

HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Astronomy

have examined a dissertation entitled

**Enhancing the analysis of the large-scale structure of the Universe for cutting-edge
cosmological surveys with two-point correlation function and beyond**

presented by Mykhailo Rashkovetskyi

candidate for the degree of Doctor of Philosophy and hereby

certify that it is worthy of acceptance.

Signature 

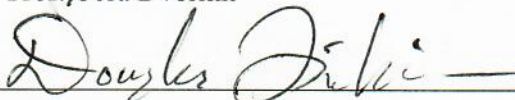
Typed name: Prof. John Kovac, Chair

Signature 

Typed name: Prof. Daniel Eisenstein

Signature 

Typed name: Prof. Cora Dvorkin

Signature 

Typed name: Prof. Douglas Finkbeiner

Signature 

Typed name: Prof. Ravi Sheth

Date: May 5, 2025

Enhancing the analysis of the large-scale structure of the Universe for cutting-edge cosmological surveys with two-point correlation function and beyond

A dissertation presented

by

Mykhailo Rashkovetskyi

to

The Department of Astronomy

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Astronomy and Astrophysics

Harvard University

Cambridge, Massachusetts

May 2025

© 2025 Mykhailo Rashkovetskyi

All rights reserved.

**Enhancing the analysis of the large-scale structure of the Universe for
cutting-edge cosmological surveys with two-point correlation function and
beyond**

Abstract

For many years, we have known that the Universe is vast and expands with acceleration on the largest scales. However, dark energy, the substance supposedly driving this acceleration, might also weaken with time. Or at least the current standard model of cosmology has a problem explaining all the highest-quality data available. We present one of the key technical ingredients that enabled the Dark Energy Spectroscopic Instrument (DESI) Baryon Acoustic Oscillation (BAO) distance measurements — the semi-analytic covariance matrices for the two-point correlation functions of point tracers. We then briefly discuss the cosmological implications of DESI BAO results, including the suggestion of dark energy. Then, we discuss a curious possibility of relieving the Hubble tension, the discrepancy in the expansion rate of the Universe today obtained directly from a Hubble diagram versus inferred indirectly from the CMB, without introducing fundamentally new physics. In the end, we explore a novel analysis technique combining galaxy redshift surveys with the data from the thermal Sunyaev-Zeldovich effect, a secondary anisotropy in the CMB. We aim to interpretably extract more valuable cosmological information from both than standard 2-point summary statistics allow. This would enable better consistency tests for the concordance model of cosmology and potentially new, exciting discoveries.

Table of Contents

Title page	i
Copyright	ii
Abstract	iii
Table of Contents	iv
List of Tables	viii
List of Figures	x
Acknowledgments	xiii
Dedication	xviii
1 Introduction	1
1.1 Cosmological big questions	1
1.2 Selected cosmological probes	2
1.2.1 Large-scale structure of the Universe: overview	2
1.2.2 Galaxy surveys	2
1.2.3 Baryon acoustic oscillations (BAO)	3
1.2.4 Cosmic microwave background (CMB)	5
1.2.5 Standard candles: Type Ia supernovae	7
1.3 Dark Energy Spectroscopic Instrument (DESI)	8
1.4 Statistics of the Universe on large scales	9
1.4.1 Overdensity field	9
1.4.2 Gaussian random field	10
1.4.3 The 2-point correlation function (2PCF)	11
1.4.4 The power spectrum	13
1.4.5 Higher-point statistics	13
1.5 Large-scale structure data analysis and covariance matrices	14
1.6 Outline of the thesis	15
2 Fast semi-analytical covariance matrices for two-point correlation functions for DESI data	17
2.1 Introduction	17
2.2 Methods of covariance matrix estimation	23

2.2.1	Overview of previous work	23
2.2.2	Split random-random computation	32
2.2.3	Reconstructed two-point function covariance	33
2.2.4	Revisited covariance for projected Legendre moments of 2PCF	34
2.3	Methods of comparison of covariance matrices	39
2.3.1	Interpretable measures of similarity for covariance matrices (original version, needs to be merged with the other one)	39
2.3.2	Internal convergence assessment	43
2.3.3	Fisher projection to the space of model parameters	44
2.4	Application to DESI-M2 LRG mocks	45
2.4.1	Mock catalogs and reconstruction method	46
2.4.2	Setup	46
2.4.3	Internal convergence checks	48
2.4.4	Shot-noise rescaling values	49
2.4.5	Measurement-space validation	51
2.4.6	Parameter-space validation	53
2.4.7	Errorbars on BAO scale parameter	55
2.5	Application to DESI DR1 mocks	58
2.5.1	DESI DR1 simulations and methods	58
2.5.2	Setup for semi-analytical covariance matrices	61
2.5.3	Intrinsic tests of the method	63
2.5.4	Comparison between semi-analytical and mock sample covariance matrices	67
2.6	Conclusions	73
3	Key DESI BAO results	81
3.1	Introduction	81
3.2	Early DESI data (2023)	82
3.3	DESI DR1 (2024)	82
3.3.1	BAO measurements	82
3.3.2	Cosmology: dynamic dark energy	84
3.4	DESI DR2 (2025)	87
3.4.1	BAO measurements	87
3.4.2	Cosmological implications	89
4	Hubble tension and small-scale clumping at recombination.	94
4.1	Introduction	94
4.2	Three-zone model (M3) for recombination	97
4.3	The effects of small-scale gas clumping on the CMB	99
4.3.1	Recombination	99

4.3.2	Sound horizon at last scattering	103
4.3.3	Silk damping	104
4.4	Results with <i>Planck</i> 2018	106
4.4.1	H_0 and model comparison	110
4.4.2	Low- ℓ <i>Planck</i> data analysis	111
4.4.3	Damping scale	113
4.4.4	S_8 tension	116
4.4.5	Baryon drag scale	118
4.5	Forecasts for future CMB experiments	125
4.5.1	Fiducial with standard recombination	126
4.5.2	Fiducial with clumping	127
4.6	Conclusions	128
4.6.1	Subsequent developments	130
5	Galaxy selection using the thermal Sunyaev-Zeldovich effect (a way beyond 2-point)	132
5.1	Introduction	132
5.2	Data	135
5.2.1	Galaxy catalog: DESI DR1 LRG	135
5.2.2	Sunyaev-Zeldovich map: ACT DR6	135
5.2.3	Simulations: ABACUSSUMMIT halo catalogs	136
5.3	Measurements	136
5.3.1	Methodology and challenges	136
5.3.2	Larger-scale clustering and galaxy bias	138
5.3.3	Small-scale line-of-sight clustering and velocity dispersions	140
5.4	Simulation-based toy model	144
5.4.1	Simple y signal map from an N -body simulation	144
5.4.2	Simple noise model	148
5.4.3	Calibration	148
5.4.4	Mock clustering	149
5.4.5	Relating halo mass to thermal Sunyaev-Zeldovich signatures	152
5.5	Conclusions	153
6	Conclusions and outlook	156
	References	161

A	Appendix to the covariance matrices (Chapter 2)	172
A.1	Estimators for multi-tracer covariance matrices	172
A.1.1	Angular bins	172
A.1.2	Projected Legendre	174
A.2	Statistics of comparison metrics for noisy sample covariance matrix	175
A.2.1	KL divergence mean and variance	175
A.2.2	Inverse test	180
A.2.3	Mean chi-squared	183
A.2.4	Validation of means and standard deviations	183
A.3	Covariance for the combination of two regions	184
A.3.1	Angular bins	185
A.3.2	Legendre	186
B	Appendix to Hubble tension with clumpy recombination (Chapter 4)	188
B.1	Justifying simplifications	188
B.1.1	RECFAST vs HyREC	188
B.1.2	Neutrino masses	190
B.2	Full contours from <i>Planck</i> runs	190
B.3	Best fit parameters	192
C	Appendix to thermal Sunyaev-Zeldovich galaxy selection (Chapter 5)	195
C.1	Distribution of numbers of close neighbors for LRG in different tSZ SNR bins	195
C.2	Efficient method for near-neighbor search in cylinders	196
C.2.1	Fixed line of sight	197
C.2.2	Varying (midpoint) line of sight	198

List of Tables

2.1	RASCALC convergence test results for DESI-M2 mocks	49
2.2	RASCALC shot-noise rescaling values for DESI-M2 mocks	50
2.3	General (measurement-space) comparison between RASCALC and sample covariance for DESI-M2 mocks before reconstruction	51
2.4	General (measurement-space) comparison between RASCALC and sample covariance for DESI-M2 mocks after reconstruction	52
2.5	Parameter-space (for BAO model) comparison between RASCALC and sample covariance for DESI-M2 mocks before reconstruction	54
2.6	Parameter-space (for BAO model) comparison between RASCALC and sample covariance for DESI-M2 mocks after reconstruction	55
2.7	Errorbars on the BAO isotropic scale parameter obtained from different covariances for the DESI-M2 mocks	56
2.8	Tracers and redshift bins used in DESI DR1 mocks validation	62
2.9	Shot-noise rescaling values for DESI DR1 mocks (one type) and data.	65
2.10	Shot-noise rescaling values for the data and two different types of mocks with two different fiber assignment models	65
2.11	Shot-noise rescaling values for the single-mock runs calibrated on jackknife (Figure 2.1) and mock sample covariances (Figure 2.2).	66
2.12	Full observable-space comparison of RASCALC covariances with the DESI DR1 mocks sample covariances	68
2.13	Observable-space comparison of RASCALC covariances with the DESI DR1 mocks sample covariances restricted to the range of ShapeFit and full modeling fits	69
2.14	Observable-space comparison of RASCALC covariances with the DESI DR1 mocks sample covariances restricted to the range of BAO fits	70
2.15	Parameter-space comparison of RASCALC covariances with the DESI DR1 mocks sample covariances projected to the BAO fit parameters	71
2.16	Parameter-space comparison of RASCALC covariances with the DESI DR1 mocks sample covariances projected to the ShapeFit parameters	73

2.17	Parameter-space comparison of RASCALC covariances with the DESI DR1 mocks sample covariances projected to the full modeling parameters	75
4.1	Parameters used in this chapter and their priors	108
4.2	Comparison between the clumping and the standard model with <i>Planck</i> 2018 data	111
4.3	Comparison between the clumping and the standard model with <i>Planck</i> $\ell < 1000$ data	113
4.4	Fiducial parameters for our forecasts.	125
4.5	Clumping vs standard model comparison forecast with a standard recombination fiducial	126
4.6	Clumping vs standard model comparison forecast with a clumpy recombination fiducial	128
5.1	Number of galaxies in our tSZ SNR bins	140
A.1	Theoretical and empirical statistical properties of covariance matrix comparison metrics	184
B.1	Our best fit parameters to (full) <i>Planck</i> and <i>Planck</i> +SH0ES.	193
B.2	Our best fit parameters to <i>Planck</i> $\ell < 1000$, without and with SH0ES.	194
C.1	Comparison of different pair pre-selection methods' efficiency	199

List of Figures

1.1	Map of the cosmic microwave background (CMB) temperature fluctuations measured by the <i>Planck</i> satellite. (Image credit: ESA and the Planck Collaboration.)	6
1.2	Cosmic microwave background (CMB) temperature power spectrum measured by the <i>Planck</i> satellite (red dots with errorbars) and predicted by the Λ CDM model (green). (Image credit: ESA and the Planck Collaboration.)	7
2.1	RASCALC jackknife (fiducial) pipeline flowchart	24
2.2	RASCALC mocks (alternative) pipeline flowchart	25
2.3	Errorbars on the BAO isotropic scale parameter obtained from different covariances for the DESI-M2 mocks	57
2.4	Comparison of the projected errorbars for the BAO scale parameters for DESI DR1 mocks	72
2.5	Comparison of the projected errorbars for the BAO scale parameters for DESI DR1 mocks from ShapeFit	74
3.1	BAO in the Early DESI data	83
3.2	Detection significance of the BAO features in various DESI DR1 galaxy tracers	84
3.3	Summary of DESI DR1 BAO measurements	85
3.4	w_0w_a CDM model implications of DESI DR1 data in combination with external datasets	86
3.5	Summary of DESI DR2 BAO measurements	88
3.6	w_0w_a CDM model implications of DESI DR2 data in combination with external datasets	90
3.7	Emerging tension identified in Λ CDM between DESI DR2 BAO and CMB	91
3.8	Hubble tension implications of DESI DR2 data	92
4.1	The spatial structure of the three-zone model: regions with average, lower and higher density	97
4.2	Demonstration of the effect of clumping on the recombination history and visibility function	100
4.3	Ionization fraction in different clumping models	102

4.4	Effects of clumping with a fixed cosmology	105
4.5	Demonstration of the effect of clumping on the CMB damping tail	107
4.6	Posteriors for H_0 and the clumping parameter of our model	109
4.7	Posterior of H_0 inferred from <i>Planck</i> data	110
4.8	Confidence regions for clumping versus H_0 from <i>Planck</i> 2018 $\ell < 1000$ data .	112
4.9	Relative difference in CMB temperature and polarization power spectra between our best fits and the Λ CDM best fit to <i>Planck</i>	114
4.10	Sound-horizon vs. damping angular scales for <i>Planck</i> 2018 + SH0ES	115
4.11	Position of the second peak in the CMB temperature power spectrum vs. damping angular scale for <i>Planck</i> $\ell < 1000$ + SH0ES	117
4.12	Confidence regions and one-dimensional posteriors for H_0 , matter fraction Ω_m and rescaled clustering amplitude S_8	119
4.13	Relative difference in the BAO distances between our best fit to <i>Planck</i> +SH0ES and the Λ CDM best fit to <i>Planck</i>	121
4.14	Transverse comoving distance versus Hubble parameter for the clumping model with <i>Planck</i> $\ell < 1000$ + SH0ES data	123
4.15	Correlation functions at $z = 0$ for Λ CDM (best fit to <i>Planck</i>) and the clumping model (best fit to <i>Planck</i> +SH0ES) are in almost perfect agreement after proper rescaling	124
4.16	Posterior for H_0 in our forecasts with a standard recombination fiducial, in all cases with CMB + SH0ES	127
4.17	Posteriors of the clumping parameter and H_0 for forecasts with a clumping fiducial	129
5.1	Large-scale cross-correlation functions of different tSZ SNR bins with the full LRG sample	139
5.2	Galaxy bias of different SNR bins relative to the full LRG sample	141
5.3	Small-scale line-of-sight cross-correlation functions of different SNR bins with the full LRG sample	142
5.4	Coordinate and velocity dispersions from small-scale line-of-sight clustering .	143
5.5	Average number of close neighbors in different tSZ SNR bins	145
5.6	Mock large-scale cross-correlation function of different SNR bins with the full LRG sample	150
5.7	Mock galaxy bias of different SNR bins relative to the full LRG sample	151
5.8	Correlations between halo mass and tSZ observables in our simulation-based model	153
B.1	Difference in CMB power spectrum multipoles computed with different recom- bination codes	189

B.2	Impact of massive neutrinos on CMB power spectrum multipoles	191
B.3	Full contour plot for <i>Planck</i> runs of the clumping model.	192
C.1	Distribution of numbers of close neighbors for LRG in different tSZ SNR bins	196

Acknowledgments

First, I thank Daniel for being the best advisor I could wish for throughout my time at Harvard. You have provided invaluable guidance and support while I was navigating grad school, Harvard, the Center for Astrophysics, DESI, and the general scientific community. You have also devoted a lot to making the Astronomy department a better place as chair and a faculty member, including the amazing introduction to observational astronomy during a yearly trip to Arizona telescopes. I have learned so much from you, and I still have a long way to go.

I deeply appreciate Cora Dvorkin co-advising my first cosmology project (on the clumpy recombination), teaching the cosmology course, and staying connected beyond the advisory committee. Julian Muñoz was a precious hands-on mentor during this project, helping me a lot to navigate my first cosmology research during the remote COVID year.

John Kovac and Doug Finkbeiner have also been on my advisory committee from the beginning, and I learned much from their different perspectives. John stressed the importance of a good explanation, the drawbacks of assuming too much prior knowledge, and the value of practicing talks to broader audiences. Doug often shared my joy of tackling a challenging problem for its own sake, and he also helped organize a wonderful day trip for me and other students on the Hawai'i island after the DESI meeting.

I thank Ravi Sheth for promptly accepting the invitation to be the external examiner and for contributing many interesting questions and suggestions at the private defense. I am looking forward to our future meetings.

Oliver Philcox also played a significant role in my PhD journey by putting together many cool codes, including RASCALC, and two iterations of EFTofLSS lectures.

I deeply value the shared knowledge, experience, interest and care for graduate students from the director of the Center for Astrophysics, Lisa Kewley, and the active faculty of the Astronomy department, particularly Charlie Conroy, Ramesh Narayan, Xingang Chen, Karin Öberg, Ashley Villar, Liam Connor and Lars Hernquist.

I have been honored and delighted to work in a very creative and supportive environment

of the Eisenstein research group. I thank my inspiring senior peers: Claire Lamman, Tanveer Karim, Boryana Hadzhiyska, Nina Maksimova, and Lehman Garrison¹; supportive junior peers: Hanyue Wang, Alex Johnson, Zihao Wu, Catherine Miller, Kyle Boone; and great postdocs around the group: Georgios Valogiannis and Carol Cuesta-Lazaro. I particularly enjoyed the many shared experiences with Claire, her passion for and great abilities in clear science communication have been a great motivation, and we shared many great fun moments around Harvard and at DESI meetings. I am looking forward to our continued collaboration and friendship at the Ohio State University.

I have had great opportunities and support through the Dark Energy Spectroscopic Instrument (DESI) collaboration. I am thankful to the DESI leadership, my mentors and peers, particularly Hee-Jong Seo, Ashley Ross, Arnaud de Mattia, Nikhil Padmanabhan, Florian Beutler, Héctor Gil-Marín, Pauline Zarrouk, Sesh Nadathur, Dragan Huterer, Will Percival; Otávio Alves, Uendert Andrade, Antón Baleato Lizancos, Xinyi Chen, Daniel Forero-Sánchez, Cristhian Garcia-Quintero, Sven Heydenreich, Jiamin Hou, Alex Krolewski, Juan Mena-Fernández, Jeongin Moon, Enrique Paillas, Alejandro Pérez-Fernández, Bernie Ried Guachalla, Christoph Saulder, David Valcin.

The great Astronomy graduate student community has been a source of immense support and pleasure. I particularly appreciate Hyerin Cho, Shelley Cheng, Alexia Simon, Chris Shallue, MJ Park; Lieke van Son, Gus Beane, Mike Foley, Charles Law, Miranda Eiben, Floor Broekgaarden, Andrew Saydjari, Jesse Han, Justina Yang, Nayantara Mudur, Zack Murray, Daniel Palumbo, Ioana Zelko; Kevin Ortiz Ceballos, Karthik Yadavalli, Maryam Hussaini, Vedant Chandra, Rebecca Woody, Olga Borodina, Ana Sofia Uzsoy, Mouza Almualla, Peixin Zhu, Erandi Chavez, Ralf Konietzka, Matthew Leung, Abby White, Tintin Nguyen and Qijia Zhou. I also value connections with Physics students who worked with Cora: Prish Chakraborty, Gemma Zhang, Anmol Raina, and Chandrika Chandrashekar.

I am deeply grateful to Peg Herlihy, our department administrator, for her constant support, care, invaluable knowledge and experience. I also thank our graduate coordinators:

¹We did not overlap at Harvard, but connected through Slack and the Michigan Cosmology School.

Robb Scholten (past) and Mark Palmer (current), and CfA administrators Lisa Catella and T.J. Martin. I am grateful for the international student support provided by the Harvard International Office.

The Derek Bok Center for Teaching and Learning, and particularly Sarah Emory, helped me a lot with English speaking and teaching culture. The Harvard Horizons symposia have also been a major inspiration.

I am deeply thankful to Vasily Beskin for organizing a range of unofficial astrophysics courses and astro coffee (all aimed at first-year undergraduates), giving me first research opportunities on radio pulsars in my second year, continuing collaboration while I was in Tel Aviv, and supporting my graduate school applications. I thank Sergey Pilipenko for the unofficial cosmology course introducing correlation function analyses during my first year at MIPT. I am grateful to Sasha Philippov for co-advising my first radio pulsar project, helping me settle professionally in Tel Aviv and then supporting my graduate school applications. I greatly appreciate my groupmates and research peers, Alisa Galishnikova and Egor Novoselov, who gave valuable information and advice about US graduate school applications. During a summer practice at the Special Astrophysical Observatory (in the North Caucasus), I gained a lot of inspiration and encouragement, for which I particularly thank Grigory Beskin and Oleg Verkhodanov. I owe much of my basis in classical and quantum field theory and general relativity to Emil Akhmedov.

My resolve to pursue cosmology research was solidified at the 15th School of Modern Astrophysics (2019). I thank the organizers and the lecturers, including Blake Sherwin, Guilherme Pimentel, Andrei Kravtsov, and Mikhail Medvedev.

I am grateful to Omer Bromberg for being an attentive advisor at Tel Aviv University and likewise supporting my graduate school applications. I value the support and company provided by Lev Pustil'nik and Anna Chashkina. My favorite advanced elective courses were paradoxes in quantum (paradoxes in physics more generally and quantum theory foundations) by Lev Vaidman and continuum theory (hydrodynamics) by Yacov Kantor.

I fondly remember the beginning of my scientific journey during the 5 years at Richelieu

Lyceum, a high school in Odesa specializing in math and natural sciences. Pavlo Viktor, our physics and astronomy teacher, fueled my love for the subjects, set a great communication example during his lessons, and greatly contributed to the extracurricular activities like the Ecological Expeditions. Vadim Manakin, my advanced physics teacher and olympiad coach, has always emphasized thinking outside of the box and good humor. Olga Korenovska, our homeroom and math teacher, as well as deputy principal for scientific activities, has always guided us with high professionalism and deep care. I am thankful to the former director of the Lyceum, Olga Palladiy, and the current director, Valerii Koleboshin, who also introduced me to the Young Physicists Tournament as early as the 7th grade. I thank our Young Physicists Tournament team mentors, Anastasia Maslechko, Tetyana Britavska, and Ivan Krychun, for their devotion and setting an example of scientific work and collaboration.

I thank Borys Kremiskiy and Ihor Anisimov for organizing the Ukrainian International Physics Olympiad (IPhO) team, and Stanislav Vilchynskiy for his fascinating lectures and problems on relativity and basic cosmology during our training and selection camps.

I owe my life to the people who organized and carried out my cancer treatment: Roman Goldman, Dr. Odelia Gur, and Dr. Freddy Aviv of Tel Aviv Sourasky Medical Center. I am deeply thankful to Dr. Natalie Buchkinsky, my primary care provider in Israel, for her sincere support and care ever since. I am grateful to our family friends who helped me and my parents settle in Israel during that frightful time: Jenny Batya Spektor and Yosef Spektor, Leonid and Irina Krasnopolsky, Sergey, Zhanna, Dmitry and Ora Levin. I appreciate the support from the Jewish community in Odesa in arranging our emergency repatriation.

I am grateful for the occasional high-stakes strategic battles in Heroes of Might and Magic III (Horn of the Abyss) with my school friends, Daniil (Danya) Selikhanovych and Alexandr (Sasha) Marunchenko; I shall not forget that “lots” is 40 (scorpicores). I also greatly enjoyed video chats with our MIT friends, Evgeny (Zhenya) Karkaryan and Ilya Simakov, which Sasha suggested during COVID. I am still eagerly waiting for an opportunity to meet all together in person. I also appreciated diverse discussions with my other school and college friend, Mykola (Kolya) Kishmar, and was happy to walk around Boston and Cambridge during

his visit with a friend.

I am thankful to my Ukrainian IPhO peer and friend, Oleksandr (Sasha) Shumaiev, for connecting me with the young Ukrainian community and his other friends at MIT and Harvard. The casual meetings with a variety of board games (including Bang, bringing back our IPhO prep days; Secret Hitler, and Codenames) and a spontaneous road trip to Plymouth (MA) and Providence (RI) have been delightful.

I thank my devoted parents, without whom none of this would be possible. They worked hard to first teach me directly and then emphasized the importance of high-quality education. They always placed me first and foremost, and undertook an emigration for my sake, leaving behind their long life in Odesa. Even after they stopped understanding most of my studies, they have always been ready to listen, share my joys and troubles, give valuable advice, help, support, and encouragement.

To my beloved parents and grandparents. My maternal grandfather loved physics and believed my mother had great potential as a physics researcher. After much persuasion, he accepted and supported her inclination and decision to pursue a degree and a career in the history of fine arts instead. We wish he could be with us longer to see his dream come true in the next generation.

Chapter 1

Introduction

1.1 Cosmological big questions

The glimpses of the Universe’s most global behavior continue to surprise and fascinate us. Early examples of such realizations include the expansion (Slipher, 1917; Hubble, 1929) and vastness (Hubble, 1926)¹ of the Universe. Then, the discovery of accelerated expansion (Perlmutter et al., 1999; Riess et al., 1998) was groundbreaking as well. Now there are indications that the accelerating force may be weakening (Brout et al., 2022; Rubin et al., 2023; DES Collaboration et al., 2024; DESI Collaboration et al., 2025e,a).

