] (cf. p. 89).

[215] Zheng ZHENG, Alison L. COIL et Idit ZEHAVI. « Galaxy Evolution from Halo Occupation Distribution Modeling of DEEP2 and SDSS Galaxy Clustering ». en. In : *The Astrophysical Journal* 667.2 (oct. 2007), p. 760-779. ISSN : 0004-637X, 1538-4357. DOI : 10.1086/521074 (cf. p. 12, 113).

[216] F. ZWICKY. « On the Masses of Nebulae and of Clusters of Nebulae ». In : *apj* 86 (oct. 1937), p. 217. DOI : 10.1086/143864 (cf. p. 30).