The standard model of cosmology, Λ CDM, emerged soon after the discovery of the accelerated expansion of the Universe. It had been remarkably successful in accommodating the different observations, but eventually some problems started to emerge. One of the most famous is the Hubble (H_0) tension — the more direct, local measurements of the current rate of expansion (most remarkably, from SH0ES collaboration, e.g. Riess et al. (2021)) give a value significantly larger than inferred from the cosmic microwave background (CMB, Planck Collaboration et al. (2020c)).

Furthermore, Λ CDM bears two phenomenological ingredients in its name. Constant dark energy (Λ) and cold dark matter (CDM) are the simplest possibilities for some fundamentally

¹Thanks to Leavitt’s law (Leavitt & Pickering, 1912) for Cepheid variable stars enabling extragalactic distance measurements.

unknown substances thought to comprise $\approx 95\%$ of total density. Cutting-edge measurements testing their properties can revolutionize our understanding of the Universe.

In this thesis, our major focus is on the expansion of the Universe: nature of its acceleration (mysterious dark energy) in Chapters 2 and 3 and discrepancy in the current speed inferred from different methods (Hubble tension) in Chapter 4. Further, in Chapter 5 we explore how to extract even more information from our data, which should help to enlighten dark matter as well.

1.2 Selected cosmological probes

1.2.1 Large-scale structure of the Universe: overview

Galaxies (and matter) in the Universe are not distributed randomly (e.g., Shapley, 1933). On a relatively smaller scale, galaxies clump together in clusters and superclusters. If we zoom out further, we can see that these groups follow a more global structure with filaments and walls – the cosmic web. This is the large-scale structure of the Universe.

In the Λ CDM model, the cosmic web is thought to form from tiny quantum fluctuations stretched by cosmic inflation, enhanced by gravitational collapse. On the very largest scales (starting from hundreds of megaparsecs), the structure is assumed to be homogeneous (largely the same from any position) and isotropic (largely the same in any directions)².

1.2.2 Galaxy surveys

Measurements of the large-scale structure of the Universe through galaxies are one of the pillars of modern cosmology. It provides access to the mass distribution³ of the Universe at different cosmic times.

The surveys systematically mapping galaxies on the sky come in two major varieties: photometric and spectroscopic. Photometric observations (photometry) record the flux (luminous

²There are claims of observations indicating inhomogeneity or anisotropy, although they have not been widely accepted.

³Not directly, because galaxies do not trace matter ideally.

energy per area) in several filters (wavelength or frequency bands), which can be wider or narrower. Spectroscopy provides finer details about the distribution of flux over wavelength (or frequency) of light; it can be loosely imagined as having a very high number of narrow, non-overlapping bands.

It generally takes more time or effort to obtain high-quality (high signal-to-noise ratio) spectroscopy than photometry of the same object, because the faint light is subdivided further. Accordingly, spectroscopic observations often use more easily obtainable photometry to select targets for the optimal use of precious telescope and instrument time.

Spectra are particularly valuable in cosmology because they allow to obtain reliable redshifts. Light propagating through the empty space is stretched (redshifted) together with the Universe. In a good spectrum, one can identify a group of spectral emission or absorption lines with known atomic lines. At least two lines are needed, because a single line can be almost anything with an unknown amount of stretching, whereas ratios of wavelengths (or frequencies) in pairs of lines are distinctive (if measured precisely enough) and do not change with an overall redshift. The redshift is then the relative difference between the emitted and the observed wavelength (known based on laboratory measurements). Redshifts add the third dimension to the more readily obtainable position on the sky (the 2-dimensional celestial sphere) and thus provide much more structural information.

There are methods of obtaining redshifts from photometry as well, but they are generally less reliable. Improving these methods is an active area of research (see e.g. Myles et al., 2021; Dey et al., 2021, 2022). For this reason, we do not explicitly use photometry in this thesis.

1.2.3 Baryon acoustic oscillations (BAO)

The early Universe was very hot and dense. High abundance of energetic photons and frequent collisions kept the ordinary matter (colloquially called “**baryons**” in cosmology⁴; mostly hydrogen) ionized. Photons continually interacted with the ordinary matter (mainly through

⁴Which is loose by particle physics standards, because electrons are strictly not baryons but leptons. However, most of the mass of the ordinary matter is from baryonic particles — protons and neutrons.

the free electrons), so that they all were in common equilibrium, comprising so-called photon-baryon plasma. Photons provided relativistically high pressure in this mixture, preventing the gravitational collapse of baryons and instead enabling sound waves (**acoustic oscillations**) propagating at roughly $1/\sqrt{3} \approx 0.58$ speed of light.

As the Universe expanded, it was diluting and becoming colder. Gradually, the photon-baryon interactions became much rarer and weaker, allowing neutral (hydrogen) atoms to form from ions and free electrons in a process called cosmic **recombination**⁵, which further decreased the collision frequency in a feedback loop. As a result, eventually the photons and baryons went their separate ways: photons propagated (almost) freely, and baryons collapsed to form galaxies in the cosmic web (not just by themselves but also into the potential wells already formed by dark matter, which was not deterred by photon pressure). Still, the acoustic oscillations are imprinted in both — mostly as it is more likely to find two higher (or two lower) densities at a distance that the sound waves traveled between the Big Bang and the recombination (the **sound horizon**, which can be computed theoretically, including numerical models). (This description largely follows Sunyaev & Zeldovich, 1970; Peebles & Yu, 1970, .)

The acoustic oscillation feature is weaker in baryons than in primordial photons. This is because the subsequent evolution of ordinary matter was more influenced by the gravitational interaction with the dark matter, which was largely unaffected by the photon pressure. But despite this and other complicating effects, it remains slightly more likely to find galaxies at the sound horizon distance from each other than at other similar distances, and this picture stretches with the expanding Universe (is largely preserved in cosmological comoving coordinates).

Thus, baryon acoustic oscillations provide a (comoving) standard ruler. Dividing the real size (sound horizon) by the angular size of the feature as observed on the sky gives the (comoving) angular diameter distance⁶ by definition. This distance is then related to the

⁵“Combination” might be a better name because the ions and electrons had never combined at an earlier point in cosmic history, to the best of our current knowledge. But the term was formed in other circumstances, in which “re-” was appropriate, and borrowed to cosmology.

⁶Comoving angular diameter distance is also called the transverse comoving distance.

integrated rate of the expansion of the Universe between now and then⁷. This measurement is called transverse BAO.

Redshift measurements also provide an additional measurement in the third dimension (line-of-sight BAO). The change in cosmological redshift is connected to the line-of-sight distance interval via the momentary rate of the expansion of the Universe. This way, measurement of the size of the BAO feature in redshift space and knowledge of the real size (sound horizon) gives the expansion rate at the representative time for the observed galaxy sample.

Without the absolute size of the sound horizon, BAO measurements with different galaxy samples can provide relative distance measurements to different redshift times and thus constrain the relative dynamics of the expansion of the Universe, but not its absolute rate.

1.2.4 Cosmic microwave background (CMB)

The cosmic microwave background is the oldest light that can be observed, often called an afterglow of the Big Bang. It is formed of photons propagating freely after last interacting with (scattering on) matter during cosmic recombination. They cooled down due to stretching with the expanding Universe to the typical millimeter (roughly microwave) wavelength. It is an on-sky image of predominantly a thin spherical shell of the Universe (surface of last scattering), from which photons originating at the times of recombination could reach us now by going straight all the time in between.

The CMB is famously an almost ideal example of blackbody (thermal) radiation with very nearly identical temperature ($T \approx 2.725$ K) in all directions. After the subtraction of the dipole, which can not be disentangled from the motions of the Earth and its surroundings, the differences are tiny at $\lesssim 100 \mu\text{K} = 10^{-4}$ K. We show the map of these fluctuations in Figure 1.1. CMB also has similarly small polarization variations across the sky. Such small perturbations can be modeled and interpreted very precisely and robustly, and they provide a great view of the early Universe.

The maps of CMB fluctuations are summarized by angular power spectra. The angular

⁷And the spatial curvature of the Universe if we do not assume its absence on the largest scales.

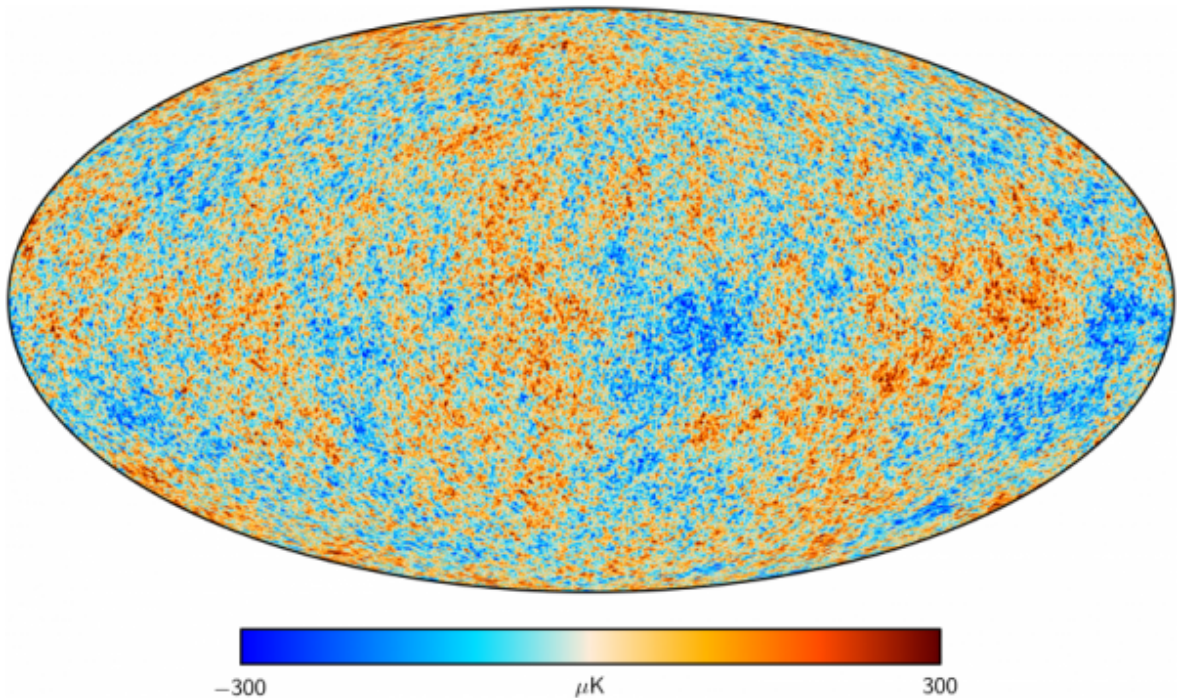


Figure 1.1: Map of the cosmic microwave background (CMB) temperature fluctuations measured by the *Planck* satellite. (Image credit: ESA and the Planck Collaboration.)

power spectrum is a statistical tool describing the amount of variations on different angular scales. We show the CMB temperature power spectrum in Figure 1.2.

Several features are apparent in Figure 1.2. One can see the damped oscillations with the first peak at an angular scale $\sim 1^\circ$. The period of these oscillations, measured very precisely, is set by the acoustic feature we described in the previous section, analogous to the transverse BAO in galaxies. The damping is due to photon diffusion before and during the cosmic recombination (the process known as Silk damping, Silk, 1968). The additional height variations in the oscillatory peaks are regulated by the amount of ordinary matter (baryons).

The formation of the CMB was intertwined with the evolution of the large-scale structure. E.g., through gravitational interaction with dark matter, including gravitational redshift from maxima and minima of the gravitational potential. CMB also has secondary anisotropies, formed by photons' interaction with matter after recombination, on their way towards us. These include gravitational lensing (deflection by masses of dark matter and baryons) and

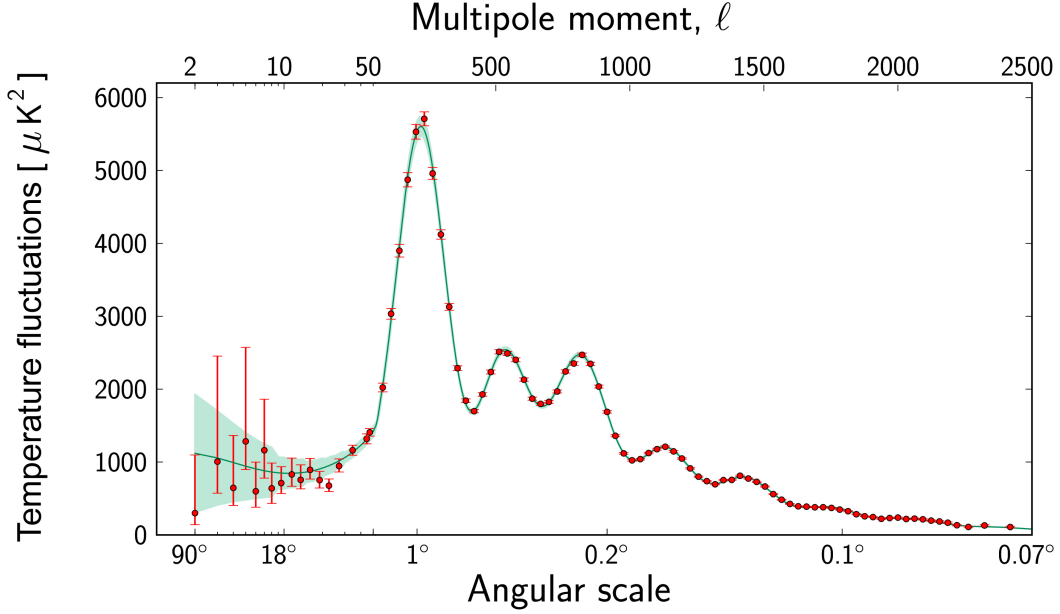


Figure 1.2: Cosmic microwave background (CMB) temperature power spectrum measured by the *Planck* satellite (red dots with errorbars) and predicted by the Λ CDM model (green). (Image credit: ESA and the Planck Collaboration.)

scattering on re-ionized baryon gas (Sunyaev-Zeldovich effect, which we will use in Chapter 5).

1.2.5 Standard candles: Type Ia supernovae

The standard candle is an object with a fixed luminous power (total or in a certain wavelength range). By knowing this power and measuring the power per area we receive from the object, we can compute the distance using the inverse-square law. More precisely, in cosmology, this gives a luminosity distance, which is related to the integrated rate of the expansion of the Universe between now and the emission of the light, like the other distances.

Supernovae of a distinct type (Ia) are nearly standard candles. They have been identified with a certain subclass of explosions of white dwarfs exceeding the Chandrasekhar limit by accretion. Type Ia supernovae provided one of the first definite pieces of evidence in favor of accelerated expansion of the Universe (Perlmutter et al., 1999; Riess et al., 1998).

For precision cosmology with Type Ia supernovae, it is necessary to go beyond the assumption of constant intrinsic power by establishing connections with other observable quantities starting from the rate of luminosity decrease (Phillips relation). Technically, this makes them standardizable candles. The light can also be partially obscured, e.g. by cosmic dust. Many of the necessary corrective relations are not understood from first principles and need to be calibrated using other distance measurements. An additional complication is that Type Ia supernovae are relatively rare and have hardly been observed within the reach of more direct distance measurements (parallax), thus, the calibration has to rely on a different type of standardizable candles, e.g., Cepheid variable stars. This thesis does not include original work on details of these measurements in particular, so we refer a curious reader to Riess et al. (2022); Freedman et al. (2024); Brout et al. (2022); Rubin et al. (2023); DES Collaboration et al. (2024).

1.3 Dark Energy Spectroscopic Instrument (DESI)

The Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration et al., 2016a) is conducting a major spectroscopic galaxy survey since 2021. The instrument (DESI Collaboration et al., 2016b, 2022) takes approximately 5000 spectra simultaneously. To deliver light to the spectrographs, it uses optical fibers individually positioned by miniature robots in the focal plane of the 4-meter Mayall telescope at Kitt Peak, Arizona. This allows to change the configuration quickly and automatically to targets selected in advance. The target selection relies on photometric Legacy Imaging Surveys (Dey et al., 2019).

During the main 5-year program, DESI is scheduled to cover 14,000 square degrees on sky and the redshift range $0 < z < 3.5$. As a result, the DESI survey will surpass the previous leading surveys (mainly SDSS/BOSS/eBOSS) by an order of magnitude in both volume and number of extragalactic objects, collecting more than 30 million galaxy and quasar spectra. One of the main scientific drivers for its design was the precision of the BAO measurements.

1.4 Statistics of the Universe on large scales

In this chapter, we provide a slightly deeper review of the statistical methodology used to describe the large-scale structure.

1.4.1 Overdensity field

By field, we mean assignment of a value (e.g., a number or a vector) to each point in space (and time). In the standard cosmological paradigm, we assume that the initial conditions are set randomly (stochastically) by the quantum fluctuations during cosmic inflation. Accordingly, we treat our observables as random fields.

Having a mass density field, we define a mass overdensity at any point:

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x})}{\bar{\rho}} - 1 \quad (1.1)$$

The average density depends on cosmic time. However, measuring the mass density in 3 dimensions can be incredibly hard.

In galaxy surveys, we instead observe galaxies mainly as points with a certain position. We can talk about their number density — number per volume. Ideal points can be represented with Dirac delta functions $\delta^D(\mathbf{x} - \mathbf{x}_p)$, which, integrated over a volume, indicate 0/1 whether a point is inside or not. Alternatively, one might imagine a fine grid of cells, each of which has a volume and a number of galaxies (possibly 0) and thus an average density.

From the number density, we define the number overdensity

$$\delta(\mathbf{x}) \equiv \frac{n(\mathbf{x})}{\bar{n}} - 1 \quad (1.2)$$

Points can be weighted; that should be taken into account in the average density. Also, the average density for a real galaxy survey should reflect the selection effects.

Because either density is non-negative, for either overdensity $\delta(\mathbf{x}) \geq -1$. By design, $\langle \delta(\mathbf{x}) \rangle = 0$, averaging over the possible realizations of the Universe from the random quantum fluctuations.

We can also describe the overdensity field as a sum of plane waves (Fourier modes) $\delta(\mathbf{k})$

with wavevector $\mathbf{k}$ using a Fourier transform:

$$\tilde{\delta}(\mathbf{k}) \equiv \int d^3\mathbf{x} \delta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} \quad (1.3)$$

$$\delta(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \tilde{\delta}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}. \quad (1.4)$$

Space of wavevectors $\mathbf{k}$ is commonly referred to as Fourier space, and the space of coordinates $\mathbf{x}$ — as configuration space; the two descriptions are equivalent.

The Fourier image is convenient because linear equations give an independent time evolution of different Fourier modes (plane waves). And linear equations are a great approximation for small perturbations.

1.4.2 Gaussian random field

A Gaussian random field is a random field in which the joint distribution of values at any number of points is a multivariate normal (Gaussian) distribution. It is a convenient and common abstraction or approximation in cosmology.

First and foremost, we mean the distribution of possible realizations of the Universe from primordial quantum fluctuations. But with homogeneity and isotropy, the spatial distribution on large scales should approach it (akin to the ergodic hypothesis).

Assuming translation invariance (homogeneity), a Gaussian random field can be obtained as a sum of independent Fourier modes. Fourier modes of a Gaussian random field also represent a Gaussian random field in Fourier space.

The early Universe is very close to Gaussian, as predicted from the inflationary foundation of Λ CDM and seen in the CMB maps (e.g., Planck Collaboration et al., 2020a). As we remarked before, linear evolution preserves the mode independence and thus Gaussianity. Later evolution by gravity and other interactions on small scales causes deviations (non-Gaussianity). Because different modes start to affect each other's evolution when perturbations grow beyond linear. Another, simpler reason is that a normal distribution with zero mean and large deviations (dispersion, variance) strongly violates $\delta(\mathbf{x}) \geq -1$. Primordial non-Gaussianity is also possible (e.g., predicted by some inflation models), and active search is ongoing using CMB (e.g., Planck

Collaboration et al., 2020e) and LSS data (e.g., Rezaie et al., 2024; Chaussidon et al., 2024).

A normal (Gaussian) distribution is completely determined by its mean and covariance. As we remarked in the previous section, cosmic overdensity fields have zero mean by construction. This leaves the covariance, the pairwise correlations. With (large-scale) homogeneity, they can depend only on the vector between the points. With isotropy, only distance matters. But the line-of-sight direction is different from the others (e.g., in redshift-based distance determination), so the angle with it matters too.

A Gaussian random field shares a convenient property with the normal distribution. Isserlis' theorem⁸ (Isserlis, 1918) states that with zero mean, all higher moments (correlations of higher-point) are determined by the covariance (2-point correlations). In particular, the odd-order (odd-point) quantities are zero.

1.4.3 The 2-point correlation function (2PCF)

The 2-point correlation function (2PCF) of discrete particles (e.g., galaxies or dark matter halos) describes the excess probability of finding two such particles separated by a certain distance, compared to the random scatter case.

$$\xi(\mathbf{r}) = \frac{dP_g(\mathbf{x}_g + \mathbf{r}|\mathbf{x}_g)}{\bar{n} \cdot d^3\mathbf{r}} - 1. \quad (1.5)$$

$dP_g(\mathbf{x}_g + \mathbf{r}|\mathbf{x}_g)$ is the conditional probability of finding a (different) galaxy in an infinitesimal volume $d^3\mathbf{r}$ around $\mathbf{x}_g + \mathbf{r}$ given a galaxy at $\mathbf{x}_g$. $\bar{n}$ is again the expected number density if the Universe were unclustered; it should reflect the survey selection effects, weighting scheme, etc.

Perhaps a more intuitive statement is that the probability of finding two particles (e.g., galaxies or halos) in infinitesimal volumes $d^3\mathbf{x}_{1,2}$ near $\mathbf{x}_{1,2}$ respectively is

$$dP_{12} = \bar{n}d^3\mathbf{x}_1 \cdot \bar{n}d^3\mathbf{x}_2 \cdot [1 + \xi(\mathbf{x}_2 - \mathbf{x}_1)], \quad (1.6)$$

where we can set $\mathbf{r} \equiv \mathbf{x}_2 - \mathbf{x}_1$.

⁸Isserlis' theorem is also known in particle physics and cosmology as Wick's (probability) theorem after Wick (1950).

A more general definition of the 2PCF is through the spatial correlations of overdensities (Equation (1.1) or Equation (1.2)):

$$\xi(\mathbf{r}) = \langle \delta(\mathbf{x})\delta(\mathbf{x} + \mathbf{r}) \rangle = \frac{1}{V} \int d^3\mathbf{x} \delta(\mathbf{x})\delta(\mathbf{x} + \mathbf{r}). \quad (1.7)$$

For the number overdensity of discrete particles (Equation (1.2)) this definition is theoretically equivalent to the probabilistic one (Equation (1.5)).

BAO feature manifests as a peak of the correlation function (scaled by r^2) at $r \approx 150$ Mpc. This feature is relatively subtle and requires large galaxy samples to detect. The first discovery this way was in Eisenstein et al. (2005).

2PCF can be estimated by counting pairs between observed galaxies (data) D and random points R distributed according to the background density $\bar{n}$. Therefore, randoms reflect the survey boundaries, subtler selection effects and weighting. It is often more efficient to generate random points than to write some functional form of $\bar{n}(\mathbf{x})$.

The probabilistic definition (Equation (1.5)) motivates the Davis-Peebles estimator (Davis & Peebles, 1983):

$$\hat{\xi}_{\text{DP}}(\mathbf{r}_{\text{bin}}) = \frac{DD(\mathbf{r}_{\text{bin}})}{DR(\mathbf{r}_{\text{bin}})} - 1, \quad (1.8)$$

where $DD(\mathbf{r}_{\text{bin}})$ is the (weighted) count of pairs of data points with separation $\mathbf{r}$ between them belonging in a certain region (bin) around $\mathbf{r}_{\text{bin}}$, and $DR(\mathbf{r}_{\text{bin}})$ is the same but for data-random pairs.

Expansion of the correlation function through overdensities (Equations (1.2) and (1.7)) motivates the Landy-Szalay estimator (Landy & Szalay, 1993):

$$\hat{\xi}_{\text{LS}}(\mathbf{r}_{\text{bin}}) = \frac{DD(\mathbf{r}_{\text{bin}}) - 2DR(\mathbf{r}_{\text{bin}}) + RR(\mathbf{r}_{\text{bin}})}{RR(\mathbf{r}_{\text{bin}})}, \quad (1.9)$$

where $RR(\mathbf{r}_{\text{bin}})$ is the same as DD and $DR(\mathbf{r}_{\text{bin}})$ but counting pairs of random points. The Landy-Szalay estimator has minimal variance and is more stable to small fluctuations in the random points (Kerscher et al., 2000), so we rely on it exclusively in Chapters 2 and 3. At the same time, the Davis-Peebles estimator is less demanding in multiple ways, one of which is important in Section 5.3.1.

1.4.4 The power spectrum

The power spectrum is defined through the correlation of the Fourier modes ((1.3)):

$$\left\langle \tilde{\delta}(\mathbf{k}) \tilde{\delta}^*(\mathbf{k}') \right\rangle = (2\pi) P(\mathbf{k}) \delta^D(\mathbf{k} - \mathbf{k}'), \quad (1.10)$$

where δ^D is the Dirac delta function, theoretically infinite at zero argument and zero elsewhere. In practice, power spectrum estimation uses a finite volume with a discrete Fourier transform, which transforms the Dirac delta function (infinite at zero) with a volume multiplier if and only if $\mathbf{k} = \mathbf{k}'$ (technically, Kronecker delta).

The BAO feature manifests as a series of peaks in the power spectrum. This was first measured in data by Cole et al. (2005).

The 2-point correlation function and the power spectrum are connected via the Fourier transform:

$$P(\mathbf{k}) = \int d^3\mathbf{r} \xi(\mathbf{r}) e^{-i\mathbf{k}\cdot\mathbf{r}} \quad (1.11)$$

$$\xi(\mathbf{r}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} P(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{r}} \quad (1.12)$$

Thus, 2PCF and power spectrum contain identical information if measured over infinite ranges. However, a cut in Fourier space does not correspond to a cut in configuration space and vice versa. Additionally, survey geometry and selection effects are more complicated for the power spectrum (Hand et al., 2017).

1.4.5 Higher-point statistics

Higher-point statistics can be obtained by generalizing Equations (1.7) and (1.10). E.g., the 3-point correlation function:

$$\zeta(\mathbf{r}_1, \mathbf{r}_2) = \langle \delta(\mathbf{x}) \delta(\mathbf{x} + \mathbf{r}_1) \delta(\mathbf{x} + \mathbf{r}_2) \rangle \quad (1.13)$$

and the bispectrum (the Fourier-space 3-point function):

$$\left\langle \tilde{\delta}(\mathbf{k}_1) \tilde{\delta}(\mathbf{k}_2) \tilde{\delta}^*(\mathbf{k}_3) \right\rangle = (2\pi) B(\mathbf{k}_1, \mathbf{k}_1) \delta^D(\mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3) \quad (1.14)$$

As we mentioned before, for a Gaussian random field, higher-point statistics are set by the 2-point statistics (Isserlis, 1918). For a non-Gaussian field, they contain additional information, approaching completeness as order increases to infinity. However, the information content does not increase very fast, whereas the number of dimensions grows quickly (even after one uses additional symmetries to reduce dimensionality). This motivates the search for alternatives.

1.5 Large-scale structure data analysis and covariance matrices

For meaningful (statistical) interpretation of data (Bayesian or frequentist), it is crucial to compute the likelihood — describing the probability of obtaining the data given a model. Fortunately, the distribution of measured clustering statistics is well described by a multivariate Gaussian (thanks to the central limit theorem and typically averaging a large number of modes), which is fully described by the mean and the covariance matrix. The mean can be set to the theoretical prediction, which leaves the covariance.

Perhaps the most intuitive approach to the covariance is scatter in repeated independent measurements. However, repeating a cutting-edge survey even once more is hardly practical. Moreover, the fact that we have only one Universe sets a limit to how many big surveys (covering the same redshifts) can be done before they have to overlap.

The next option is simulating the Universe in computer models (ultimately, making mock galaxy catalogs). This solution is far from ideal. Detailed simulations are computationally expensive, more so as they need to cover larger volumes for bigger surveys. A precise covariance estimate requires a large number of samples, increasingly higher as more quantities are measured. This forces a hard compromise between quality and quantity, keeping the total computation time very long. Such catalogs need to capture key aspects of clustering and be representative of the data or the assumed theoretical model. The mock-based covariance production is therefore not characterized by flexibility, as the generation and processing of numerous simulations for an updated dataset or alternative model requires a huge effort.

Another method involves making multiple samples by selecting different subsets of the data. The lack of dependence on mocks and model assumptions makes them attractive. These are

represented by resampling techniques like jackknife and bootstrap, which involve splitting the data into parts. However, it is often hard to divide an intricately shaped survey volume into equivalent regions. These regions would not be independent of each other because large-scale correlations exist. Subsets of the survey may have more complicated boundaries, affecting the covariances in ways not described by simple scalings.

From a theoretical point of view, covariance of 2-point functions ((1.7)) involves expectation values at 4 points. In a Gaussian random field, by Isserlis' theorem (Isserlis, 1918), they can be reduced to 2-point functions between different pairs of points. The 2-point function could be measured from the data or modeled theoretically and substituted into its covariance computation. However, non-Gaussianities develop over cosmic time. Galaxies correspond to high matter overdensities, evolving in a crucially nonlinear regime. This makes the derivation of important higher-point corrections to the ideal Gaussian picture also very challenging. Measuring the necessary 3- and 4-point functions from the data reliably is hard as well because of the large number of bins needed.

Each of these approaches has its advantages and disadvantages. It would be nice to develop an optimal combination. This is what we approach in a major part of this thesis, Chapter 2, by combining theory with data resampling (or simulated Universes).

1.6 Outline of the thesis

Chapter 2 is dedicated to the important technical work on the fast semi-theoretical, semi-empirical covariance matrices for 2-point correlation functions. Chapter 3 then shows a selection of global DESI BAO results, which relied on these covariance matrices (along with multiple other supporting studies). We focus on the nature of dark energy and the Hubble tension. Chapter 4 follows with our earlier work on a potential relief of the Hubble tension through inhomogeneous recombination, which does not require fundamentally new physics. Chapter 5 presents our work in progress on extracting more information than standard 2-point functions allow without higher-point functions, by using the DESI spectroscopic galaxies together with secondary CMB anisotropies maps. In Chapter 6 we reflect on the achieved

results and discuss future directions.

Chapter 2

Fast semi-analytical covariance matrices for two-point correlation functions for DESI data¹

2.1 Introduction

It is a particularly exciting time for observational cosmology due to the transition from Stage III to Stage IV dark energy experiments. The Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2016a, 2022) belongs to this newer generation and is actively operating. Last year saw the validation of its scientific program (DESI Collaboration et al., 2024d) and the early data release (DESI Collaboration et al., 2024e). In particular, the sample of Luminous Red Galaxies (LRG, Zhou et al. (2023)) observed during the first two months of DESI main survey operations (DESI-M2) yields a BAO scale measurement with 1.7% precision (Moon et al., 2023), which is already comparable to the aggregate precision of 0.77% of preceding leading surveys, BOSS and eBOSS (Alam et al., 2021). A larger 1-year dataset (DR1) (DESI Collaboration et al., 2025c) has been released, along with two-point clustering (DESI Collaboration et al., 2024a), inverse distance ladder measurements (and thus

¹This chapter is a compilation of Rashkovetskyi et al. (2023) and Rashkovetskyi et al. (2025b).

the expansion history of the Universe) using the baryon acoustic oscillations (BAO) of galaxies, quasars (DESI Collaboration et al., 2025f), and Lyman- α (DESI Collaboration et al., 2025d), full-shape analysis of the 2-point statistics for galaxies and quasars constraining the growth of cosmic structure (DESI Collaboration et al., 2024b), and implications for cosmological models (DESI Collaboration et al., 2025e, 2024c; Chaussidon et al., 2024). These DESI results have unprecedented precision for their kind of measurement, providing unique new opportunities to test our understanding of the Universe.

Data interpretation with a physical model requires a covariance matrix model, which can be hard to obtain. An intuitive way to do so is from the scatter in repeated independent identical measurements. However, only one Universe is accessible for us to probe in cosmology. Moreover, it is not feasible to replicate a state-of-the-art experiment exactly. Therefore, covariance matrix estimation in cosmology requires elaborate techniques.

The standard method for computing covariance matrices in large-scale structure studies has been based on scatter within large sets of simulated (mock) catalogs (e.g. Chuang et al. (2015); Zhao et al. (2021, 2025); Ereza et al. (2023); Variu et al. (2024)). They need to be both highly accurate representations of the data (in particular, large enough to cover the survey volume), and numerous enough to give a good estimate of the covariance matrix of the vector of observables. The relative precision is primarily determined by the number of samples and the dimension of the matrix. If the number of samples is smaller than the number of bins plus one, the resulting covariance matrix estimate is not invertible. Moreover, sample noise biases the estimate of the inverse covariance matrix (Hartlap et al., 2007) and causes the widening of model parameter errorbars (Percival et al., 2022). Because highly detailed simulations require a lot of time even for volumes much smaller than a survey like DESI covers, it is unavoidable to rely on approximations, limiting the realism of the simulated catalogs. Even so, generating and calibrating an adequate mock suite is very hard and expensive.

The mock-based covariance matrix estimation is becoming increasingly challenging with time. First, as the surveys improve, each mock catalog needs to include more galaxies and/or more volume, thereby taking longer to generate and process. Second, with more data, we

aim to include a longer vector of observables in the analysis, requiring a larger covariance matrix for it, which in turn demands a higher number of mocks for adequate precision (Hartlap et al., 2007; Taylor et al., 2013; Percival et al., 2014). Third, the substantial time to generate mocks typically means that one cannot produce enough simulations for many separate sets of cosmological, galaxy-halo connection models and selections of tracer galaxy samples. This creates a potential systematic error when extrapolating the covariances derived in one scenario to another. Fourth, as a specific example of this, the long timeframe to generate mocks can even create a situation where schedule concerns force the mocks to be calibrated on early inputs that do not match the final version of the analysis. The problem is especially severe when blinding is employed, and the simulation teams should not see the true and complete data clustering before the analysis methodology is frozen. This creates a need and opportunity for faster alternative methods.

A promising alternative is a theoretical derivation of the covariance matrix, which is more conveniently performed in Fourier space for cosmological clustering. Analytical covariance matrices for galaxy power spectrum have been developed using perturbation theory (Wadekar & Scoccimarro, 2020; Kobayashi, 2023; Alves et al., 2025). However, perturbative expansions only work for a limited range of wavenumbers. Unfortunately, the translation of these results to the configuration space introduces an additional precision loss due to the nonlocality of the Fourier transform. The accuracy of the transformation can be improved with the generalization of the FFTLog algorithm to covariance matrices (Fang et al., 2020), but this approach has not yet been applied to and validated for redshift-space clustering of galaxies in 3 dimensions. Therefore, the analytical methods are not yet directly applicable to the correlation functions.

Another approach is resampling the data and considering the scatter between different parts for the covariance estimation. The jackknife technique is based on this idea. Unfortunately, dividing the spectroscopic survey volume into equivalent parts is challenging because of boundary effects and multiple other factors varying across the survey. This problem becomes more severe with a larger number and accordingly smaller size of regions. More samples are still desirable for higher covariance matrix precision, analogously to the mock case. In

addition, regions are not completely independent from each other. Mohammad & Percival (2022) attempted to correct for the last factor by dividing the jackknife pair counts into different categories and re-weighting some of them. However, Trusov et al. (2024) showed that some of the former assumptions are violated in higher-density setups, leading to significant biases. To solve this problem and obtain a more precise, better-conditioned matrix, they proposed a hybrid approach, combining jackknife with mocks (requiring fewer realizations than for the sample covariance). This illustrates the promise of hybrid approaches incorporating different techniques.

Here, we focus on a combination of analytical and jackknife elements in configuration space. The theoretical component hinges on the relation between the covariance matrix of the 2-point correlation function (2PCF) and the 4-, 3- and 2-point functions. This method has been developed in a series of papers (O’Connell et al., 2016; O’Connell & Eisenstein, 2019; Philcox et al., 2020; Philcox & Eisenstein, 2019) and implemented as the RASCALC code². This approach enables the computation of covariance matrices for correlation functions using only the data, without extra effort, assumptions and approximations for generating mocks.

Multipole (Legendre) moments of the correlation function are preferable for redshift-space analysis and theoretical modeling (e.g. Chuang & Wang (2013)). Philcox et al. (2020) only developed an estimator for the covariance matrix of the 2PCF in angular bins. With a large number of angular bins required to estimate multipoles adequately, the computation time of the covariance matrix becomes infeasibly long. Philcox & Eisenstein (2019) introduced a direct covariance model for Legendre moments, but its final calibration depends on a separate computation for angular bins, which is inconvenient in practice. We extend the methods of Philcox et al. (2020); Philcox & Eisenstein (2019), developing a more practical and exact estimator for the 2PCF multipole covariance compatible with DESI’s correlation function measurement code, PYCORR³ (De Mattia et al., 2024).

We also note the prospects of standard reconstruction techniques that aim to reverse the

²<https://github.com/oliverphilcox/RascalC>

³<https://github.com/cosmodesi/pycorr>

large-scale displacements during the times after the drag epoch. Such subsequent evolution leads to broadening and contamination of the BAO peak, and one can partially undo it to sharpen the feature (Eisenstein et al., 2007). The RASCALC formalism is applicable to reconstructed 2PCF covariance as well, with minor adjustments.

The method has been closely integrated with DESI. We have contributed RASCALC covariances for the BAO analysis of DESI-M2 data (Moon et al., 2023). A part of this chapter accompanies it, focusing on the validation of the approach in realistic circumstances. We limit ourselves to analogs of the DESI LRG sample (Zhou et al., 2023) due to the availability of a large suite of mocks with corresponding cuts, providing a good sample covariance matrix for reference. Similarly to Philcox et al. (2020), we process a single mock catalog in essentially the same manner as data and compare the resulting covariance with the sample covariance of clustering measurements in all available mocks, which gives a fair proxy of the pipeline performance on data and is also robust to the mismatch between data and mock clustering. We repeat the procedure multiple times, taking a different catalog each time to assess the accuracy of the method, its stability, and fluctuations. In addition, we pay extra attention to the formation of a covariance matrix comparison toolkit. We focus on the meaning of the numbers used and derive reference values for the ideal case when the semi-analytical prediction matches the true underlying covariance. Due to sample variance, these expectation values can be nontrivial, and understanding the noise in the comparison measures is crucially important as well. We choose a smaller number of observables for lower noise and clearer interpretation, and further project the covariances into the lower-dimensional and more physically meaningful space of model parameters.

Following that, it was embraced as part of a coordinated covariance matrix effort for DESI DR1 two-point clustering measurements (DESI Collaboration et al., 2024a), along with analytical covariance matrices for power spectra (Alves et al., 2025) and the general comparison focusing on the consistency of the model fits (Forero-Sánchez et al., 2024). We have benefited enormously from synergies with other supporting studies for the galaxies and quasars BAO (DESI Collaboration et al., 2025f): optimal reconstruction (Chen et al., 2024a; Paillas

et al., 2025), combined tracers (Valcin et al., 2025), halo occupation distribution systematics (Mena-Fernández et al., 2025; Garcia-Quintero et al., 2025), fiducial cosmology systematics (Pérez-Fernández et al., 2025) and theoretical systematics (Chen et al., 2024b).

Fiber assignment incompleteness effects are a big challenge with DESI. For a spectrograph with 5000 robotic fibers, the choice of a target for each is highly complex. Some fibers cannot reach all targets in the field of view, which leads to incomplete coverage, dependent on the priority of different targets and the density of the targets on the sky. To improve the completeness, DESI is scheduled to make 7 dark-time and 4 bright-time passes over each area during its full 5-year program (Schlafly et al., 2023). However, the coverage achieved during the first year of observations (DR1) varies across the footprint (Ross et al., 2025) between complete and only a single pass in many areas. In multi-pass regions, the assignment also depends on previous DESI exposures in the area. These effects imprint on 2-point clustering (Bianchi et al., 2024) and must also affect the higher-point correlations entering the covariance. The exact algorithm can be applied to mocks, but it is not fast enough to obtain as many catalogs (~ 1000) as needed for the estimation or high-quality validation of the full covariance matrix (Lasker et al., 2025). A faster approximation has been developed (Hanif et al., 2025), enabling a suite of 1000 mocks with fiber assignment modeled. We aim to validate the semi-analytical covariance matrices, taking fiber assignment incompleteness into account for the first time.

We organize this chapter in the following way. Section 2.2 is dedicated to the semi-analytical covariance matrix estimation: Section 2.2.1 summarizes the previously developed methodology, Section 2.2.2 discusses a modification of random counts computation and Section 2.2.3 introduces a formal extension to reconstructed data. In Section 2.2.4 we derive the new estimators for covariance matrices of Legendre multipoles of the correlation function, used in Section 2.5. In Section 2.3 we discuss the problem of covariance matrix comparison, presenting our selection of generic compact measures in Section 2.3.1, explaining their application to the internal convergence checking in Section 2.3.2 and the Fisher projection to a space of physical model parameters in Section 2.3.3.

In Section 2.4 we apply the previously discussed methods to RASCALC validation with

DESI-M2 LRG mocks. We briefly describe the mock catalogs and standard BAO reconstruction methods in Section 2.4.1, explain the covariance matrix setup in Section 2.4.2, check the internal convergence in Section 2.4.3 and finally compare the semi-analytic covariances with the mock sample covariances in the measurement space (the correlation function bins) in Section 2.4.5, in the BAO model parameter space in Section 2.4.6 and particularly for the BAO scale in Section 2.4.7.

In Section 2.5 we validate the updated semi-analytical covariance matrix methodology for DESI DR1 using mocks. Section 2.5.1 provides details on the DESI DR1 mock catalogs, fiber assignment incompleteness modeling, standard BAO reconstruction and analysis techniques informing our comparison of covariance matrices. Section 2.5.2 explains the setup for semi-analytic covariance matrix estimation. In Section 2.5.3 we check the consistency of the method’s application to the simulations: intrinsic numerical stability and the values of the key parameter of the covariance matrix model. In Section 2.5.4 we compare our semi-analytical covariance matrix estimates with the sample covariance of the mocks in terms of correlation function multipoles and the cosmological parameters, following both BAO and full-shape analyses.

Section 2.6 concludes the main text with a summary and an outlook. Section A.1 gives the covariance matrix estimators generalized for multi-tracer analysis. Section A.2 provides an overview and derivations of useful properties of covariance matrix comparison metrics. Section A.3 explains the procedure to obtain the covariance for the combination of two disjoint regions (volumes).

2.2 Methods of covariance matrix estimation

2.2.1 Overview of previous work

We start with a summary of the methodology developed in O’Connell et al. (2016); O’Connell & Eisenstein (2019); Philcox et al. (2020). The RASCALC code builds single-parameter⁴

⁴For single tracer; for multiple tracers it is one parameter per tracer.

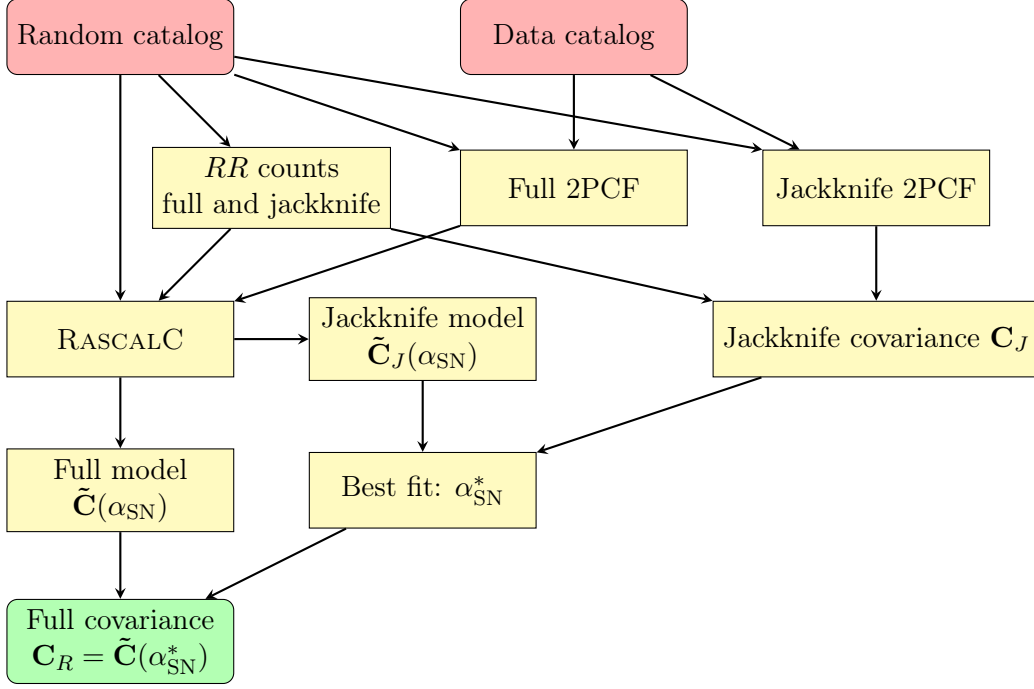


Figure 2.1: Flowchart of RASCALC jackknife pipeline (fiducial; developed in O’Connell & Eisenstein (2019); Philcox et al. (2020)). This process is used for DESI data and most of the mock tests in this paper. In the latter case, a single mock catalog and its corresponding random catalog(s) are provided as data and randoms.

covariance matrix models⁵ based on a random catalog and a table of 2-point correlation function values. Then we fit a model to a reference covariance (defined later) to obtain the optimal parameter value for the final prediction. In the fiducial data pipeline, shown schematically in Figure 2.1, we measure the correlation function directly from the data, the code produces separate models for full and jackknife covariance matrices, we fit the latter to the data jackknife covariance matrix, and plug the resulting optimal parameter into the full model. Figure 2.2 shows an alternative pipeline where we use the best fit of the full covariance model to the mock sample covariance instead. The jackknife methodology has a significant advantage: it only requires mocks for the initial validation, and then allows to generate covariance matrices for different setups without additional simulations. We now provide more details on both pipelines, focusing on the single-tracer case; the generalization

⁵This step is the most computationally heavy and is implemented in C++.

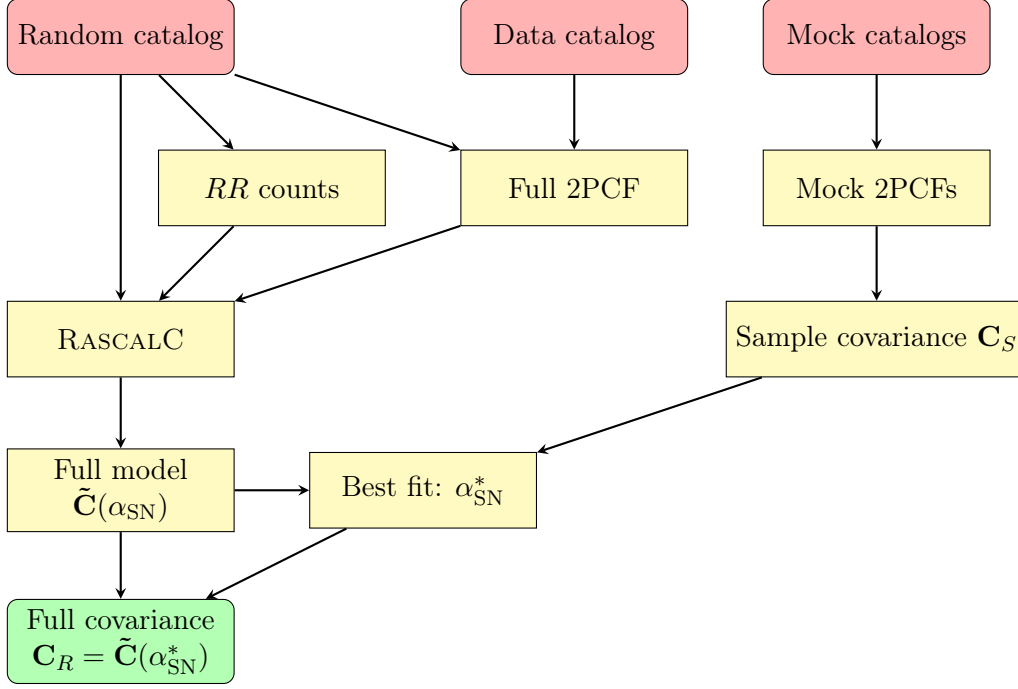


Figure 2.2: Flowchart of RASCALC mocks pipeline (alternative; principal idea from O’Connell et al. (2016)). The full covariance model can be reused from a jackknife computation (Figure 2.1), provided that the randoms and full 2PCF were the same.

to multiple tracers can be found in Section A.1.

To explain the covariance matrix model, we need to start with the definition of the 2-point correlation function (2PCF). The standard Landy-Szalay estimator (Landy & Szalay, 1993) in radial bin a and angular bin c ⁶ is

$$\hat{\xi}_a^c = \frac{(NN)_a^c}{(RR)_a^c} \quad (2.1)$$

where $N = D - R$, R are random points and D are data points (galaxies).

The random counts RR are determined by the survey geometry. We assume that the survey design choices are independent of the random realization of the Universe. Consequently, we treat the random counts as fixed.

The numerator, however, is determined by structure formation processes, which are

⁶RASCALC assumes uniform binning in $|\mu|$, μ being the cosine of the angle between the line of sight and the pair separation (assuming symmetry with respect to $\mu \rightarrow -\mu$). The line of sight is the direction to the midpoint of the galaxy pair in an aperiodic survey, and a fixed coordinate direction ($\hat{z}$) in periodic boxes.

considered stochastic in the cosmological paradigm. Therefore, it is crucial for scatter in the correlation function measurements.

The NN counts can be expanded as

$$(NN)_a^c = \sum_{i \neq j} n_i n_j w_i w_j \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) \delta_i \delta_j. \quad (2.2)$$

The survey has been divided into cells indexed by i and j . n_i is the ensemble average number of galaxies in the cell i , w_i is the weight for the random point in the cell, δ_i is the fractional galaxy overdensity in the cell, μ_{ij} is the absolute value of the cosine of the angle between the line of sight and the separation vector $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$, r_{ij} is the length of that vector, and Θ are binning functions (unity if the argument fits into the bin and zero otherwise).

Hereafter, we use the following shorthand notation for the covariance matrix:

$$C_{ab}^{cd} \equiv \text{cov} [\hat{\xi}_a^c, \hat{\xi}_b^d] \equiv \langle \hat{\xi}_a^c \hat{\xi}_b^d \rangle - \langle \hat{\xi}_a^c \rangle \langle \hat{\xi}_b^d \rangle. \quad (2.3)$$

This covariance matrix involves an ensemble average of 4 overdensities at up to 4 different positions. The reason is that each correlation function estimator (Equations (2.1) and (2.2)) contains 2 overdensities at 2 different positions, and they are substituted into Equation (2.3). Some of the 4 positions can be the same, resulting in only 3 or 2 distinct positions⁷ (O’Connell et al., 2016).

We then use the following shot-noise approximation to eliminate repeated overdensities at the same position and arrive at the N -point correlation functions O’Connell et al. (2016):

$$(\delta_i)^2 \approx \frac{\alpha_{\text{SN}}}{n_i} (1 + \delta_i). \quad (2.4)$$

The original motivation (O’Connell et al., 2016) refers to the Poisson sampling. We repeat it with slight modifications because the approximation is crucial to the method. Assuming similar weights for galaxies and randoms, the overdensity in cell i is

$$\delta_i = \frac{b_i}{n_i} - 1, \quad (2.5)$$

⁷However, all 4 cannot be the same, because the correlation at zero separation is excluded ($i \neq j$ in Equation (2.2)).

where b_i is the actual number of galaxies in the cell and $n_i = \langle b_i \rangle$ is the expectation value (ensemble average) of that number⁸. We can choose a sufficiently small cell size so that $n_i \ll 1$. Assuming that the galaxies appear in the cell independently, b_i follows the Poisson distribution, and then $\langle b_i^2 \rangle = n_i$. Substituting Equation (2.5) into Equation (2.4) and taking the ensemble average of both sides gives $1/n_i - 1 \approx \alpha_{\text{SN}}/n_i$. This sets the baseline expectation for $\alpha_{\text{SN}} = 1$.

We further explain the overdensity δ_i in the right-hand side of Equation (2.4), because it cancels in the previous argument. With $n_i \ll 1$, the average of $(\delta_i)^2$ is dominated by cells with $b_i = 1$. Whereas almost all the cells have $b_i = 0$, the corresponding $(\delta_i)^2 = 1$ (Equation (2.5)). The fraction of cells with $b_i = 1$ is only $\approx n_i$, but they have a large $(\delta_i)^2 \approx 1/n_i^2$. For $b_i = 2$, the fraction drops significantly to $\approx n_i^2$, whereas the $(\delta_i)^2 \approx 4/n_i^2$ does not increase as much. The contributions to the average of $(\delta_i)^2$ (fraction of cells times the $(\delta_i)^2$ value) decrease further for higher values of b_i . Thus substituting the most significant case, $b_i = 1$, into Equations (2.4) and (2.5) gives $1/n_i^2 - 2/n_i + 1 \approx \alpha_{\text{SN}}/n_i^2$. This is again true for $\alpha_{\text{SN}} = 1$ because $n_i \ll 1$. The right-hand side of Equation (2.4) is dominated by $\delta_i \gg 1$ for $b_i = 1$, therefore, this overdensity is necessary.

Several factors can cause the shot-noise rescaling, i.e., effectively shift α_{SN} from 1. First, the observations of galaxies in the same small volume of a real survey are not independent. There are fundamental limitations to their number due to resolution, number of observations, and number and size of optical fibers in a fiber spectroscopic instrument. Furthermore, in DESI, each fiber is attached to a robotic positioner confined to a specific area of the focal plane (Silber et al., 2023). Therefore, the selection of a target for the fiber (fiber assignment) must depend on other objects present within this patrol area. As a result, the number of observed galaxies b_i is not guaranteed to follow the Poisson distribution, and $\langle b_i^2 \rangle = n_i$ may not hold. Second, different weights on galaxies and random points can alter the cell overdensity estimate $\delta_i = b_i/n_i - 1$. Such weighting is introduced, in particular, for the mitigation of fiber assignment effects on DESI clustering measurements (Bianchi et al., 2024; Ross et al., 2025). The combination of these effects can decrease or increase the shot-noise rescaling.

⁸ n_i can vary from cell to cell.

To recapitulate, the shot-noise approximation (Equation (2.4)) allows us to remove the repeated same-cell overdensities from the covariance matrix estimator. The ensemble averages of a product of overdensities at N different positions are the N -point correlation function by definition. Therefore, the resulting expression for the covariance matrix involves certain sums with the 4-, 3- and 2-point correlation functions. We separate the model covariance matrix into three parts:

$$\tilde{C}_{ab}^{cd}(\alpha_{\text{SN}}) = {}^4C_{ab}^{cd} + \alpha_{\text{SN}} {}^3C_{ab}^{cd} + \alpha_{\text{SN}}^2 {}^2C_{ab}^{cd}. \quad (2.6)$$

These parts, or the d -point terms ${}^d\mathbf{C}$ have the following theoretical expressions⁹:

$$\begin{aligned} {}^4C_{ab}^{cd} &= \frac{1}{(RR)_a^c (RR)_b^d} \sum_{i \neq j \neq k \neq l} n_i n_j n_k n_l w_i w_j w_k w_l \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) \Theta^b(r_{kl}) \Theta^d(\mu_{kl}) \\ &\quad \times \left[\eta_{ijkl}^{(c)} + 2\xi_{ik}\xi_{jl} \right] \\ {}^3C_{ab}^{cd} &= \frac{4}{(RR)_a^c (RR)_b^d} \sum_{i \neq j \neq k} n_i n_j n_k w_i w_j^2 w_k \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) \Theta^b(r_{jk}) \Theta^d(\mu_{jk}) \\ &\quad \times [\zeta_{ijk} + \xi_{ik}] \\ {}^2C_{ab}^{cd} &= \frac{2\delta^{ab}\delta^{cd}}{[(RR)_a^c]^2} \sum_{i \neq j} n_i n_j w_i^2 w_j^2 \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) [1 + \xi_{ij}], \end{aligned} \quad (2.7)$$

where δ^{ab} and δ^{cd} are Kronecker deltas; $\xi_{ij} = \xi(r_{ij}, \mu_{ij})$ is the 2PCF evaluated¹⁰ at the separation between points number i and j .

ζ_{ijk} and $\eta_{ijkl}^{(c)}$ are the 3-point and connected 4-point correlation functions. They are evaluated at the separations between i, j, k and i, j, k, l points, respectively. These non-Gaussian higher-point functions are included in the theoretical expression for completeness. However, evaluating them in practice is challenging. Theoretical models may not cover the necessary range of scales and might also require additional assumptions. Direct measurements from data are costly and noisy because the number of bins increases for higher-point correlation functions. Consequently, the 3- and connected 4-point correlation functions are not used in

⁹In the RASCALC code ${}^d\mathbf{C}$ are estimated using Monte Carlo importance sampling of points from random catalogs (Philcox et al., 2020).

¹⁰To compute ξ_{ij} , the RASCALC code builds a bicubic interpolator based on an input grid of $\xi(r, \mu)$ values.

the practical implementation.

Instead of evaluating them, the covariance matrix has been adjusted with the shot-noise rescaling parameter according to Equation (2.6). The d -point terms have been computed solely with the 2-point function (as designated by crossing out ζ and η in Equation (2.7)). This 2PCF may, however, include non-linear effects, e.g., due to being estimated from the data or a realistic simulation.

The resulting expressions are theoretically analogous to the Gaussian covariances, neglecting the trispectrum contribution and super-sample covariance (e.g., Grieb et al., 2016). With them, one can also rescale the shot noise as a free parameter. The constant shot-noise term in the power spectrum corresponds (via the Fourier transform) to a delta-function addition to the correlation function (which is not commonly considered because we do not measure the correlation function at zero separation). One can then eliminate the delta functions by integrating their contributions to the full 4-point covariance matrix integral over 1 or 2 of the positions and obtain the 3- and 2-point integrals analogous to sums in Equation (2.7). The only difference we found is that the $1 + \xi_{ij}$ factor does not appear in the 2-point term ${}^2C_{ab}^{cd}$. But working in configuration space is advantageous because it allows to naturally account for survey geometry and selection, including the variation of the expected density $\bar{n}$ across the survey volume, by sampling points from the random catalog.

The key shot-noise rescaling parameter value can be chosen to optimally match a reference covariance obtained from a set of mocks. The best match is quantified by minimal Kullback-Leibler divergence¹¹:

$$D_{\text{KL}}\left[\tilde{\mathbf{C}}^{-1}(\alpha_{\text{SN}}), \mathbf{C}_S\right] = \frac{1}{2} \left[\text{tr}\left(\tilde{\mathbf{C}}^{-1}(\alpha_{\text{SN}})\mathbf{C}_S\right) - N_{\text{bins}} - \ln \det\left(\tilde{\mathbf{C}}^{-1}(\alpha_{\text{SN}})\mathbf{C}_S\right) \right]. \quad (2.8)$$

In its computation, the model covariance $\tilde{\mathbf{C}}(\alpha_{\text{SN}})$ (with elements given by Equation (2.6)) is inverted because the sample covariance of the simulations $\mathbf{C}_S$ is more noisy. With a smooth theoretical model and only one parameter (α_{SN}) to adjust, fewer mocks are required than for

¹¹Which is determined by the covariance matrices for two multivariate normal distributions with the same mean.

direct use of their sample covariance matrix (O’Connell et al., 2016). The process is illustrated in Figure 2.2. In other words, fitting the RASCALC covariance model to mocks is akin to template-based smoothing of the simulation-based sample covariance matrix.

Theoretically, the inversion of the RASCALC covariance gives a slightly biased estimate of the precision matrix. However, the Hartlap factor (Hartlap et al., 2007) is not applicable since it is not a sample covariance. The relevant second-order bias correction matrix accounting for importance sampling noise has been worked out (O’Connell & Eisenstein, 2019):

$$\tilde{\Psi} = (\mathbb{I} - \tilde{\mathbf{D}}) \tilde{\mathbf{C}}^{-1} \quad (2.9)$$

$$\tilde{\mathbf{D}} = \frac{N_{\text{subsamples}} - 1}{N_{\text{subsamples}}} \left[-\mathbb{I} + \frac{1}{N_{\text{subsamples}}} \sum_{i=1}^{N_{\text{subsamples}}} \tilde{\mathbf{C}}_{[i]}^{-1} \tilde{\mathbf{C}}_i \right],$$

which uses the partial covariance estimates $\tilde{\mathbf{C}}_i$ from $N_{\text{subsamples}}$ distinct sets of configurations resulting from importance sampling in the estimation of sums (Eq. (2.7) or (2.13)) and mean of all the partial estimates but the i ’th $\tilde{\mathbf{C}}_{[i]}$. However, we find it practically insignificant: the eigenvalues of $\tilde{\mathbf{D}}$ are $\lesssim 10^{-3}$.

Unfortunately, the mock-fitting approach does not solve all issues with the mocks. The simulations still take extra time to produce and impose further assumptions and approximations. Therefore, a different method is desirable to reduce the dependence on simulations.

An alternative reference covariance can be obtained from the data with jackknife resampling. However, the jackknife covariance matrix is not perfectly representative of the true, full-survey covariance for several reasons. First, the resampled pieces of realistic data have different geometry from the full dataset and each other, which affects the covariance in a complicated way. Second, the pieces of the data are correlated (Mohammad & Percival, 2022; Trusov et al., 2024). Therefore, it is safer to develop a separate theoretical model for jackknife covariance (O’Connell & Eisenstein, 2019), which we briefly explain in the following.

The method uses a slightly non-standard formalism, dubbed *unrestricted jackknife* (Philcox et al., 2020). There the jackknife correlation function estimate ξ_A is not the auto-correlation of the whole survey excluding the jackknife region A (as in exclude-one, *restricted* jackknife),

but the cross-correlation function between that region and the whole survey. Equivalently, this means that the additional jackknife weighting factor for the pair of points i, j , q_{ij}^A , is 1 if both of them belong to the jackknife region A , $1/2$ if only one, and 0 if neither. The pair counts can be converted from different terms (auto and cross jackknife counts), often saved separately in light of Mohammad-Percival correction (Mohammad & Percival, 2022).

The unrestricted jackknife is convenient since the full pair counts of each type are the sum of all the jackknife pair counts of the same type. Then if one weights the regions by the RR pair counts,

$$(w_A)_a^c = \frac{(RR_A)_a^c}{(RR)_a^c}, \quad (2.10)$$

the weighted mean correlation function is equal to the full-survey one. This simplifies the theoretical modeling of the jackknife covariance.

The data jackknife covariance estimate is then

$$(C_J)_{ab}^{cd} = \frac{\sum_A (w_A)_a^c (w_A)_b^d \left[\left(\hat{\xi}_A \right)_a^c - \hat{\xi}_a^c \right] \left[\left(\hat{\xi}_A \right)_b^d - \hat{\xi}_b^d \right]}{1 - \sum_A (w_A)_a^c (w_A)_b^d}, \quad (2.11)$$

the corresponding theoretical estimate, $(\tilde{C}_J)_{ab}^{cd}$, is constructed analogously to Equation (2.6):

$$(\tilde{C}_J)_{ab}^{cd}(\alpha_{\text{SN}}) = ({}^4C_J)_{ab}^{cd} + \alpha_{\text{SN}} ({}^3C_J)_{ab}^{cd} + \alpha_{\text{SN}}^2 ({}^2C_J)_{ab}^{cd}. \quad (2.12)$$

with terms defined as (Philcox et al., 2020):

$$\begin{aligned} ({}^4C_J)_{ab}^{cd} &= \frac{1}{(RR)_a^c (RR)_b^d \left[1 - \sum_A (w_A)_a^c (w_A)_b^d \right]} \sum_{i \neq j \neq k \neq l} n_i n_j n_k n_l w_i w_j w_k w_l \\ &\quad \times \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) \Theta^b(r_{kl}) \Theta^d(\mu_{kl}) \left[\cancel{\eta_{ijkl}^{(c)}} + \xi_{ij} \xi_{kl} + 2\xi_{ik} \xi_{jl} \right] (\omega_{ijkl})_{ab}^{cd} \\ ({}^3C_J)_{ab}^{cd} &= \frac{4}{(RR)_a^c (RR)_b^d \left[1 - \sum_A (w_A)_a^c (w_A)_b^d \right]} \sum_{i \neq j \neq k} n_i n_j n_k w_i w_j^2 w_k \\ &\quad \times \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) \Theta^b(r_{jk}) \Theta^d(\mu_{jk}) [\cancel{\xi_{ijk}} + \xi_{ik}] (\omega_{ijk})_{ab}^{cd} \\ ({}^2C_J)_{ab}^{cd} &= \frac{2\delta^{ab}\delta^{cd}}{[(RR)_a^c]^2 \left\{ 1 - \sum_A [(w_A)_a^c]^2 \right\}} \sum_{i \neq j} n_i n_j w_i^2 w_j^2 \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) [1 + \xi_{ij}] (\omega_{ijij})_{ab}^{cd}, \end{aligned} \quad (2.13)$$

where $(\omega_{ijkl})_{ab}^{cd}$ is an additional jackknife weight tensor:

$$(\omega_{ijkl})_{ab}^{cd} = \sum_A [q_{ij}^A - (w_A)_a^c] [q_{kl}^A - (w_A)_b^d]. \quad (2.14)$$

To sum up, in the fiducial (jackknife) pipeline (Figure 2.1), we obtain the shot-noise rescaling by fitting the model for its jackknife covariance (Equations (2.12) and (2.13)) to the data (Equation (2.11)). As in the mock approach, this specifically means minimizing the Kullback-Leibler (KL) divergence between the covariance matrices:

$$D_{\text{KL}}[\tilde{\mathbf{C}}_J^{-1}(\alpha_{\text{SN}}), \mathbf{C}_J] = \frac{1}{2} \left[\text{tr}(\tilde{\mathbf{C}}_J^{-1}(\alpha_{\text{SN}}) \mathbf{C}_J) - N_{\text{bins}} - \ln \det(\tilde{\mathbf{C}}_J^{-1}(\alpha_{\text{SN}}) \mathbf{C}_J) \right]. \quad (2.15)$$

where the model jackknife covariance $\tilde{\mathbf{C}}_J(\alpha_{\text{SN}})$ (Equation (2.12)) is inverted (Philcox et al., 2020), because the data jackknife covariance matrix $\mathbf{C}_J$ (Equation (2.11)) is often not invertible. The final covariance is obtained by plugging the resulting shot-noise rescaling values into the full covariance model (Equation (2.6)).

To summarise, the key assumption is that shot-noise rescaling of purely Gaussian contributions (i.e., ignoring 3-point and connected 4-point functions) can produce a realistic covariance matrix in configuration space. A theoretical motivation is that non-Gaussian contributions primarily affect the squeezed configurations involving small-scale correlations, below the bin width for the 2-point function, therefore not distinguishable from shot noise operating on infinitesimally small scales. The method has been empirically shown to agree well with mock-based covariances (O’Connell et al., 2016; Vargas-Magaña et al., 2018; Rashkovetskyi et al., 2025b).

2.2.2 Split random-random computation

Keihänen et al. (2019) showed that splitting the random catalog into a number of sub-catalogs of the same size as the data catalog when calculating random–random pairs and excluding pairs across different sub-catalogs provides the optimal error at a fixed computational cost. The splitting can be used in RASCALC. It gives little to no speed-up and impact on results because the importance sampling is too far from complete. However, it can be useful for multi-node

parallelization. This approach has been used for the data-based RASCALC computation in Moon et al. (2023).

A robust implementation of split random-random pair calculations in RASCALC would require considering only quadruples of random points where members of each pair are from the same sub-catalog, but the pairs can be from different catalogs. However, this has been found to have little to no impact on the results, probably due to the fact that importance sampling covers only a small fraction of all possible configurations. At the same time, such implementation makes the code less efficient and makes it impossible to split the computation of different catalogs between nodes.

2.2.3 Reconstructed two-point function covariance

Standard BAO reconstruction modifies the correlation function estimator and thus requires adjustments in the covariance. Standard BAO reconstruction procedures shift the positions of both the data and random points. Only shifted data (D) is used, whereas the randoms are kept in two variants: original (R) and shifted (S). The correlation function is estimated via the Landy-Szalay estimator (Equation (2.1)) with $N = D - S$ instead of $D - R$, but still RR in the denominator. This means shifted randoms are to be used in sums representing NN (starting from Equation (2.2)). Through the expansion of Equation (2.3) these propagate into the sums for the covariance matrix terms (Equations (2.7) and (2.13)).

Thus in the computations of the covariance matrix terms (Equations (2.7) and (2.13)) for reconstructed catalogs we use

- the shifted randoms S in the sums (i.e. $\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \mathbf{r}_l$ used for binning functions and correlation function interpolation are the shifted random positions) because they come from expanding NN , which now involves S and not R ;
- non-shifted random counts (i.e., still RR) for normalization¹²;

¹²Or correction function(s) in the original implementation of Legendre moments.

- two-point correlation functions ξ_{ij} etc. with a different normalization¹³: having SS instead of RR in the denominator, i.e.

$$\left(\hat{\xi}_{\text{in}}\right)_a^c = \frac{(NN)_a^c}{(SS)_a^c}. \quad (2.16)$$

For data jackknife covariance (Equation (2.11)), we still use the ordinary normalization of 2PCF (Equation (2.1) with RR in the denominator but $N = D - S$ instead of $D - R$).

In this approach, we treat the shifted randoms as fixed independently of the realization of the Universe’s density field. This is not precisely true because the shifts applied to these randoms depend on the data (observed galaxies). However, this should be a small effect, because we find the resulting covariances match the mock-based ones well.

Shifted randoms are individual for each mock catalog. Therefore, they can not be defined clearly for mock-averaged computations. In those cases, we continue to use the non-shifted randoms everywhere for consistency.

2.2.4 Revisited covariance for projected Legendre moments of 2PCF

Theoretical models of large-scale structure often use multipole (Legendre) moments of the 2-point correlation function instead of its angularly (μ) binned estimates (e.g. Chuang & Wang (2013); Chen et al. (2024b)). Moreover, the number of multipoles of interest is typically low — monopole, quadrupole, and sometimes hexadecapole (Maus et al., 2025a,b; Noriega et al., 2025; Lai et al., 2025; Ramirez-Solano et al., 2025)¹⁴. It is then convenient to compress the correlation by converting the angularly-binned correlation function (with more than 3 angular bins) to the Legendre moments. Covariance matrices for Legendre multipoles have a lower dimension, which causes fewer numerical problems and makes them easier to estimate directly.

The covariance matrix model for Legendre moments was developed in RASCALC previously (Philcox & Eisenstein, 2019), but this implementation has several disadvantages.

¹³In the code, this is implemented by renormalizing the input grid of correlation values $\xi(r, \mu)$.

¹⁴These references discuss the power spectrum modeling, but the ℓ ’th Legendre moment of the correlation function ξ_ℓ is determined only by the same-order multipole of the power spectrum P_ℓ via a spherical Bessel j_ℓ transform.

First, it is not directly compatible with jackknives. In practice, producing the Legendre moment covariance with optimal shot-noise rescaling based on data (not relying on a mock sample) requires two separate computations: one for angular (μ) bins with jackknives to tune the shot-noise rescaling, and another to construct the full covariance matrix model for Legendre multipoles. This is an inconvenience when one does not intend to use the angularly-binned correlation function in cosmological inference.

Second, the 2PCF estimation library used in DESI, PYCORR¹⁵ (De Mattia et al., 2024), operates under slightly different assumptions. PYCORR uses the angularly binned 2-point correlation function $\hat{\xi}_a^c$ (Equation (2.1)) estimates with a large but finite number (~ 100) of angular bins to compute the radially binned Legendre moments:

$$\hat{\xi}_a^\ell = (2\ell + 1) \sum_c \hat{\xi}_a^c \int_{\Delta\mu_c} d\mu L_\ell(\mu) = \sum_c \hat{\xi}_a^c F_c^\ell; \quad (2.17)$$

where

$$F_c^\ell \equiv (2\ell + 1) \int_{\Delta\mu_c} d\mu L_\ell(\mu) \quad (2.18)$$

are the projection factors, which do not depend on radial bins or the tracers involved in the correlation function. The equations above assume even multipole index ℓ , and binning in $|\mu| \in [0, 1]$ ¹⁶. In contrast, the RASCALC estimators developed earlier assume weighting by Legendre polynomials during pair counting (Philcox & Eisenstein, 2019). This is equivalent to using infinitesimally narrow angular bins in Equation (2.17). The mismatch in assumptions between PYCORR and RASCALC may not cause significant differences in practice, but it is not desirable.

Third, an additional step was needed to account for realistic survey geometry. The previous RASCALC realization relies on the survey correction function — the ratio of pair counts in a real survey and a periodic box with the same volume (Xu et al., 2010). This function needed to be modeled for arbitrary angles ($|\mu|$). A piecewise-polynomial form has been assumed

¹⁵<https://github.com/cosmodesi/pycorr>

¹⁶PYCORR is capable of binning in $\mu \in [-1, 1]$, retaining the sign information. But all auto-correlation functions are necessarily symmetric, and it is easy to “wrap” the counts to $|\mu|$ bins in any case as long as the number of μ bins is even.

(Philcox & Eisenstein, 2019). Whereas they explain the need for two different polynomials by the particularly strong redshift-space distortions near the line of sight ($|\mu| = 1$), the choice of the partition point¹⁷ at $|\mu| = 0.75$ has not been motivated and may not be the best. This might be a minor issue as well, but still a source of extra uncertainty.

We have seen an opportunity to address all three issues and streamline the covariance matrix computation procedure for extensive usage with DESI. Since the projection in Equation (2.17) is linear, the covariance matrix for these Legendre moments estimators can be obtained from the r, μ -binned one given by Equation (2.6):

$$\tilde{C}_{ab}^{\ell\ell'} \equiv \text{cov} \left[\hat{\xi}_a^\ell, \hat{\xi}_b^{\ell'} \right] = \sum_{c,d} \tilde{C}_{ab}^{cd} F_c^\ell F_d^{\ell'}. \quad (2.19)$$

A major technical result of this paper is a methodology to compute this covariance matrix of the Legendre multipoles directly at the level of the summation over point configurations, rather than having to compute and then project the much larger covariance matrix of fine angular bins. For this, several quantities need to be inserted into the sums of Equation (2.7), and we obtain the following 4, 3, and 2-point terms:

$$\begin{aligned} {}^4C_{ab}^{\ell\ell'} &= \sum_{i \neq j \neq k \neq l} n_i n_j n_k n_l w_i w_j w_k w_l \Theta^a(r_{ij}) \Theta^b(r_{kl}) \left[\cancel{\eta_{ijkl}^{(c)}} + 2\xi_{ik}\xi_{jl} \right] \\ &\quad \times \sum_c \frac{\Theta^c(\mu_{ij}) F_c^\ell}{(RR)_a^c} \sum_d \frac{\Theta^d(\mu_{kl}) F_d^{\ell'}}{(RR)_b^d}, \\ {}^3C_{ab}^{\ell\ell'} &= 4 \sum_{i \neq j \neq k} n_i n_j n_k w_i w_j^2 w_k \Theta^a(r_{ij}) \Theta^b(r_{jk}) [\cancel{\zeta_{ijk}} + \xi_{ik}] \sum_c \frac{\Theta^c(\mu_{ij}) F_c^\ell}{(RR)_a^c} \sum_d \frac{\Theta^d(\mu_{jk}) F_d^{\ell'}}{(RR)_b^d}, \\ {}^2C_{ab}^{\ell\ell'} &= 2\delta^{ab} \sum_{i \neq j} n_i n_j w_i^2 w_j^2 \Theta^a(r_{ij}) [1 + \xi_{ij}^{XY}] \sum_c \frac{\Theta^c(\mu_{ij}) F_c^\ell F_c^{\ell'}}{[(RR)_a^c]^2}. \end{aligned} \quad (2.20)$$

As in Equation (2.7), we include the non-Gaussian higher-point functions in these theoretical equations. However, we drop them in the current implementation, which we signify by crossing them out. Sums like $\sum_c \Theta^c(\mu) \dots$ practically mean finding the angular bin $\tilde{c}$ to which the μ value belongs and then evaluating the rest only for that one bin. Within the code, we sample a quadruplet, triplet, or pair of particles and then accumulate their contribution to all the

¹⁷I.e. the boundary between the two polynomials, where they join continuously and smoothly.

Legendre multipole moments in their radial bins.

These 4, 3, and 2-point terms can be combined to the full theoretical estimate analogously to Equation (2.6), i.e.

$$\tilde{C}_{ab}^{\ell\ell'}(\alpha_{\text{SN}}) = {}^4C_{ab}^{\ell\ell'} + \alpha_{\text{SN}} {}^3C_{ab}^{\ell\ell'} + \alpha_{\text{SN}}^2 {}^2C_{ab}^{\ell\ell'}. \quad (2.21)$$

This is the single-tracer expression; the version for multiple tracers is provided in Section A.1.2.

For simplicity, we have decided to reuse the r, μ binned jackknife covariance matrix estimate (Equation (2.11)). An alternative could be a tedious re-derivation of the theoretical jackknife covariance model with some weights for individual jackknife multipole estimators. The method only requires this step to calibrate the shot-noise rescaling parameter. The precise choice of the reference jackknife covariance should not matter as long as the data and model estimators are treated consistently. Consequently, we project the angularly-binned data jackknife covariance matrix (Equation (2.11)) similarly to the full covariance (Equation (2.19)):

$$(C_J)_{ab}^{\ell\ell'} = \sum_{c,d} (C_J)_{ab}^{cd} F_c^\ell F_d^{\ell'} \quad (2.22)$$

and do the same with the theoretical prediction (Equation (2.12)):

$$(\tilde{C}_J)_{ab}^{\ell\ell'}(\alpha_{\text{SN}}) = ({}^4C_J)_{ab}^{\ell\ell'} + \alpha_{\text{SN}} ({}^3C_J)_{ab}^{\ell\ell'} + \alpha_{\text{SN}}^2 ({}^2C_J)_{ab}^{\ell\ell'}. \quad (2.23)$$

The jackknife d -point terms (Equation (2.13)) accordingly are transformed to

$$\begin{aligned} ({}^4C_J)_{ab}^{\ell\ell'} &= \sum_{i \neq j \neq k \neq l} n_i n_j n_k n_l w_i w_j w_k w_l \Theta^a(r_{ij}) \Theta^b(r_{kl}) \left[\cancel{\eta_{ijkl}^{(\varphi)}} + \xi_{ij} \xi_{kl} + 2\xi_{ik} \xi_{jl} \right] \\ &\times \sum_{c,d} \frac{(\omega_{ijkl})_{ab}^{cd} \Theta^c(\mu_{ij}) \Theta^d(\mu_{kl}) F_c^\ell F_d^{\ell'}}{(RR)_a^c (RR)_b^d \left[1 - \sum_A (w_A)_a^c (w_A)_b^d \right]}, \\ ({}^3C_J)_{ab}^{\ell\ell'} &= 4 \sum_{i \neq j \neq k} n_i n_j n_k w_i w_j^2 w_k \Theta^a(r_{ij}) \Theta^b(r_{jk}) [\cancel{\xi_{ijk}} + \xi_{ik}] \\ &\times \sum_{c,d} \frac{(\omega_{ijjk})_{ab}^{cd} \Theta^c(\mu_{ij}) \Theta^d(\mu_{jk}) F_c^\ell F_d^{\ell'}}{(RR)_a^c (RR)_b^d \left[1 - \sum_A (w_A)_a^c (w_A)_b^d \right]}, \\ ({}^2C_J)_{ab}^{\ell\ell'} &= 2\delta^{ab} \sum_{i \neq j} n_i n_j (w_i w_j)^2 \Theta^a(r_{ij}) [1 + \xi_{ij}] \sum_c \frac{(\omega_{ijij})_{ab}^{cc} \Theta^c(\mu_{ij}) F_c^\ell F_c^{\ell'}}{[(RR)_a^c]^2 \left\{ 1 - \sum_A [(w_A)_a^c]^2 \right\}}. \end{aligned} \quad (2.24)$$

Similarly to Equation (2.20), sums of the form $\sum_{c,d} \Theta^c(\mu_1) \Theta^d(\mu_2) \dots$ practically mean finding the angular bins $\tilde{c}, \tilde{d}$ to which the $\mu_{1,2}$ values belong correspondingly and then evaluating the rest (...) only for that pair of bins. A given set of cells contributes to the covariance of one pair of angular bins (c, d in Equation (2.7)), but all multipoles (ℓ, ℓ').

We have also omitted the small disconnected part of the 4-point jackknife term ($\xi_{ij}\xi_{kl}$) in the main code implementation for practical reasons. This part is estimated in r, μ bins (Equation (2.13)) by separating it into a product of two sums that need to be computed for each jackknife region (Philcox et al., 2020). The evaluation becomes less convenient in Legendre multipoles due to the projection factors inserted into the sum over cells/particles. We have computed the disconnected term in a couple of realistic setups by using the technique for r, μ bins¹⁸ and projecting the resulting covariance matrix part into the multipole moments (analogously to Equation (2.22)). We have found that the inclusion of the disconnected term does not change the shot-noise rescaling values up to the 6th digit after the decimal point. The optimal shot-noise rescaling parameter is the only link between the disconnected jackknife term and the final covariance matrix. Therefore, we concluded that the impact of the disconnected term is practically negligible.

In addition, we give a theoretical justification for neglecting the disconnected term, although not completely strict. The disconnected jackknife term vanishes exactly if either the correlation function is constant, the jackknife regions are identical, or the jackknife counts in each region are the same in each fine bin (Philcox et al., 2020). An arbitrary survey would not meet any of these conditions exactly, but they likely hold approximately. Therefore, the disconnected term is expected to be small.

This concludes the description of the new covariance estimators we apply to DESI data and mocks. To reiterate, the key practical advantage is that a single computation provides the Legendre covariance with shot-noise rescaling tuned on jackknives. Additionally, the new estimators use the random-random counts directly, instead of relying on a survey correction

¹⁸Storing more data due to keeping many (100) angular (μ) bins for each of 60 jackknife regions. The number of angular bins could be reduced, but at the price of potential loss of precision.

function fit for realistic survey geometry (Philcox & Eisenstein, 2019). Therefore, we use this method for covariance matrix estimation in the rest of this paper.

2.3 Methods of comparison of covariance matrices

Since a covariance matrix is a high-dimensional object, it can be hard to explore and interpret. Moreover, we run the pipeline multiple times independently and aim to study all the covariance matrix products to assess their stability and fluctuations. Thus, compact and numerical comparison measures are instructive.

2.3.1 Interpretable measures of similarity for covariance matrices (original version, needs to be merged with the other one)

The first characteristic we consider is the Kullback-Leibler (KL) divergence, a measure of distance between distributions used to fit covariances in RASCALC (Section 2.2.1 and Equation (2.26)). It is generally defined as the expectation value of the logarithm of the ratio of the two probability distribution functions according to the first distribution:

$$D_{\text{KL}}(P_1||P_2) = \int \ln \left(\frac{P_1(x)}{P_2(x)} \right) P_1(x) dx \quad (2.25)$$

By this expression, KL divergence can be seen as an average difference in log-likelihood. For two Gaussian distributions with covariance matrices $\mathbf{C}_i$ and precision matrices $\mathbf{\Psi}_i = \mathbf{C}_i^{-1}$ describing N_{bins} observables (correlation function bins in our setup), it can be found as

$$D_{\text{KL}}(\mathbf{\Psi}_1, \mathbf{C}_2) = \frac{1}{2} [\text{tr}(\mathbf{\Psi}_1 \mathbf{C}_2) - N_{\text{bins}} - \ln \det(\mathbf{\Psi}_1 \mathbf{C}_2)]. \quad (2.26)$$

The KL divergence computed using the RASCALC precision matrix and mock sample covariance is related to the log-likelihood of the sample covariance under the assumption that the RASCALC covariance truly describes the distribution of mock clustering measurements¹⁹ (O’Connell et al., 2016) This is very appropriate for testing the hypothesis that the RASCALC precision

¹⁹This log-likelihood relation does not hold if the KL divergence is computed between sample precision (inverse covariance) and theoretical covariance matrices.

matrix is a precise, unbiased estimate.

The next metric assesses how close the first precision matrix is to the inverse of the second covariance matrix, and at the same time, a “directional” root-mean-square relative difference in χ^2 given by the two covariance matrices (explained in more detail in Section A.2.2):

$$\begin{aligned} R_{\text{inv}}(\Psi_1, \mathbf{C}_2) &= \frac{1}{\sqrt{N_{\text{bins}}}} \left\| \mathbf{C}_2^{1/2} \Psi_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right\|_F \\ &= \sqrt{\frac{\text{tr} \left[(\Psi_1 \mathbf{C}_2 - \mathbb{I})^2 \right]}{N_{\text{bins}}}}. \end{aligned} \quad (2.27)$$

This measure can also be seen as the average relative difference in the errorbars. Moreover, if the covariance matrix is estimated from a sample with multivariate normal distribution and the precision matrix is assumed to be true, R_{inv}^2 is proportional to the χ^2 computed using the covariance of independent covariance matrix elements (Section A.2.2). Thus, in this case, it can serve as an approximation of log-likelihood for optimization.

The last metric is akin to the mean reduced χ^2 of samples corresponding to one covariance matrix with respect to the other precision matrix:

$$\chi_{\text{red}}^2(\Psi_1, \mathbf{C}_2) = \frac{1}{N_{\text{bins}}} \text{tr}(\Psi_1 \mathbf{C}_2). \quad (2.28)$$

It can be seen as the mean ratio of χ^2 given by the two covariance/precision matrices.

All three metrics are not symmetric, meaning that values for $(\Psi_1, \mathbf{C}_2)$ and $(\Psi_2, \mathbf{C}_1)$ may be different, so in principle, it might be informative to consider differences both ways. On the other hand, the sample covariance is less robust than the RASCALC result, and its inversion can be less stable. Moreover, computing each metric twice makes the results more numerous and less clear. Finally, some additional properties do not hold with the other one, as noted above. Therefore, we decided to limit ourselves to RASCALC precision matrices and sample covariance matrices.

For a better understanding of the metrics, let us consider the eigenvalues of $\Psi_1 \mathbf{C}_2$ (alternatively, one can use $\Psi_1^{1/2} \mathbf{C}_2 \Psi_1^{1/2}$ which is symmetric) and denote them as λ_a . We would like

$\Psi_1 \rightarrow \mathbf{C}_2^{-1}$ thus all $\lambda_a \rightarrow 1$. The metrics then can be expressed as

$$D_{\text{KL}}(\Psi_1, \mathbf{C}_2) = \frac{1}{2} \sum_{a=1}^{N_{\text{bins}}} [\lambda_a - 1 - \ln \lambda_a] \approx \frac{1}{4} \sum_{a=1}^{N_{\text{bins}}} (\lambda_a - 1)^2, \quad (2.29)$$

$$R_{\text{inv}}(\Psi_1, \mathbf{C}_2) = \sqrt{\frac{1}{N_{\text{bins}}} \sum_{a=1}^{N_{\text{bins}}} (\lambda_a - 1)^2}, \quad (2.30)$$

$$\chi_{\text{red}}^2(\Psi_1, \mathbf{C}_2) = \frac{1}{N_{\text{bins}}} \sum_{a=1}^{N_{\text{bins}}} \lambda_a. \quad (2.31)$$

Thus D_{KL} and R_{inv} accumulate any deviation of λ_a from 1, although they cannot indicate the direction of such differences. Note that the quadratic expression for D_{KL} is approximate, so it is not generally degenerate with R_{inv} , although as the covariance matrices approach each other, these two measures become more redundant:

$$D_{\text{KL}}(\Psi_1, \mathbf{C}_2) \approx \frac{N_{\text{bins}}}{4} R_{\text{inv}}^2(\Psi_1, \mathbf{C}_2). \quad (2.32)$$

χ_{red}^2 can show which covariance matrix is “larger” on average, while deviations in opposite directions may cancel each other.

Finally, it is important to set expectations for these three comparison measures in the perfect case. By this, we mean comparing the true precision matrix with the sample covariance estimated from n_S multivariate normal samples following the same covariance. The distribution of clustering statistics can be assumed normal. This allows us to test the hypothesis that RASCALC can predict the true covariance of the mocks.

In this case, we need to focus on the noise properties of the sample covariance matrix $\mathbf{C}_S$ obtained via the standard unbiased estimator for the case when the true mean is not known:

$$C_{S,ab} = \frac{1}{n_S - 1} \sum_{i=1}^{n_S} (\xi_{a,i} - \bar{\xi}_a)(\xi_{b,i} - \bar{\xi}_b) \quad (2.33)$$

where a, b denote bin numbers, i, j index sample numbers, and $\bar{\xi}_a$ is the estimate of the mean:

$$\bar{\xi}_a \equiv \frac{1}{n_S} \sum_{i=1}^{n_S} \xi_{a,i}. \quad (2.34)$$

Since the clustering measurements are described well by a multivariate normal distribution,

their sample covariance matrix follows the Wishart distribution. This provides a reference of how the metrics behave when the perfect precision matrix is compared to a covariance matrix estimated from n_S samples with N_{bins} bins (or any other Gaussian observables).

Full derivations are presented in Section A.2; here, we will only provide the results for mean/expectation values and standard deviations:

$$\begin{aligned}\langle D_{\text{KL}}(\Psi_0, \mathbf{C}_S) \rangle &\approx \frac{N_{\text{bins}}(N_{\text{bins}} + 1)}{4(n_S - 1)}, \\ \sigma[D_{\text{KL}}(\Psi_0, \mathbf{C}_S)] &\approx \frac{1}{2} \sqrt{\frac{N_{\text{bins}}[(N_{\text{bins}} + 1)(n_S + 2N_{\text{bins}} + 2) + 2]}{(n_S - 1)^3}},\end{aligned}\quad (2.35)$$

$$\begin{aligned}\langle R_{\text{inv}}(\Psi_0, \mathbf{C}_S) \rangle &\approx \sqrt{\frac{N_{\text{bins}} + 1}{n_S - 1}}, \\ \sigma[R_{\text{inv}}(\Psi_0, \mathbf{C}_S)] &\approx \frac{1}{n_S - 1} \sqrt{\frac{(N_{\text{bins}} + 1)(n_S + 2N_{\text{bins}} + 2) + 2}{N_{\text{bins}}(N_{\text{bins}} + 1)}}.\end{aligned}\quad (2.36)$$

Naively, one could expect D_{KL} and R_{inv} to become arbitrarily small as $\Psi_1 \rightarrow \Psi_0$. However, in reality, they can have large expectation values, especially as the number of bins increases.

χ_{red}^2 , however, would behave like the reduced χ^2 with $N_{\text{bins}} \times (n_S - 1)$ degrees of freedom in this case (see Section A.2.3):

$$\begin{aligned}\langle \chi_{\text{red}}^2(\Psi_0, \mathbf{C}_S) \rangle &= 1, \\ \sigma[\chi_{\text{red}}^2(\Psi_0, \mathbf{C}_S)] &= \sqrt{\frac{2}{N_{\text{bins}}(n_S - 1)}}.\end{aligned}\quad (2.37)$$

It might seem like D_{KL} and R_{inv} could be unbiased by multiplying one of the matrices by a factor similar to the Hartlap factor (Hartlap et al., 2007), but a lack of bias in χ_{red}^2 expectation value suggests that this is not true.

It is notable that a deviation of χ_{red}^2 from 1 would contribute to R_{inv} – see Equations (2.30) and (2.31); their consequence is also that

$$|\chi_{\text{red}}^2(\Psi_1, \mathbf{C}_2) - 1| \leq R_{\text{inv}}(\Psi_1, \mathbf{C}_2). \quad (2.38)$$

However, the direction of such deviation will not be clear in R_{inv} , and a nontrivial expectation

value can make it harder to interpret. This keeps χ_{red}^2 useful in many cases.

For additional validation, we also compute the statistical means and standard deviations for our comparison measures empirically (Monte-Carlo method) with a large number of multivariate normal samples. We generate $N_r = 10,000$ chunks of n_S independent samples of N_{bins} -dimensional multivariate normal vectors with a (true) unit covariance matrix²⁰. In each chunk, we estimate the sample covariance and compute the comparison measures between their true (unit) precision matrix and the sample covariance estimate. Finally, we estimate the mean and standard deviation for each comparison measure using the obtained N_r random realizations²¹.

We choose a preferred (fiducial) value between the theoretical estimate and the corresponding empirical figure in each case. We prefer the theoretical value unless the difference is more than 3σ . Otherwise, we select the empirical estimate. As suspected, we only find $> 3\sigma$ deviations for KL divergence (with larger numbers of bins) and R_{inv} (for smaller numbers of bins). The fiducial values are used in Tables 2.12 to 2.17 as reference for comparison measures between RASCALC precision matrices and sample covariances.

2.3.2 Internal convergence assessment

Internal consistency of RASCALC covariance matrices in one run and the convergence of the Monte-Carlo integration procedure are also important to assess quantitatively. We propose to employ the above-mentioned methods to accomplish this and provide valuable diagnostics that do not rely on a reference (e.g., sample) covariance and thus can be used in any run, including the pure data-based one. However, we need to note that such a test can only quantify limited sources of uncertainty or error, leaving aside the factors like adequacy of the approximations in the formalism, the precision of the input clustering, and noise in the jackknife covariance

²⁰The value of the true covariance matrix does not matter because it cancels out in each comparison measure (ignoring the numerical instabilities). The dimension, however, is crucially important. Therefore, we repeat the procedure for each value of N_{bins} relevant for comparisons in this work.

²¹These realizations of covariance matrix comparison measures (for $n_S = 1000$), as well as all estimates of their means and standard deviations (theoretical, empirical and our fiducial choice between the two for each case), are provided in the supplementary material: doi:10.5281/zenodo.10895161.

estimated from the data.

The RASCALC code provides multiple partial intermediate results corresponding to practically non-overlapping sets of quadruples, triples, and pairs of points. These resulting covariance matrices can be split into two distinct sets of similar size, averaged within them, and compared using the three metrics. In this case, however, the arguments for the χ_{red}^2 weaken – we can expect R_{inv} to become arbitrarily low as the number of Monte-Carlo samples increases, which would limit the reduced chi-squared via Equation (2.38), and it is not as interesting to understand which of the halves gives a “smaller” matrix. Then D_{KL} also becomes more redundant with R_{inv} via Equation (2.32). Therefore, it is reasonable to only show R_{inv} , which can be seen as an estimate of root-mean-square relative precision (considered over all directions in measurement space).

2.3.3 Fisher projection to the space of model parameters

Comparing the full covariance matrices for observables (e.g., binned 2PCF or 2PCF multipoles) may be overly generic. In such a case, we consider every arbitrary “direction” in the high-dimensional linear space of these observables. Some of these directions may be unphysical, whereas others can have little or no connection to the data analysis. To highlight a smaller number of meaningful, impactful directions, Philcox et al. (2021); Ferreira & Marra (2022) suggested projecting the covariance matrices into a model-dependent subspace.

We project the covariance matrices to the model parameters through the inverse of the Fisher matrix for simplicity. For RASCALC results, one can use a simple expression neglecting the inversion bias (Hartlap-like) corrections²²:

$$\mathbf{C}_R^{\text{par}} = \left[\mathbf{M} \left(\mathbf{C}_R^{\text{obs}} \right)^{-1} \mathbf{M}^T \right]^{-1}. \quad (2.39)$$

$\mathbf{C}_R^{\text{obs}}$ is the RASCALC covariance matrix restricted to the separation and multipole range of the respective model. This range constitutes N'_{bins} observables. $\mathbf{M}$ is the Jacobian of the model, i.e., the matrix of derivatives of radially-binned 2PCF multipoles ξ with respect to the

²²As we mentioned in Section 2.2.1, such corrections are small for RASCALC.

parameter vector $\boldsymbol{\theta}$:

$$M_{pa} \equiv \frac{\partial \xi_a}{\partial \theta_p}. \quad (2.40)$$

For the mock sample covariance, we use a similar expression:

$$\mathbf{C}_S^{\text{par}} = \frac{n_S - 1}{n_S - N'_{\text{bins}} + N_{\text{pars}} - 1} \left[\mathbf{M} \left(\mathbf{C}_S^{\text{obs}} \right)^{-1} \mathbf{M}^T \right]^{-1}. \quad (2.41)$$

n_S is the number of mock samples used to estimate the sample covariance, and N_{pars} is the number of model parameters.

The important difference from Equation (2.39) is that the sample covariance estimate is more noisy than the semi-analytical model result. Therefore, non-linear matrix inversion operations cause significant biases. Paillas et al. (2023) considered this problem and gave a single (Hartlap-like) correction factor in their Eq. (B6), which we use here. The additional multiplier removes the inversion biases to leading order.

In Section 2.5.4 we additionally exclude the nuisance parameters. We leave out the corresponding rows and columns from the covariance matrices given by Equations (2.39) and (2.41). The remaining sub-matrix represents the covariance of the parameters of interest marginalized over the nuisance parameters.

2.4 Application to DESI-M2 LRG mocks

In this section, we use the described methods on DESI-M2 mocks to assess the performance and stability of the approach on the actual dataset. We describe the setup first, then perform intrinsic validation described in Section 2.3.2, look at the shot-noise rescaling values resulting from jackknife calibration used for the final covariance estimates, validate the RASCALC results by comparison with the mock sample covariance in measurement/observable and parameter space and finally focus specifically on errorbars on BAO scale.

2.4.1 Mock catalogs and reconstruction method

We use the 999 effective Zel’dovich (EZ) mocks (Chuang et al., 2015; Zhao et al., 2021) with cuts corresponding to the DESI LRG sample (Zhou et al., 2023) (described in more detail in Moon et al. (2023)), which will be referred to as **DESI-M2 Firstgen EZ mocks**. Sample covariance based on these does not provide a perfect reference because both the number of mock catalogs and the level of details in each simulation are limited, but the best one can have realistically, because increasing one without making the other worse would require even more significant computational resources. Comparing these is also robust to the mismatch between the data and mock clustering.

The reconstruction method is also the same as in Moon et al. (2023): the iterative procedure (Burden et al., 2015) implemented in the **IterativeFFTReconstruction** algorithm of the PYRECON package²³ with the **RecIso** convention. Three iterations are used with a Gaussian smoothing kernel of width $15h^{-1}\text{Mpc}$. An approximate growth rate and the expected bias are assumed.

2.4.2 Setup

For this study, we have performed separate runs using 2PCF measured from single LRG DESI-M2 **Firstgen EZ mocks** catalogs. This has been repeated 10 times for pre- and post-recon. In the latter case, individual shifted random catalogs have been used for each mock following the procedure we described in Section 2.2.3.

Pre-reconstruction galaxies and randoms were assigned unity weights. Post-reconstruction, FKP weights (Feldman et al., 1994) were used, given by

$$w_{\text{FKP}} = \frac{1}{1 + n(z)CP_0} \quad (2.42)$$

where $n(z)$ is the weighted number density (per volume), C is the mean completeness for the sample, and P_0 is a fiducial power-spectrum amplitude. For LRG, $C = 0.579$ and $P_0 = 10^4(h^{-1}\text{Mpc})^3$ (Moon et al., 2023). We note that the weighting schemes are not exactly

²³<https://github.com/cosmodesi/pyrecon>

the same as for real data, but since weights are included explicitly in the covariance estimators, we expect RASCALC to work with any fixed choice applied consistently for 2PCF measurements and Monte-Carlo integration.

For the importance sampling input, 10 random catalogs were used in pre-recon computations and 20 in post-recon, like for the RR (SS) pair counts computation for the 2PCF estimates. These randoms have been concatenated before being provided to RASCALC executable. We assign 60 jackknife regions assigned by a K -means subsampler based on data positions (but not weights) as in DESI-M2 data, compute the jackknife covariance matrix and use it to calibrate the shot-noise rescaling.

We note that some validation has been performed in Moon et al. (2023): RASCALC covariance matrix based on 2PCF averaged over all LRG DESI-M2 `Firstgen EZ mocks` with no shot-noise rescaling applied (and using unshifted randoms in the post-reconstruction case) has been compared to the sample covariance matrix in terms of χ^2 of BAO fits, best-fit values and standard deviations of BAO isotropic scale parameter α , yielding a good agreement. However, there are significant limitations to this approach:

- effect of noise in the input clustering is significantly smaller in 2PCF averaged over ≈ 1000 mocks than in the real data, which is close to a single mock catalog;
- possible differences in other parameters of the BAO model or more generic aspects of the correlation function have not been assessed.

Single-mock runs address the first issue since each of them is a fair proxy of the data. The use of covariance matrix comparison metrics from Section 2.3.1 expands on the second one. We keep the result with mock-average 2PCF, labeled “Average G” (Gaussian), to assess the importance of the precision of input clustering.

We also consider a shot-noise-rescaled version of the run with mock-averaged clustering, labeled “Average NG” (non-Gaussian). We note that calibration of an all-mocks run on a jackknife estimate can be ambiguous or require a repeated computation of all the pair counts with jackknives, which was not done before because jackknives are not necessary for the sample

covariance. Thus, we choose to fit the full covariance matrix to the mock sample covariance by minimizing the KL divergence between them (analogously to the jackknife procedure described in Section 2.2.1 and Equation (2.26)). Since only one parameter is varied in the fit, a perfect agreement is still not guaranteed. On the other hand, this setup is clearly idealized and would be closer to the closest possible match to the mock covariance the RASCALC method can provide. It comprises another useful reference to compare to the data-like performance on single mocks.

We only consider 45 radial bins, spanning $4 h^{-1}\text{Mpc}$ each from 20 to $200 h^{-1}\text{Mpc}$. The RASCALC covariances are produced with single angular bins, which is a simplifying assumption since treating the monopole more precisely in Legendre mode with shot-noise rescaling would require 2 runs per dataset. The 2PCF measurements for the mock sample covariance use a more precise monopole estimate provided by PYCORR²⁴. We also project them into a BAO model parameter space using the derivatives near the best fit (Fisher forecast). The model uses only a separation range from 48 to $148 h^{-1}\text{Mpc}$.

2.4.3 Internal convergence checks

We perform an intrinsic diagnostic procedure (as described in Section 2.3.2) to ensure that RASCALC integrals converged well in each run and exclude importance sampling random noise from significant error factors. We found that pre-recon mock 6 and post-recon mock 5 showed significantly worse consistency than all the rest. Therefore, we have run them twice longer.

After that, all the RASCALC results have reached a high and quite uniform level of internal consistency, as presented in Table 2.1. We only show R_{inv} , which are easier to interpret as the root-mean-square relative deviation between different partial estimates of the covariance matrix (considered over all directions in measurement space). χ_{red}^2 are limited via Equation (2.38), and D_{KL} values are quite close to estimates from Equation (2.32). Due to high consistency in measurement space, we have not performed a projection to parameter space here. The splitting of Monte-Carlo subsamples has been done in a few different ways and the non-symmetric

²⁴<https://github.com/cosmodesi/pycorr>

Mock no.	R_{inv} pre	R_{inv} post
Average G	1.3×10^{-3}	3.0×10^{-3}
Average NG	1.2×10^{-3}	8.1×10^{-3}
1	3.7×10^{-3}	2.3×10^{-3}
2	3.1×10^{-3}	2.2×10^{-3}
3	3.8×10^{-3}	2.1×10^{-3}
4	6.3×10^{-3}	1.9×10^{-3}
5	4.4×10^{-3}	1.3×10^{-3}
6	4.4×10^{-3}	1.9×10^{-3}
7	3.7×10^{-3}	2.0×10^{-3}
8	3.3×10^{-3}	1.7×10^{-3}
9	3.5×10^{-3}	2.5×10^{-3}
10	3.5×10^{-3}	2.2×10^{-3}

Table 2.1: Intrinsic pre- and post-reconstruction convergence test results in measurement space. R_{inv} estimate root-mean-square relative precision, averaged over all different directions in the measurement space. The numbers provided here are for the full covariance, but the consistency levels of the jackknife covariance prediction from RASCALC are similar. They demonstrate sub-percent stability in RASCALC integrals and ensure that random noise in importance sampling is not a significant source of error, as these numbers are smaller than deviations observed in further comparisons (Tables 2.3 to 2.6).

metric has been computed both ways ($\Psi_1 C_2$ and $\Psi_2 C_1$), but all the values were very close²⁵ and thus have been averaged to one number for each metric. These low (sub-percent) internal deviations give us confidence that the Monte-Carlo integration procedure in RASCALC has converged well and it will not be a significant error source in further comparison. After ascertaining this, we have not touched the covariance matrix products to be fair – with real survey and no mocks, other validation procedures described in this paper are not available.

2.4.4 Shot-noise rescaling values

Next, we look into the shot-noise rescaling values because the final covariance estimates (with approximate non-Gaussianity) are based on them. The shot-noise rescaling values for single mocks are obtained by fitting the separate RASCALC jackknife covariance prediction to the jackknife covariance estimate for each mock. For the mock-average clustering, the full RASCALC covariance was fit to the mock sample covariance instead, as discussed in

²⁵In the case of disjoint running (Section 2.2.2) there can be a meaningful difference between splittings within or between different random sub-catalogs, due to fluctuations in pair counts between these. But here all the randoms have been concatenated together.

Mock no.	Pre-recon α_{SN}	Post-recon α_{SN}
Average NG	1.096	1.038
1	1.096	1.062
2	1.074	1.043
3	1.040	1.034
4	1.077	1.051
5	1.079	1.033
6	1.089	1.030
7	1.080	1.041
8	1.102	1.058
9	1.116	1.033
10	1.080	1.041
1-10 mean $\pm$ std	1.083 ± 0.020	1.043 ± 0.011

Table 2.2: Shot-noise rescaling values for the 10 mocks, on which the final covariance predictions are based, pre- and post-reconstruction. The “Average NG” value is fit to mock sample covariance, and the 1-10 are fit to jackknife covariance estimates from single mocks, and the resulting numbers are consistent. The standard deviation for single mocks is $\approx 2\%$ before and $\approx 1\%$ after reconstruction, which translates into a similar effect on relative precision of the final covariance matrices due to rescaling.

Section 2.4.2.

The shot-noise rescaling values are gathered in Table 2.2. We note that all of them are greater than one (which corresponds to purely Gaussian covariance), in accordance with our expectation that the non-Gaussianity expands the errorbars. Moreover, the mean shot-noise rescaling of the 10 single mocks is ≈ 4 standard deviations larger than 1 both before and after reconstruction. The values obtained from the jackknife and mock covariance are consistent. After reconstruction, the shot-noise rescaling decreases for every mock. The pre-recon mean is larger than the post-recon one by ≈ 1.8 standard deviations. The scatter after reconstruction is also smaller than before. These deviations can be caused by the random fluctuations in the input 2PCF estimates, noise in jackknife covariances, and differences in shifted randoms (for post-recon only).

The key conclusion is that we have obtained the shot-noise rescaling parameter for a data-like setup (single mock runs) with a percent-level precision. This maps into a similar or smaller relative deviation in the rescaled covariance matrices since the 2-point term has the strongest scaling, $\propto \alpha_{\text{SN}}^2$, and the 4-point term remains the same (Equation (2.6)).

Mock no.	$D_{\text{KL}}(\Psi_R, \mathbf{C}_S)$	$R_{\text{inv}}(\Psi_R, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\Psi_R, \mathbf{C}_S)$
Average G	0.793	0.2922	1.1422
Average NG	0.537	0.2136	0.9906
1	0.721	0.2252	0.9600
2	0.602	0.2234	0.9992
3	0.571	0.2329	1.0477
4	0.548	0.2191	0.9991
5	0.755	0.2302	0.9830
6	0.727	0.2315	0.9723
7	0.695	0.2254	0.9675
8	0.530	0.2085	0.9695
9	0.610	0.2140	0.9291
10	0.518	0.2106	0.9895
1-10	0.628 ± 0.089	0.2221 ± 0.0087	0.982 ± 0.031
Perfect Ψ	0.519 ± 0.024	0.2147 ± 0.0049	1.0000 ± 0.0067

Table 2.3: Results of general measurement-space comparison between the RASCALC results (R) and mock sample covariance (S) before reconstruction. The last row provides the perfect-case reference – expectation values and standard deviations for the three metrics, if the RASCALC precision matrix truly described the distribution of the mock correlation functions.

2.4.5 Measurement-space validation

Now we proceed to comparison with the sample covariance matrices as reference, keeping in mind they are not devoid of noise so not all the comparison measures can be ideal. We consider the higher-dimensional space of observables first, where the effects of sample variance are quite significant. It consists of 45 bins of 2PCF, spanning 20–200 $h^{-1}\text{Mpc}$ linearly with a bin width of 4 $h^{-1}\text{Mpc}$.

Sample covariances for all original bins have been estimated using $n_S = 999$ 2PCF measurements from all the DESI-M2 **Firstgen EZ mocks** and the standard unbiased estimator (Equation (2.33)). The procedures have been similar for pre- and post-recon.

The comparison measures between the RASCALC precision matrices (estimated via Equation (2.9)) and the sample covariance matrices have been computed and are presented in Table 2.3 for pre-reconstruction and Table 2.4 for post-reconstruction. First of all, there are fluctuations in the comparison measures involving the single-mock results, stemming from the input 2PCF estimates, jackknife covariances, and differences in shifted randoms (for post-recon only) – the same causes as for scatter in α_{SN} discussed in Section 2.4.4. The

Mock no.	$D_{\text{KL}}(\Psi_R, \mathbf{C}_S)$	$R_{\text{inv}}(\Psi_R, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\Psi_R, \mathbf{C}_S)$
Average G	0.62	0.247	1.0653
Average NG	0.57	0.225	1.0028
1	0.94	0.256	0.9463
2	0.63	0.229	0.9725
3	0.72	0.266	1.0031
4	0.61	0.222	0.9657
5	0.62	0.229	0.9924
6	0.65	0.231	0.9817
7	0.62	0.226	0.9706
8	0.61	0.222	0.9631
9	0.77	0.281	1.0214
10	0.88	0.314	1.0162
1-10	0.70 ± 0.12	0.247 ± 0.031	0.983 ± 0.024
Perfect Ψ	0.519 ± 0.024	0.2147 ± 0.0049	1.0000 ± 0.0067

Table 2.4: Results of general measurement-space comparison between RASCALC results (R) and mock sample covariance (S) after reconstruction. The last row provides the perfect-case reference – expectation values and standard deviations for the three metrics, if the RASCALC precision matrix truly described the distribution of the mock correlation functions.

individual pre-reconstruction covariances appear to agree with the mock sample better than the post-reconstruction covariances. The covariance with mock-averaged clustering and no shot-noise rescaling, on the contrary, gives a closer agreement after reconstruction. Before reconstruction, any individual shot-noise rescaled covariance shows better agreement than the Gaussian mock-averaged clustering run; after reconstruction, it is very often worse. With mock-averaged clustering, the shot-noise rescaling is clearly beneficial for pre-reconstruction and less so for post-reconstruction. This may be a hint that the shot-noise rescaling might not be doing as well after reconstruction as before.

Compared to the perfect case, RASCALC typically performs worse (higher D_{KL} and R_{inv} , reduced chi-squared further from one), which we expect since the code (and the mocks) involve (different) approximations. We note that, on average, the comparison metrics are within a couple of standard deviations of the expectation value for the true underlying covariance matrix. On the other hand, it is the larger standard deviation in RASCALC results that is allowing this conclusion, and reducing the noise factors causing it (input 2PCF fluctuations, single jackknife covariance) may allow us to reach a closer agreement in future works.

2.4.6 Parameter-space validation

In this section, we project the covariance into a lower-dimensional and more physically meaningful space of BAO model parameters. Lower dimensionality makes the reference values for comparison metrics clearer, and the results become easier to interpret. In addition, the correlation function modes that are not physically possible or do not affect the parameter constraints are removed from consideration, which leaves only real and important “directions” for consideration.

We choose a commonly used BAO model²⁶ (Ross et al., 2017; Ata et al., 2018) with a scalable template ξ_0 and three nuisance polynomial terms:

$$\xi_{\text{mod}}(r) = B\xi_0(\alpha_{\text{BAO}}r) + A_0 + A_1/r + A_2/r^2, \quad (2.43)$$

comprising $N_{\text{pars}} = 5$ parameters: $B, A_0, A_1, A_2, \alpha_{\text{BAO}}$.

Instead of performing full fits, we use the Fisher matrix formalism. This can be seen as less precise than full fits on every mock or MCMC using a 2PCF likelihood. On the other hand, the parameter distribution is not Gaussian when the model is not a linear function of parameters, which makes linear approximation within the Fisher matrix formalism more suitable for the comparison methods we have discussed.

We estimate the parameter covariance matrix as the inverse of the Fisher matrix:

$$\mathbf{C}_S^{\text{par}} = \frac{n_S - 1}{n_S - N'_{\text{bins}} + N_{\text{pars}} - 1} \left[\mathbf{M}(\mathbf{C}'_S^{\text{meas}})^{-1} \mathbf{M}^T \right]^{-1} \quad (2.44)$$

where the measurement-space mock sample covariance matrix $\mathbf{C}'_S^{\text{meas}}$ is cut to the $N'_{\text{bins}} = 25$ bins spanning separations from 48 to 148 $h^{-1}\text{Mpc}$ used in BAO fits, and $\mathbf{M}$ is the matrix of derivatives of binned 2PCF vector $\boldsymbol{\xi}$ with respect to parameters $\mathbf{p}$:

$$M_{ca} \equiv \frac{\partial \xi_a}{\partial p_c}. \quad (2.45)$$

The derivatives have been taken at the best-fit parameters for the mock-averaged clustering measurements (separate before and after reconstruction).

²⁶https://github.com/cosmodesi/BAOfit_xs/

Mock no.	$D_{\text{KL}}(\Psi_R, \mathbf{C}_S)$	$R_{\text{inv}}(\Psi_R, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\Psi_R, \mathbf{C}_S)$
Average G	0.026	0.138	0.992
Average NG	0.027	0.135	0.925
1	0.104	0.236	0.855
2	0.065	0.193	0.911
3	0.022	0.122	0.941
4	0.013	0.099	0.962
5	0.127	0.250	0.883
6	0.091	0.232	0.866
7	0.095	0.229	0.861
8	0.025	0.128	0.918
9	0.042	0.165	0.876
10	0.016	0.107	0.944
1-10	0.060 ± 0.042	0.176 ± 0.059	0.902 ± 0.039
Perfect Ψ	0.0075 ± 0.0028	0.078 ± 0.014	1.000 ± 0.020

Table 2.5: Results of BAO parameter space $(B, A_0, A_1, A_2, \alpha_{\text{BAO}})$ comparison between RASCALC results (R) and mock sample covariance (S) before reconstruction. The last row provides the perfect-case reference – expectation values and standard deviations for the three metrics, if the RASCALC precision matrix truly described the distribution of the mock correlation functions.

Note that Equation (2.44) is scaled by a correction factor according to Eq. (B6) in Paillas et al. (2023) to account for biases caused by both matrix inversions. This provides an unbiased (although not noiseless) estimate of the true underlying covariance in parameter space as validated in Section A.2.4.

A similar but simpler procedure was performed with RASCALC products:

$$\Psi_{R,cd}^{\text{par}} \approx \mathbf{M}(\mathbf{C}_R^{\text{meas}})^{-1} \mathbf{M}^T, \quad (2.46)$$

where $\mathbf{C}_R^{\text{meas}}$ was also cut to the 25 bins spanning separations from 48 to 148 $h^{-1}\text{Mpc}$ used in BAO fits. There is a bias correction matrix $\mathbf{D}$ for RASCALC (Equation (2.9)), but for the results presented here, absolute values of its eigenvalues are $\lesssim 10^{-3}$; thus, we have decided to neglect this correction factor.

The comparison measures have been computed between the projected matrices and are presented in Table 2.5 for pre-recon and Table 2.6 for post-recon. Generally, lower expectation values of D_{KL} and R_{inv} for perfect precision make these numbers for RASCALC easier to interpret. There is a less apparent difference between pre- and post-recon. A notable exception

Mock no.	$D_{\text{KL}}(\Psi_R, \mathbf{C}_S)$	$R_{\text{inv}}(\Psi_R, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\Psi_R, \mathbf{C}_S)$
Average G	0.020	0.134	1.087
Average NG	0.014	0.113	1.049
1	0.122	0.252	0.907
2	0.036	0.171	1.002
3	0.041	0.202	1.082
4	0.015	0.106	0.981
5	0.030	0.149	0.999
6	0.029	0.144	0.974
7	0.021	0.125	0.981
8	0.014	0.105	0.977
9	0.095	0.329	1.154
10	0.096	0.327	1.155
1-10	0.050 ± 0.039	0.191 ± 0.085	1.021 ± 0.082
Perfect Ψ	0.0075 ± 0.0028	0.078 ± 0.014	1.000 ± 0.020

Table 2.6: Results of BAO parameter space ($B, A_0, A_1, A_2, \alpha_{\text{BAO}}$) comparison between RASCALC results (R) and mock sample covariance (S) after reconstruction. The last row provides the perfect-case reference – expectation values and standard deviations for the three metrics, if the RASCALC precision matrix truly described the distribution of the mock correlation functions.

is that all χ_{red}^2 for rescaled (with mock-averaged and single-mock clusterings) pre-recon are significantly less than 1 (meaning RASCALC “overestimates” the covariance then). In other cases, mimicking non-Gaussianity gives a slight improvement for mock-averaged clustering, but higher noise in 2PCF and jackknife covariance in single-mock estimates often drives the agreement with mock sample covariance worse than in the mock-averaged Gaussian estimate.

Overall, RASCALC single-mock results are within $< 2\sigma$ (dominated by the standard deviation of the perfect reference values, except the reduced chi-squared before reconstruction, which deviates by ≈ 2.2 std (combined)). However, the scatter in these numbers is quite significant (e.g. a few percent in root-mean-square relative error R_{inv}), and we should try to reduce it in future work.

2.4.7 Errorbars on BAO scale parameter

Since the scale parameter α_{BAO} is the important output of the current BAO analysis (Equation (2.43)), we have decided to extract its errorbar, marginalized over the other four parameters. This is quite trivial after the previous subsection – we only needed to invert the RASCALC

$\sigma(\alpha_{\text{BAO}})$	Pre-recon	Post-recon
Sample cov	0.01590 ± 0.00036	0.01459 ± 0.00033
Average G	0.01522	0.01360
Average NG	0.01596	0.01392
1	0.01616	0.01424
2	0.01598	0.01395
3	0.01583	0.01410
4	0.01609	0.01405
5	0.01595	0.01385
6	0.01613	0.01396
7	0.01594	0.01389
8	0.01621	0.01400
9	0.01652	0.01402
10	0.01618	0.01405
1-10 mean $\pm$ std	0.01610 ± 0.00019	0.01401 ± 0.00011

Table 2.7: Pre- and post-reconstruction errorbars on α_{BAO} from Fisher forecast. The mean of 10 single pre-reconstruction catalogs agrees with the sample covariance within a standard deviation. For post-reconstruction, the difference is within 2 standard deviations.

parameter-space precisions

$$\mathbf{C}_R^{\text{par}} = (\Psi_R^{\text{par}})^{-1} \quad (2.47)$$

neglecting the inversion bias, since it is expected to be even smaller than before with the smaller size of the matrices. Then we extract the marginalized errorbars from all the parameter covariances as

$$\sigma(\alpha_{\text{BAO}}) = \sqrt{C_{\alpha\alpha}^{\text{par}}}. \quad (2.48)$$

For the sample covariance, we expect the variance of $C_{S,\alpha\alpha}$ to nearly follow Equation (A.10):

$$\text{Var} [C_{S,\alpha\alpha}^{\text{par}}] \approx \frac{2}{n_S - 1} (C_{S,\alpha\alpha}^{\text{par}})^2 \quad (2.49)$$

and therefore the standard deviation of $\sigma(\alpha_{\text{BAO}})$ of

$$\sigma[\sigma(\alpha_{\text{BAO}})] \approx \frac{\sigma(\alpha_{\text{BAO}})}{\sqrt{2(n_S - 1)}}, \quad (2.50)$$

resulting in relative precision of 2.2%. This has been confirmed in Section A.2.4.

The resulting errorbar (Fisher) forecasts are provided in Table 2.7 and also presented as a scatter plot in Figure 2.3. We can notice that in any case, the post-recon precision is expected

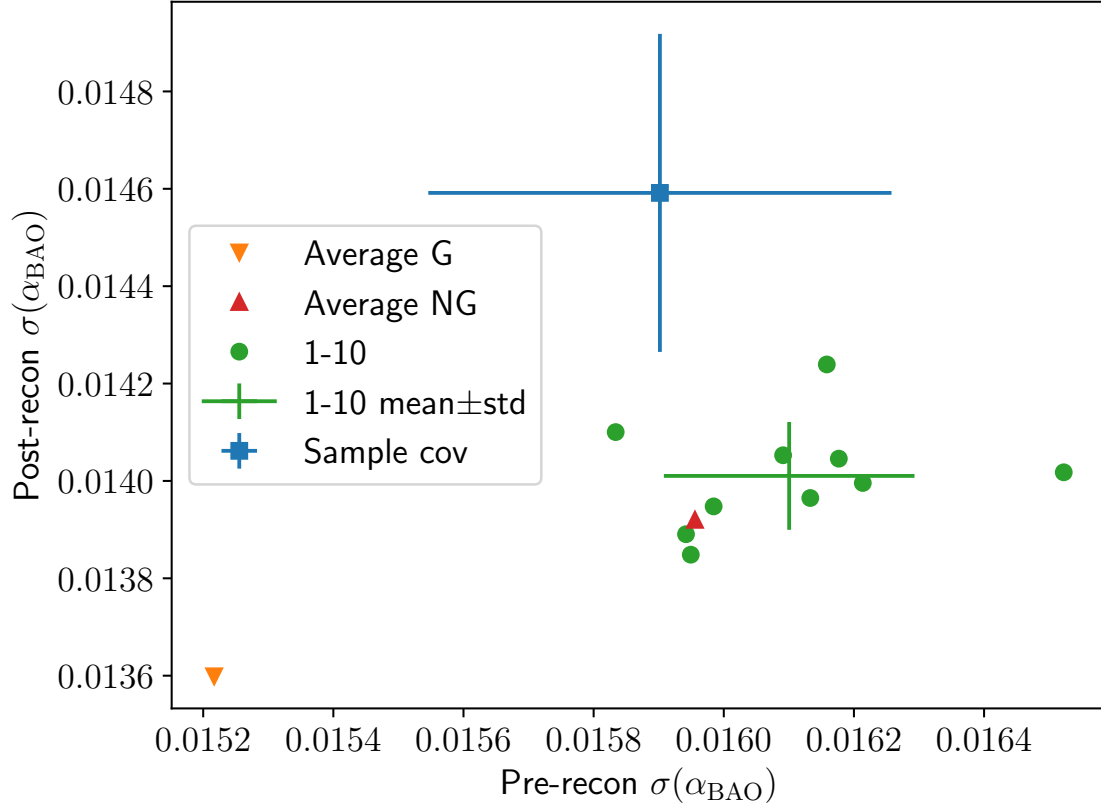


Figure 2.3: Pre- and post-reconstruction errorbars on α_{BAO} from Fisher forecast plotted against each other. Compared to the mock-averaged Gaussian run (Average G), the mock-averaged rescaled run (Average NG) and single-mock predictions with rescaling are closer to the sample covariance. The horizontal (pre-reconstruction) agreement is closer than the vertical (post-reconstruction) one, but even the latter falls within $\approx 2\sigma$. Note that the range of the axes is quite narrow, comprising only $\approx 9\%$ relative difference in errorbar before reconstruction and $\approx 7\%$ after.

to be higher than pre-recon. In both pre-recon and post-recon, $\sigma(\alpha_{\text{BAO}})$ in the mock-averaged clustering run without shot-noise rescaling are noticeably smaller than predicted from the sample covariance and are brought closer in the rescaled results. Mock-averaged clustering with fit shot noise and single mock runs gives very similar numbers. The key conclusion is that single-mock runs are in good agreement with the sample covariance on $\sigma(\alpha_{\text{BAO}})$, with a remarkably close match before reconstruction (just fractions of standard deviation) and a difference of ≈ 2 standard deviation after reconstruction. This gives assurance that data-based RASCALC covariances are on par with mock sample one for isotropic BAO fits.

2.5 Application to DESI DR1 mocks

2.5.1 DESI DR1 simulations and methods

In this section, we briefly describe the simulations we use and their processing steps.

Mocks

In this work, we mainly rely on effective Zel’dovich mocks (EZmocks) (Chuang et al., 2015; Zhao et al., 2021). We use the suite of 1000 catalogs representative of DESI DR1 (Zhao et al., 2025). These mocks are more approximate than those relying on full N -body simulations. However, they are fast enough to make a large number of simulations in $6\ h^{-1}\text{Gpc}$ boxes covering the whole volume of DESI DR1 without replications.

For shot-noise rescaling investigation (Section 2.5.3) we also used more realistic Abacus-2 mocks. They are based on the ABACUSSUMMIT suite of N -body simulations (Maksimova et al., 2021) produced with the ABACUS code (Garrison et al., 2021). The halos have been identified with the COMPASO halo finder (Hadzhiyska et al., 2022). The galaxy catalogs have been produced within halo occupancy distribution (HOD) formalism using the ABACUSHOD framework (Yuan et al., 2022). More details on the suite representing DESI DR1 can be found in DESI Collaboration et al. (2025f). These simulations have two disadvantages: only 25 realizations and smaller ($2\ h^{-1}\text{Gpc}$) boxes requiring replications to cover the DESI DR1

volume.

The factors of number and volume are crucial for covariance matrices. The number of mocks sets the relative precision of the sample covariance estimate, which is the only reference we have for comparison. By replications, we mean different parts of the final galaxy catalog made from the same part of the original simulation. These parts can become too strongly correlated and therefore bias the sample covariance estimate. This motivates our choice of **EZmocks** for the majority of this work.

Fiber assignment modeling

An important novel aspect of this work is the application of semi-analytical covariance matrices to mocks that incorporate a model of fiber assignment effects.

We mainly use the approximate algorithm called “fast fiber assign” or “fast fiber assignment” (FFA) (Hanif et al., 2025; Bianchi et al., 2024). It emulates the DESI fiber assignment algorithm using less computational resources (DESI Collaboration et al., 2025f).

For shot-noise rescaling investigation (Section 2.5.3) we also used the mock realizations of the DESI fiber assignment algorithm. They involve alternate merged target ledgers, thus the shortcut AMTL (Lasker et al., 2025). However, this method is prohibitively slow to process all 1000 **EZmocks** (DESI Collaboration et al., 2025f).

These fiber assignment algorithms vary in how they assign weights to galaxies and random points to mitigate incompleteness effects. As discussed in section 5.1 of Ross et al. (2025), the DESI assignment completeness can be split into two components, one of which can be modeled by assigning weights to randoms. However, the FFA algorithm only determines the total assignment completeness per galaxy. Consequently, the weights per galaxy resulting from FFA have greater variance than for AMTL mocks and the DESI DR1 LSS catalogs. As we mentioned earlier, different weighting methods for data and random points can affect the shot-noise rescaling values (Equation (2.4)).

Effects of fiber assignment may pose additional challenges to the method because it involves anisotropic pair-wise sampling, depending both on the density of the targets and

the number of survey passes in the region (Bianchi et al., 2024). We provide the RASCALC code with the random catalog and clustering estimate affected by fiber assignment (see the flowchart in Figure 2.1), but the expansion leading to Equation (2.7) (and Equation (2.13)) uses the survey-wide correlation function(s) to calculate the ensemble averages of products of overdensities, and the shot-noise rescaling (Equation (2.4)) is also global. It is challenging to let the correlation function or shot-noise rescaling vary across the survey (with on-sky position or redshift) without complicating the covariance matrix model too much and introducing too many parameters.

Fiber assignment incompleteness might also cause issues with the jackknife. Ideally, we would like each sub-region to have a distribution of the number of passes representative of the full survey. This is challenging to achieve, and the current jackknife assignment does not guarantee that.

Reconstruction

We apply the semi-analytical covariance matrix estimators before and after standard BAO reconstruction (pre- and post-recon). The reconstruction methodology follows the findings of the DESI DR1 optimal reconstruction task force (Chen et al., 2024a; Paillas et al., 2025): the `RecSym` mode of the `IterativeFFTReconstruction` algorithm (Burden et al., 2015) from the `PYRECON` package²⁷ with smoothing scale of $15 h^{-1}\text{Mpc}$.

Theoretical modeling and fitting

Our first analysis of interest is DESI 2024 baryon acoustic oscillations (DESI Collaboration et al., 2025f). We use the same anisotropic (2D) model with BAO power spectrum template and spline-based broadband terms (Chen et al., 2024b). The fit uses monopole and quadrupole in radial bins spanning $s = 48 - 152 h^{-1}\text{Mpc}$.

The second analysis of interest is DESI 2024 full-shape (DESI Collaboration et al., 2024b). This presents more difficulties because it primarily uses power spectra and the methodology

²⁷<https://github.com/cosmodesi/pyrecon>

has not been standardized for correlation functions. Nevertheless, we use similar models relying on the **velocileptors** Lagrangian perturbation theory model (Chen et al., 2020, 2021; Maus et al., 2025b) with maximum freedom and standard prior basis. Two variants correspond to the power spectrum template choice. One approach is “ShapeFit”. It is a compression method using parametric variations of a single power spectrum template evaluated at the reference cosmology (Brieden et al., 2021). The other is “full modeling”, meaning a linear power spectrum from **CLASS** Boltzmann code (Blas et al., 2011). The fit uses monopole, quadrupole and hexadecapole in radial bins spanning $s = 28 - 152 h^{-1}\text{Mpc}$.

These analyses inform our work in two ways.

First, the two fit ranges set our choice of sub-matrices of correlation function multipoles covariances for comparison. In other words, we restrict the covariance matrices to the multipoles and radial bins used in the fitting. This gives us an informed choice for smaller sub-matrices to see which parts of the covariance matrix are captured better by the semi-analytic method.

Second, we project the covariance matrices into the parameter spaces of these three models. We do this with the inverse Fisher matrix as described in Section 2.3.3. It is very important to see that the errorbars on physical parameters are predicted reliably through the RASCALC covariances, although this analysis may not generalize to alternative models.

We use the **DESILIKE** package²⁸ for the evaluation of all three models (including polynomial emulators for the full-shape models) and fitting the data.

2.5.2 Setup for semi-analytical covariance matrices

We consider three galaxy types (tracers) in three different redshift ranges. We use luminous red galaxies (LRG) with $z = 0.8 - 1.1$, emission line galaxies (ELG) (Raichoor et al., 2023) with $z = 1.1 - 1.6$ and magnitude-limited bright galaxy survey (BGS) (Hahn et al., 2023) with $z = 0.1 - 0.4$. These correspond to **LRG3**, **ELG2** and **BGS** samples in the main DESI BAO papers (DESI Collaboration et al., 2025f,e) respectively. The summary of tracers and redshift bins is provided in Table 2.8.

²⁸<https://github.com/cosmodesi/desilike>

Tracer	LRG	ELG	BGS
z range	(0.8, 1.1)	(1.1, 1.6)	(0.1, 0.4)
Designation in DESI Collaboration et al. (2024a, 2025f)	LRG3	ELG2	BGS
EZmocks snapshot z	1.1	1.325	0.2

Table 2.8: Tracers and redshift bins used in this chapter. For LRG and ELG, which have multiple bins unlike BGS, we have selected the densest ones, as shot-noise seems easier to capture with RASCALC. We did not include the quasars (QSO) for the same reason. The snapshot redshifts were used to construct the power spectrum templates.

We apply the data pipeline²⁹ (Figure 2.1) to single mock catalogs, as was done in previous works. This implies using the random files, full and jackknife correlation function estimates specific to that catalog. We use 10^{30} different realizations of EZmocks to quantify the fluctuations in the semi-analytical results due to realistic random variations in the input quantities.

We process the North and South Galactic Caps (NGC and SGC) separately. DESI DR1 data has been processed in the same manner. This allows different shot-noise rescaling values, reflecting different completeness patterns in these parts of DESI DR1. Then we combine the two covariances into a single matrix for the full survey, assuming the Galactic Caps are uncorrelated³¹ (Section A.3).

We create the covariance matrices for monopole, quadrupole and hexadecapole in 45 radial bins between 20 and 200 h^{-1} Mpc (each 4 h^{-1} Mpc wide). We exclude the $s < 20 h^{-1}$ Mpc bins because they impede the convergence of the covariance matrices. Moreover, we expect the shot-noise rescaling to become inadequate on small scales.

²⁹The data scripts are available at https://github.com/cosmodesi/RascalC-scripts/tree/DESI2024/DESI/Y1_pre and [post](https://github.com/cosmodesi/RascalC-scripts/tree/DESI2024/DESI/Y1_post) directories for single-tracer covariances before and after reconstruction, respectively. Our analogous single-mock scripts (with only minor differences) are in https://github.com/cosmodesi/RascalC-scripts/tree/DESI2024/DESI/Y1/EZmocks/single_pre and [post](https://github.com/cosmodesi/RascalC-scripts/tree/DESI2024/DESI/Y1/EZmocks/single_post) folders.

³⁰The number is a compromise between getting good statistics and saving computing time.

³¹In the EZmocks NGC and SGC are constructed from separate realizations Zhao et al. (2025), so they are independent. As a result, we do not validate this assumption for the data.

2.5.3 Intrinsic tests of the method

Having described the setup for the current RASCALC application to EZmocks, we detail the quality checks we perform before comparing the results with the mock sample covariance.

Intrinsic convergence and computation time

We have two specific convergence criteria for each RASCALC computation. The first is the eigenvalue test performed by the code on both full and jackknife covariance matrix terms (Equations (2.20) and (2.24)). The desired condition is that the minimal eigenvalue of the 4-point term is larger than minus the minimal eigenvalue of the 2-point term (Philcox et al., 2020). The second is the positive definiteness of the final covariance matrix estimate. We do not use the covariance matrix model that fails either of these tests.

In addition, we have a quantitative measure of convergence without strict thresholds. As in Section 2.4.3, we use the R_{inv} comparison measure (Equation (2.27)) between the different estimates of each RASCALC covariance matrix from separate halves of the Monte-Carlo integration samples. We found several high outliers compared to other mock realizations of the same tracer.

We repeat the computation on the mocks with the convergence issues mentioned before. Then we post-process the data from the second computation and the data combined from the two computations. We choose the one that first satisfies the two strict criteria and then gives a lower R_{inv} value.

After this, we find that LRG covariance matrices reached $R_{\text{inv}} \leq 2.0\%$ convergence within 4 node-hours (512 core-hours³²) on the NERSC Perlmutter supercomputer. The ELG covariance matrices reached $R_{\text{inv}} \leq 2.9\%$ convergence within 10 node-hours (1280 core-hours). BGS reached $R_{\text{inv}} < 11.5\%$ convergence within 12 node-hours (1536 core-hours).

The computations have become longer relative to Section 2.4.3, whereas R_{inv} convergence figures have worsened (i.e. increased). In this previous work, the maximum R_{inv} was 0.63%.

³²Hereafter the figure is for the single computation, i.e., effectively twice longer in a few cases. We count physical cores and not hyperthreads. Later profiling also showed a possibility of $\approx 2\times$ improvement by using 64 threads and physical cores instead of 128.

The number of 4-point configurations contributing to the sums (Equations (2.20) and (2.24)) for LRG NGC or SGC in this work (2.4×10^{12}) is very close to the analogous number for full LRG in Section 2.4.3 (2.5×10^{12}), where the sums (Equations (2.7) and (2.13)) were evaluated for a single angular bin.

The processing of each 4-point configuration becomes longer in Legendre mode. This can be expected because, as mentioned earlier, a given configuration contributes only to one pair of angular bins, but to all Legendre multipoles. Other factors, like CPU differences, parallelism efficiency, and a larger number of 3- and 2-point configurations in this work, may also affect the runtimes.

The increase of the R_{inv} convergence measure can be primarily attributed to three times more observables in the covariance matrix. In this section, we have the same number of radial bins as in Section 2.4, but three multipoles instead of only the monopole. A higher-dimensional covariance matrix requires more configurations sampled for the same relative precision. Additionally, ELG and especially BGS are more challenging due to the increasing importance of the 4-point term relative to the 3- and 2-point terms. The more points, the more configurations need to be sampled for the same precision of the term. Still, the R_{inv} we obtained are significantly smaller than the expected deviation of the mock sample covariance estimate from the true covariance matrix ($R_{\text{inv}} \approx 37\%^{33}$).

Shot-noise rescaling values

Our method uses a rescaling of the shot noise contribution to account for differences between the true small-scale contributions and our Gaussian approximation, as we pointed out in Section 2.2.1. After reaching a relatively uniform convergence level in the previous section, we should investigate the values of the shot-noise rescaling parameter.

Table 2.9 shows the shot-noise rescaling values obtained for mocks and DESI DR1 data according to the fiducial procedure (Figure 2.1). Interestingly, we find that all the shot-noise

³³As we show later in Table 2.12, this number is based on 1000 mock realizations. For a fixed number of samples, the relative precision of the sample covariance also worsens with the number of observables.

α_{SN}	NGC		SGC	
	Mocks	Data	Mocks	Data
LRG pre-recon	0.743 ± 0.012	0.865	0.7935 ± 0.0081	0.953
LRG post-recon	0.770 ± 0.010	0.836	0.809 ± 0.011	0.969
ELG pre-recon	0.3757 ± 0.0043	0.687	0.4018 ± 0.0014	0.718
ELG post-recon	0.3789 ± 0.0044	0.693	0.4051 ± 0.0044	0.737
BGS pre-recon	0.792 ± 0.012	0.855	0.8198 ± 0.0091	0.897
BGS post-recon	0.812 ± 0.012	0.872	0.8447 ± 0.0095	0.934

Table 2.9: Shot-noise rescaling values α_{SN} for mocks (10 realizations of FFA EZmocks) and data.

α_{SN}	Data	AMTL (correct)		FFA (approximate)		Complete
		ABACUS	EZmocks	ABACUS	EZmocks	ABACUS
LRG NGC	0.865	0.845	0.886	0.721	0.758	0.934
LRG SGC	0.953	0.954	0.961	0.781	0.784	0.962
ELG NGC	0.687	0.649	0.672	0.378	0.380	0.965
ELG SGC	0.718	0.707	0.744	0.405	0.403	0.968

Table 2.10: Shot-noise rescaling values α_{SN} (before reconstruction) for the data and two different types of mocks (Section 2.5.1) with two different fiber assignment models (Section 2.5.1). “Complete” designates mocks before fiber assignment. We use a single mock realization for each category.

rescaling values are smaller than 1; i.e., the jackknife variations are smaller than predicted by the Gaussian approximation with standard Poisson shot noise. Previous mock studies (O’Connell et al., 2016; O’Connell & Eisenstein, 2019; Philcox et al., 2020, our Section 2.4.4), on the contrary, obtained shot-noise rescaling values greater than 1.

We also see that the shot-noise rescaling values are significantly lower for mocks than for the data. The difference is most pronounced for ELGs, which have the lowest shot-noise rescaling values for both mocks and data. ELGs are also impacted most by the fiber incompleteness effect due to their lower priority compared to other dark-time targets, LRG and QSO Raichoor et al. (2023). This suggests that the low shot-noise rescaling is related to fiber assignment.

To test whether fiber assignment is the main cause for low and different shot-noise rescaling values, we have performed additional RASCALC computations using more realistic ABACUS mocks and the DESI fiber assignment algorithm (AMTL). We only used one realization for LRG and ELG in each case to save computing time.

Table 2.10 shows the shot-noise rescaling values for data and different mocks (ABACUS or

α_{SN}	NGC		SGC	
	Jackknife	Mock sample	Jackknife	Mock sample
LRG pre	0.743 ± 0.012	0.7417 ± 0.0038	0.7935 ± 0.0081	0.7906 ± 0.0045
LRG post	0.770 ± 0.010	0.7446 ± 0.0029	0.809 ± 0.011	0.7945 ± 0.0032
ELG pre	0.3757 ± 0.0043	0.3757 ± 0.0013	0.4018 ± 0.0014	0.4077 ± 0.0016
ELG post	0.3789 ± 0.0044	0.3751 ± 0.0013	0.4051 ± 0.0044	0.4067 ± 0.0017
BGS pre	0.792 ± 0.012	0.7916 ± 0.0068	0.8198 ± 0.0091	0.827 ± 0.014
BGS post	0.812 ± 0.012	0.8148 ± 0.0070	0.8447 ± 0.0095	0.844 ± 0.013

Table 2.11: Shot-noise rescaling values for the single-mock runs calibrated on jackknife (Figure 2.1) and mock sample covariances (Figure 2.2).

EZmocks) with different fiber assignment models (FFA or AMTL). We also include an ABACUS mock before fiber assignment (“complete”). We see that the fiber assignment modeling method makes a bigger difference than the type of mocks. The approximate fast fiber assignment gives the lowest shot-noise rescaling values. The DESI fiber assignment algorithm applied to the mocks (AMTL) gives α_{SN} very similar to the data. The complete ABACUS mocks have larger shot-noise rescaling values than data and fiber-assigned mocks. This confirms the fiber assignment as a key factor affecting the α_{SN} parameter.

The discrepancy in shot-noise rescaling values also raises concerns about the quality of approximations in the FFA algorithm. However, of all the mock types, only **EZmocks** processed with FFA are numerous enough for a precise sample covariance estimate. Therefore, we continue using them in the remainder of this paper.

We performed the final consistency tests of the shot-noise optimization procedure in light of the concerns about fiber assignment and jackknife discussed in Section 2.5.1. We optimized the shot-noise rescaling based on mock sample covariance (the process is shown schematically in Figure 2.2). We show the resulting values along with the baseline, jackknife-based ones in Table 2.11, and find them very close for all cases. In other words, calibration of shot-noise rescaling on jackknives still brings us close to an optimal fit on the mocks. With that, we have decided to proceed further with the validation process using the jackknife-calibrated shot-noise rescaling values. This will show how close this nearly optimal fit is to the mock sample covariance.

2.5.4 Comparison between semi-analytical and mock sample covariance matrices

Thus, we reach the last validation part: comparison of the semi-analytical covariances obtained from single mock catalogs with the mock sample covariance matrices. We compute three similarity measures (Section 2.3.1) between these covariance matrices for each tracer before and after standard BAO reconstruction (pre- and post-recon). Instead of presenting 10 numbers (corresponding to each RASCALC realization) for each case in the paper, we provide their mean and standard deviation. The full set of comparison measure values is available in the supplementary material³⁴. We also provide the “perfect” reference row, listing the statistical properties of the similarity measures between the true covariance and a mock sample covariance matrix based on their dimension (also obtained in Section 2.3.1).

We note that the standard deviation in the “perfect” row is different and independent from the others. The “perfect” standard deviation characterizes the distribution of the random difference between the sample covariance estimate from 1000 realizations and the true covariance. In each other row (tracer + pre- or post-recon combination), the sample covariance estimate is fixed, and the standard deviation describes the scatter resulting from 10 different RASCALC covariance realizations.

Ideally, every RASCALC covariance matrix would be consistent with the true covariance. To see whether this is the case, we compute the difference between each comparison measure for each RASCALC covariance realization and the corresponding “perfect” mean in “perfect” standard deviations. We summarize the 10 resulting quantities by the mean and standard deviation as well, and provide them below the summary of the similarity measure itself. This sets the common structure for all tables in this section (i.e., Tables 2.12 to 2.17.).

Observables

We begin with the covariances for correlation function multipoles, in other words, in the space of observables.

³⁴[doi:10.5281/zenodo.10895161](https://doi.org/10.5281/zenodo.10895161)

	$D_{\text{KL}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$R_{\text{inv}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\mathbf{C}_R^{-1}, \mathbf{C}_S)$
Perfect	4.817 ± 0.070	0.3690 ± 0.0031	1.0000 ± 0.0039
LRG pre-recon	4.856 ± 0.029 $(0.55 \pm 0.41)\sigma$	0.3678 ± 0.0049 $(-0.4 \pm 1.6)\sigma$	0.989 ± 0.016 $(-2.7 \pm 4.1)\sigma$
LRG post-recon	4.977 ± 0.050 $(2.27 \pm 0.70)\sigma$	0.3581 ± 0.0034 $(-3.5 \pm 1.1)\sigma$	0.957 ± 0.014 $(-11.3 \pm 3.6)\sigma$
ELG pre-recon	4.811 ± 0.024 $(-0.09 \pm 0.34)\sigma$	0.3700 ± 0.0055 $(0.3 \pm 1.8)\sigma$	1.000 ± 0.015 $(0.1 \pm 3.8)\sigma$
ELG post-recon	5.001 ± 0.018 $(2.61 \pm 0.26)\sigma$	0.3701 ± 0.0042 $(0.4 \pm 1.4)\sigma$	0.986 ± 0.012 $(-3.6 \pm 3.1)\sigma$
BGS pre-recon	5.129 ± 0.052 $(4.43 \pm 0.73)\sigma$	0.3824 ± 0.0081 $(4.4 \pm 2.6)\sigma$	0.997 ± 0.016 $(-0.8 \pm 4.2)\sigma$
BGS post-recon	5.177 ± 0.077 $(5.1 \pm 1.1)\sigma$	0.3810 ± 0.0079 $(3.9 \pm 2.6)\sigma$	0.994 ± 0.016 $(-1.4 \pm 4.1)\sigma$

Table 2.12: Summary of full observable-space comparison of RASCALC covariances with the sample covariances (135 bins, $s = 20 - 200 \ h^{-1}\text{Mpc}$, monopole, quadrupole and hexadecapole).

Table 2.12 shows the comparison measures for the full covariance matrices, covering $s = 20 - 200 \ h^{-1}\text{Mpc}$ radial bins for all three multipoles. We can see that some RASCALC results deviate significantly from the “perfect” (i.e., the true covariance behavior). We remind that the KL divergence and R_{inv} accumulate deviations in all “directions”. The KL divergences exceed the ideal expectation value by nearly 3σ for LRG and ELG post-recon, whereas for BGS they are even further from perfect. R_{inv} values are high with a larger scatter for BGS. In the reduced chi-squared, which captures the overall “scaling” with higher accuracy, the mean values for the RASCALC runs are shifted significantly for LRG and ELG post-recon, and the mock-to-mock scatter is high in all the cases.

We continue the comparisons in Table 2.13, now cutting the range to $s = 28 - 152 \ h^{-1}\text{Mpc}$ as is common for full-shape fits (ShapeFit and direct). The KL divergences and R_{inv} become more consistent with the perfect reference cases for LRG and ELG, but remain high for BGS. The scaling difference (in reduced chi-squared) remains similar.

We perform the final set of observable-space comparisons in Table 2.14, further restricting the range to $s = 48 - 152 \ h^{-1}\text{Mpc}$ and using only monopole and quadrupole, without hexadecapole. We do not see significant consistency changes from the previous case.

We note that the abovementioned differences are relatively small. In the reduced chi-

	$D_{\text{KL}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$R_{\text{inv}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\mathbf{C}_R^{-1}, \mathbf{C}_S)$
Perfect	2.260 ± 0.049	0.3067 ± 0.0036	1.0000 ± 0.0046
LRG pre-recon	2.307 ± 0.022 $(0.99 \pm 0.46)\sigma$	0.3056 ± 0.0037 $(-0.3 \pm 1.0)\sigma$	0.983 ± 0.016 $(-3.7 \pm 3.4)\sigma$
LRG post-recon	2.333 ± 0.027 $(1.52 \pm 0.56)\sigma$	0.2998 ± 0.0024 $(-1.93 \pm 0.68)\sigma$	0.960 ± 0.014 $(-8.7 \pm 3.0)\sigma$
ELG pre-recon	2.2578 ± 0.0095 $(-0.04 \pm 0.20)\sigma$	0.3050 ± 0.0044 $(-0.5 \pm 1.2)\sigma$	0.995 ± 0.014 $(-1.0 \pm 3.1)\sigma$
ELG post-recon	2.292 ± 0.013 $(0.67 \pm 0.28)\sigma$	0.3044 ± 0.0033 $(-0.65 \pm 0.93)\sigma$	0.987 ± 0.012 $(-2.8 \pm 2.6)\sigma$
BGS pre-recon	2.414 ± 0.025 $(3.18 \pm 0.52)\sigma$	0.3140 ± 0.0066 $(2.0 \pm 1.9)\sigma$	0.987 ± 0.016 $(-2.7 \pm 3.4)\sigma$
BGS post-recon	2.479 ± 0.038 $(4.52 \pm 0.78)\sigma$	0.3202 ± 0.0065 $(3.8 \pm 1.8)\sigma$	0.993 ± 0.016 $(-1.5 \pm 3.4)\sigma$

Table 2.13: Summary of observable-space comparison of RASCALC covariances with the sample covariances restricted to the range of ShapeFit and full modeling fit (93 bins, $s = 28 - 152 \ h^{-1}\text{Mpc}$, monopole, quadrupole and hexadecapole).

squared, they are at most $(4.3 \pm 1.4)\%$ for LRG post-recon in the widest range, and in R_{inv} – no more than a percent or two on top of $23 - 37\%$ caused by the finite sample size. We should ask whether we trust the realism of the mocks to such a high level in all aspects of the correlation function multipoles. Moreover, for the real survey, matching the clustering between data and simulations will become an additional issue for mocks. On the other hand, our RASCALC computations use the correlation function measured directly from single mock catalogs similar to real data. Thus, we conclude that the semi-analytic method’s performance is very compelling.

Model parameters

We proceed to project the covariance matrices (as described in Section 2.3.3) to the parameters of the models listed in Section 2.5.1.

We use the same model Jacobian (Equation (2.40)) for projecting RASCALC covariances (Equation (2.39)) and the mock sample covariances (Equation (2.41)) into the parameter space. We compute the partial derivatives at the best fit of each model to the mean clustering of all the available mocks for each galaxy type. Since BAO reconstruction changes clustering,

	$D_{\text{KL}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$R_{\text{inv}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\mathbf{C}_R^{-1}, \mathbf{C}_S)$
Perfect	0.702 ± 0.027	0.2303 ± 0.0046	1.0000 ± 0.0062
LRG pre-recon	0.763 ± 0.014 $(2.29 \pm 0.54)\sigma$	0.2351 ± 0.0023 $(1.04 \pm 0.49)\sigma$	0.982 ± 0.016 $(-2.9 \pm 2.6)\sigma$
LRG post-recon	0.732 ± 0.013 $(1.14 \pm 0.50)\sigma$	0.2281 ± 0.0019 $(-0.48 \pm 0.40)\sigma$	0.964 ± 0.015 $(-5.7 \pm 2.3)\sigma$
ELG pre-recon	0.7195 ± 0.0083 $(0.65 \pm 0.31)\sigma$	0.2317 ± 0.0039 $(0.30 \pm 0.86)\sigma$	0.999 ± 0.015 $(-0.2 \pm 2.4)\sigma$
ELG post-recon	0.6903 ± 0.0088 $(-0.45 \pm 0.33)\sigma$	0.2278 ± 0.0029 $(-0.54 \pm 0.62)\sigma$	0.995 ± 0.012 $(-0.8 \pm 1.9)\sigma$
BGS pre-recon	0.796 ± 0.021 $(3.54 \pm 0.78)\sigma$	0.2419 ± 0.0079 $(2.5 \pm 1.7)\sigma$	0.982 ± 0.018 $(-2.8 \pm 2.8)\sigma$
BGS post-recon	0.777 ± 0.031 $(2.8 \pm 1.2)\sigma$	0.2462 ± 0.0090 $(3.4 \pm 1.9)\sigma$	1.011 ± 0.017 $(1.8 \pm 2.7)\sigma$

Table 2.14: Summary of observable-space comparison of RASCALC covariances with the sample covariances restricted to the range of BAO fits (52 bins, $s = 48 - 152 \ h^{-1}\text{Mpc}$, monopole and quadrupole).

we compute separate Jacobians before and after reconstruction. We use the mock sample covariance matrix in this fit. Using different Jacobian matrices for each covariance matrix would complicate the comparison.

A more thorough investigation of the covariance matrix’s impact on the model fits is presented in the companion paper (Forero-Sánchez et al., 2024). They perform fits to each mock clustering using different covariance matrices (mock sample and semi-analytical) and compare the resulting best parameter values as well as errorbar estimates. On the flip side, with such a detailed approach Forero-Sánchez et al. (2024) can not test multiple RASCALC single-mock covariances.

BAO

In Table 2.15 we compare the covariances projected to the BAO parameters, the scaling parameters α_{iso} and α_{AP} ³⁵. The comparison measures look consistent with the perfect reference case. The most significant deviations are seen in BGS pre-recon: both the KL divergence and

³⁵Because we use Fisher matrix formalism, the covariance matrices projected for $\alpha_{\parallel}$ and $\alpha_{\perp}$ should have the same comparison measures.

	$D_{\text{KL}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$R_{\text{inv}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\mathbf{C}_R^{-1}, \mathbf{C}_S)$
Perfect	0.0015 ± 0.0012	0.051 ± 0.021	1.000 ± 0.032
LRG pre-recon	0.00103 ± 0.00066 $(-0.39 \pm 0.53)\sigma$	0.042 ± 0.014 $(-0.39 \pm 0.67)\sigma$	0.960 ± 0.014 $(-1.27 \pm 0.45)\sigma$
LRG post-recon	0.00110 ± 0.00026 $(-0.32 \pm 0.21)\sigma$	0.0458 ± 0.0051 $(-0.23 \pm 0.24)\sigma$	0.9734 ± 0.0098 $(-0.84 \pm 0.31)\sigma$
ELG pre-recon	0.00069 ± 0.00036 $(-0.66 \pm 0.29)\sigma$	0.0367 ± 0.0093 $(-0.65 \pm 0.44)\sigma$	1.033 ± 0.010 $(1.04 \pm 0.32)\sigma$
ELG post-recon	0.00078 ± 0.00045 $(-0.59 \pm 0.37)\sigma$	0.037 ± 0.012 $(-0.63 \pm 0.55)\sigma$	0.967 ± 0.012 $(-1.05 \pm 0.38)\sigma$
BGS pre-recon	0.0042 ± 0.0016 $(2.2 \pm 1.3)\sigma$	0.088 ± 0.015 $(1.74 \pm 0.71)\sigma$	0.915 ± 0.014 $(-2.68 \pm 0.46)\sigma$
BGS post-recon	0.00027 ± 0.00024 $(-1.00 \pm 0.19)\sigma$	0.0213 ± 0.0093 $(-1.38 \pm 0.44)\sigma$	0.992 ± 0.013 $(-0.24 \pm 0.40)\sigma$

Table 2.15: Summary of parameter-space comparison of RASCALC covariances with the sample covariances projected to the BAO fit parameters, α_{iso} and α_{AP} .

R_{inv} are high, while the reduced chi-squared is almost 3 sigma low on average (meaning that the RASCALC covariance is “larger” than the mock sample one). The mock-to-mock scatter in RASCALC results is always lower than the noise expected from the finite mock sample size.

We have also plotted the errorbars on α_{iso} and α_{AP} against each other in Figure 2.4. They corroborate Table 2.15. Almost all deviations are within the 99.7% contour. The only exceptions are two BGS pre-recon realizations. However, we note that these plots do not show the covariance of the parameters, which is taken into account with KL divergence and R_{inv} .

Full shape: ShapeFit and full modeling

Table 2.16 shows the comparison measures for the covariances projected to the ShapeFit parameters: α_{iso} , α_{AP} , dm and df . We do not see significant statistical deviations from the perfect case. This is the only case when BGS (both pre- and post-recon) are not particularly far from the reference. The mock-to-mock scatter in RASCALC results is smaller than or comparable with the noise in the sample covariances.

Additionally, in Figure 2.5 we compare the projected errorbars for the scaling parameters, α_{iso} and α_{AP} . All the points fall within the 99.7% confidence region in this 2D space, although the other parameters and cross-correlations are ignored in this picture, unlike in Table 2.16.

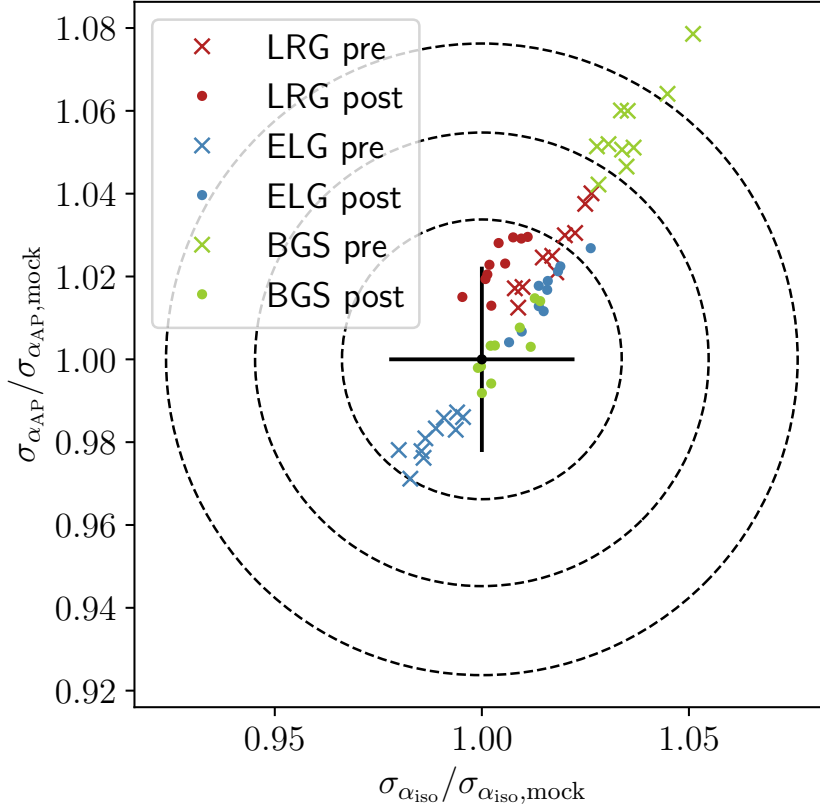


Figure 2.4: Comparison of the projected errorbars for the BAO scale parameters, normalized to the values obtained from the mock sample covariance. The cross shows one-dimensional relative precisions $((2(n_S - 1))^{-1/2} \approx 2.2\%$, Equation (2.50)) following from the **EZmocks** sample size ($n_S = 1000$), and the dashed ellipses approximately correspond to two-dimensional 68%, 95% and 99.7% confidence regions in this 2D space of errorbars. Here, the correlation of errorbars is ignored; it varies in different cases but is too small to notice ($\lesssim 0.04$).

Finally, in Table 2.17 we provide the comparison results for the covariances projected to the full modeling parameters, h , ω_{cdm} , ω_b and $\log A_s$. We see deviations exceeding 3 sigma in most of the measures for LRG pre-recon and BGS; LRG post-recon and ELG look consistent. Still, the differences we see are at a few percent level, and may be caused by the limited realism of the mocks.

	$D_{\text{KL}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$R_{\text{inv}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\mathbf{C}_R^{-1}, \mathbf{C}_S)$
Perfect	0.0050 ± 0.0023	0.069 ± 0.016	1.000 ± 0.022
LRG pre-recon	0.0059 ± 0.0023 $(0.4 \pm 1.0)\sigma$	0.073 ± 0.013 $(0.25 \pm 0.81)\sigma$	0.969 ± 0.020 $(-1.39 \pm 0.89)\sigma$
LRG post-recon	0.00739 ± 0.00087 $(1.06 \pm 0.39)\sigma$	0.0840 ± 0.0052 $(0.95 \pm 0.33)\sigma$	0.984 ± 0.022 $(-0.73 \pm 0.98)\sigma$
ELG pre-recon	0.0063 ± 0.0029 $(0.6 \pm 1.3)\sigma$	0.078 ± 0.015 $(0.59 \pm 0.94)\sigma$	1.005 ± 0.027 $(0.2 \pm 1.2)\sigma$
ELG post-recon	0.0029 ± 0.0013 $(-0.95 \pm 0.58)\sigma$	0.052 ± 0.011 $(-1.06 \pm 0.68)\sigma$	0.999 ± 0.017 $(-0.06 \pm 0.78)\sigma$
BGS pre-recon	0.0071 ± 0.0030 $(0.9 \pm 1.3)\sigma$	0.083 ± 0.019 $(0.9 \pm 1.2)\sigma$	0.996 ± 0.031 $(-0.2 \pm 1.4)\sigma$
BGS post-recon	0.0067 ± 0.0027 $(0.7 \pm 1.2)\sigma$	0.080 ± 0.017 $(0.7 \pm 1.1)\sigma$	0.992 ± 0.030 $(-0.3 \pm 1.3)\sigma$

Table 2.16: Summary of parameter-space comparison of RASCALC covariances with the sample covariances projected to the ShapeFit parameters: α_{iso} , α_{AP} , dm and df .

2.6 Conclusions

We present and validate the DESI DR1 pipeline for the semi-analytical covariance matrices of the galaxy 2-point correlation functions on realistic mock catalogs, including a model of fiber assignment.

We develop a streamlined procedure for the estimation of the semi-analytical covariance matrices for Legendre moments of the 2PCF in separation bins with the RASCALC code (Philcox et al., 2020). The previous implementation (Philcox & Eisenstein, 2019) required an additional computation with angular bins to mimic the non-Gaussian effects by calibrating the shot-noise rescaling value on the jackknife covariance matrix estimate. Now we can perform this calibration together with the construction of the covariance model for Legendre moments. This allowed for more efficient massive production of covariance matrices for all the tracers, redshift bins, and galactic caps of DESI DR1 galaxies and quasars data.

Importantly, we reconsidered the methods for covariance matrix comparison, paying great attention to their meaning, interpretation, and noise stemming from mock sample variance. We have also discussed the implications of split random-random counts computation and made a slight modification to the formalism to cover the reconstructed 2PCF estimates.

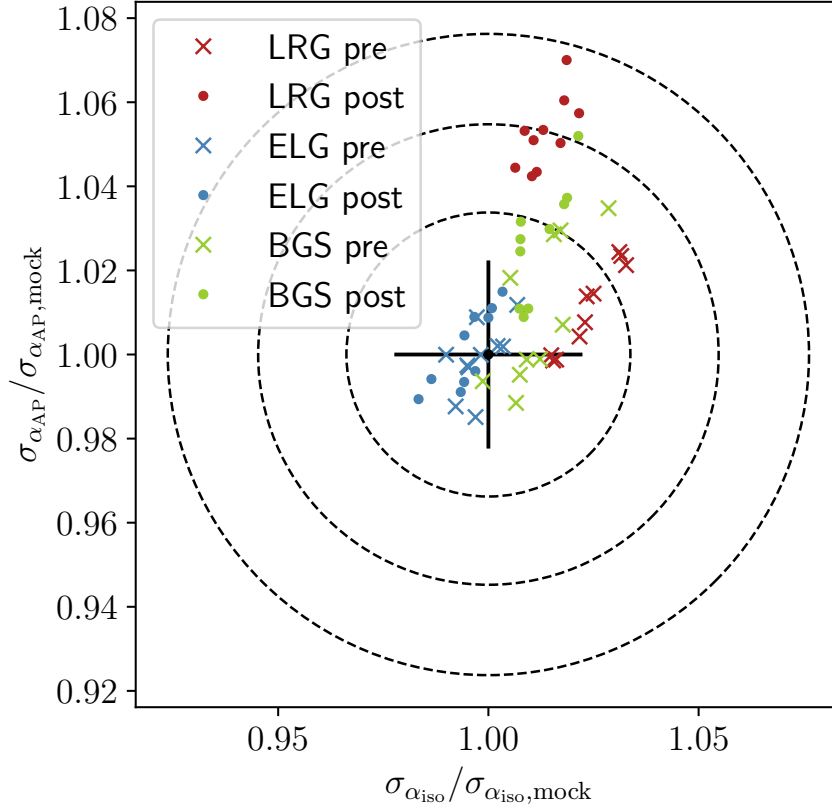


Figure 2.5: Same as Figure 2.4 but with errorbars for the scale parameters following from ShapeFit.

We have applied the selected approaches to the validation of RASCALC on single DESI-M2 Firstgen EZ mock catalogs (using their individual clustering measurements and shifted random catalogs after reconstruction), each representing a reasonable proxy for DESI-M2 data, by comparison with full mock sample covariance. We find a close agreement (maximum deviation $\approx 2.2\sigma$) with a perfect case, although much of this deviation is due to scatter in RASCALC results. The preceding discussion about the interpretations of the metrics, focusing on a smaller number of observables and even fewer parameters, allowed us to obtain a clearer quantitative assessment of the precision and accuracy of RASCALC results than in previous works. One should keep in mind that the mocks are approximate, and this can partially account for the imperfection of the match with the reference statistics.

Focusing on the errorbar of the BAO scale, we found a very close, percent-level agreement

	$D_{\text{KL}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$R_{\text{inv}}(\mathbf{C}_R^{-1}, \mathbf{C}_S)$	$\chi_{\text{red}}^2(\mathbf{C}_R^{-1}, \mathbf{C}_S)$
Perfect	0.0050 ± 0.0023	0.069 ± 0.016	1.000 ± 0.022
LRG pre-recon	0.0099 ± 0.0034 $(2.2 \pm 1.5)\sigma$	0.093 ± 0.014 $(1.50 \pm 0.89)\sigma$	0.951 ± 0.020 $(-2.20 \pm 0.91)\sigma$
LRG post-recon	0.00360 ± 0.00068 $(-0.63 \pm 0.30)\sigma$	0.0584 ± 0.0055 $(-0.66 \pm 0.35)\sigma$	0.976 ± 0.016 $(-1.08 \pm 0.71)\sigma$
ELG pre-recon	0.0061 ± 0.0013 $(0.48 \pm 0.60)\sigma$	0.0783 ± 0.0094 $(0.59 \pm 0.59)\sigma$	1.013 ± 0.024 $(0.6 \pm 1.1)\sigma$
ELG post-recon	0.0033 ± 0.0014 $(-0.75 \pm 0.60)\sigma$	0.055 ± 0.011 $(-0.85 \pm 0.72)\sigma$	0.985 ± 0.016 $(-0.68 \pm 0.72)\sigma$
BGS pre-recon	0.0101 ± 0.0073 $(2.3 \pm 3.2)\sigma$	0.099 ± 0.040 $(1.9 \pm 2.5)\sigma$	1.013 ± 0.035 $(0.6 \pm 1.5)\sigma$
BGS post-recon	0.0083 ± 0.0045 $(1.5 \pm 2.0)\sigma$	0.090 ± 0.027 $(1.3 \pm 1.7)\sigma$	1.003 ± 0.035 $(0.1 \pm 1.5)\sigma$

Table 2.17: Summary of parameter-space comparison of RASCALC covariances with the sample covariances projected to the full modeling parameters: h , ω_{cdm} , ω_b and $\log A_s$.

with the sample covariance from mocks. It is on par with the accuracy that a set of ≈ 1000 simulations can provide. The number of available mocks thus limits the precision of the validation at the current level.

In the full measurement space before reconstruction, using the shot-noise rescaling is particularly clearly beneficial compared to the pure Gaussian estimate, even with less noisy (mock sample average) input clustering. The discrepancies and fluctuations are likely to be impacted by the precision of correlation function estimates from data, which will improve with its size in the future. Further validations with larger mocks, corresponding to a year and/or full five years of DESI data, will follow.

Then, we apply the updated pipeline to more advanced mock catalogs with fast (approximate) fiber assignment, representative of DESI DR1. We cover 3 selected tracers and redshift bins (LRG3, ELG2 and BGS according to DESI Collaboration et al. (2025f)), without and with standard BAO reconstruction applied. We use a single mock catalog and repeat the procedure for 10 different realizations to assess the impact of realistic fluctuations in the input quantities.

First, we note the difference between the shot-noise rescaling values obtained for the mocks with fast fiber assignment and the data in Section 2.5.3. We show that the fiber assignment modeling (and the associated weighting scheme) is the main factor causing the difference.

Then we find the parameter calibration on jackknife and mock sample covariance yields very close results in the mock runs. This motivates us to proceed with the comparison of the covariance matrices.

Then, we apply the set of compact measures of covariance matrix similarity from Section 2.3.1. We use covariances in the observable space (correlation function multipoles) in Section 2.5.4, as well as projected linearly (through Fisher matrix formalism) to the parameters of different models in Section 2.5.4. We find some deviations that cannot be explained solely by the finite sample size limiting the accuracy of the mock-based covariance. However, these differences are at a few percent level.

We argue that this level of agreement is sufficient for real-world applications. First, mocks necessarily involve approximations to make sufficiently many catalogs. Second, simulations are never perfectly representative of data in terms of clustering due to the limited precision of the measurements. Blinding can aggravate this issue by forcing the mock tuning to rely on earlier, smaller samples with larger uncertainties in their clustering. Additionally, some simulations, including **EZmocks** (DESI Collaboration et al., 2025f; Zhao et al., 2025), are only matched to the 1- and 2-point statistics of the data, whereas the covariance matrices are impacted by 3- and 4-point functions. Therefore, matching the mocks perfectly is not necessarily a reasonable goal.

Our results for the BAO model (Section 2.5.4) are particularly important because RASCALC covariance matrices are used in the DESI DR1 baryon acoustic oscillations analysis (DESI Collaboration et al., 2025f). The errorbars of the scale parameters (α_{iso} and α_{AP}) predicted from RASCALC agree to $\leq 8\%$ with the mock sample covariance. The standard deviation expected from the sample covariance of 1000 mocks itself is $\approx 2.4\%$. When we exclude BGS before reconstruction (not used for the main DESI DR1 BAO), the errorbars agree within $\approx 5\%$. Therefore, we report a close match between the semi-analytic and mock sample covariance.

We find the covariance matrices for the BGS sample less consistent in most comparisons³⁶. We expected them to be more challenging for RASCALC because of a higher number density and

³⁶A notable exception is ShapeFit parameters in Section 2.5.4.

thus higher significance of the 4-point term compared to the 3- and 2-point terms. This already caused slower convergence and could further demonstrate the limitations of the shot-noise rescaling. However, BGS has been challenging for the **EZmocks** as well (Zhao et al., 2025)³⁷, so their sample covariance is likely to be a less robust reference than for LRG and ELG.

The observable-space results (Section 2.5.4) may leave an impression that RASCALC performance worsened with fiber assignment³⁸. However, we think the change of interpretation primarily causes this. The previous RASCALC validation for early DESI data (Section 2.4) used an earlier version of **EZmocks** without any fiber assignment effects. The covariance matrices there also showed statistically significant variations from the sample covariance in observable space. They were deemed acceptable as comparable to the scatter in semi-analytical results. In this work, we use a stricter interpretation, testing whether every RASCALC single-mock result is consistent with the perfect-case reference. In future work, we could repeat the tests on the mocks before fiber assignment with all the tracers (and multipoles) used in this paper.

One perspective improvement is the development of alternatives to shot-noise rescaling within the more generic semi-analytic configuration-space formalism. The usage of fully empirical higher-point functions is likely to be not viable, due to a significantly higher number of bins and accordingly lower signal-to-noise. Precise theoretical modeling of non-Gaussian correlation functions is also very challenging. Instead, we might include a basic prescription for non-Gaussian covariance contribution inferred from a set of detailed simulations, or use approximate expressions for higher-point functions like $\zeta(\mathbf{r}_1, \mathbf{r}_2) = Q[\xi(r_1)\xi(r_2) + \xi(r_1)\xi(|\mathbf{r}_1 - \mathbf{r}_2|) + \xi(r_2)\xi(|\mathbf{r}_1 - \mathbf{r}_2|)]$ motivated by hierarchical models (Peebles & Groth, 1975), and possibly a similar structure for the 4PCF. This might provide better accuracy than rescaling the Gaussian terms while keeping the number of parameters low and thus still allowing us to fit them to a reference (e.g., jackknife) covariance. On the other hand, the aforementioned 3PCF prescription is known to be far from exact with constant Q (Takada & Jain, 2003), and the computations may suffer from slower convergence due to additional large values of

³⁷Moreover, DESI BGS **EZmocks** had not been used in Moon et al. (2023).

³⁸Or due to including not only monopole, but also quadrupole and hexadecapole.

small-scale 2PCF compared to Gaussian parts.

Another direction for further improvement is to study the dependence of shot-noise rescaling on fiber assignment incompleteness. This could be achieved by running RASCALC on survey sub-regions with different completeness patterns, set by the number of passes. It is also possible that a prescription for higher-point non-Gaussian correlations with a low number of parameters would give a better consistency with the reference than rescaling the covariance matrix terms, in which they are nulled. However, for such precision studies, it is instructive to have extremely reliable, realistic and numerous mocks.

The semi-analytic approach can extend beyond the standard 2-point function. First, the cross-covariances of different tracers are provided in Section A.1. The full cross-covariance has more bins and requires more mocks than a single covariance for validation at similar precision. Consistent simulated catalogs for different tracers are also crucial for capturing realistic cross-correlations. Second, Philcox & Eisenstein (2020a) have introduced the covariance matrices for the modified power spectra, including the cross-covariance with correlation functions. Third, Philcox & Eisenstein (2019) have derived the covariance matrices for isotropic 3-point correlation functions. It involves higher-order correlation functions up to 6 points. Approximating all of them in fast mocks is challenging. Thus, these extensions require extra work, particularly on the mock side for validation.

In summary, we have confirmed RASCALC semi-analytical covariance matrices for 2PCF as a very viable alternative to the mock-based ones. Despite the increase in the runtime of the RASCALC code (from 100-300 core-hours in Section 2.4 to 500-1500 core-hours in Section 2.5.3³⁹), the method is far faster than calibrating, generating, and processing a suite of mocks numerous enough to give an adequate covariance matrix precision. We have shown that the two methods produce similar results given the requirements of the DESI 2024 BAO analysis. The speed advantage of the semi-analytic method permits easier exploration of situations where one cannot afford to regenerate mock catalogs, such as different assumptions

³⁹As we remarked before, the later profiling results suggests that our covariance matrices could potentially be generated using ≈ 2 times less computational resources.

about cosmology, galaxy-halo connection, or non-standard sample selections. We therefore expect that such semi-analytic methods can be of broad utility for computing large-scale covariance matrices in wide-field surveys.

Acknowledgments

We thank Daniel Forero-Sánchez, Arnaud de Mattia, Nikhil Padmanabhan, Hee-Jong Seo, Ashley Ross, Violeta Gonzalez-Perez, Will Percival, Oliver Philcox, and Martin White for useful comments and fruitful discussions. MR and DJE are supported by the U.S. Department of Energy (DOE) grant DE-SC0007881 and by the Simons Foundation Investigator program. DFS acknowledges support from the Swiss National Science Foundation (SNF) "Cosmology with 3D Maps of the Universe" research grant, 200020_175751 and 200020_207379. H-JS acknowledges support from the U.S. Department of Energy, Office of Science, Office of High Energy Physics under grant No. DE-SC0023241. H-JS also acknowledges support from Lawrence Berkeley National Laboratory and the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123 during the sabbatical visit.

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of High-Energy Physics, under Contract No. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: <https://www.desi.lbl.gov/collaborating-institutions>. Any opinions, findings, and conclusions or recommendations expressed in this material are

those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation, the U. S. Department of Energy, or any of the listed funding agencies.

The authors are honored to be permitted to conduct scientific research on Iolkam Du’ag (Kitt Peak), a mountain with particular significance to the Tohono O’odham Nation.

This work has used the following software packages: `ASTROPY` (Astropy Collaboration et al., 2013, 2018, 2022), `JUPYTER` (Perez & Granger, 2007; Kluyver et al., 2016), `MATPLOTLIB` (Hunter, 2007), `NUMPY` (Harris et al., 2020), `PYCORR` (De Mattia et al., 2024; Sinha & Garrison, 2020, 2019), `PYTHON` (Van Rossum & Drake, 2009), `SCIPY` (Virtanen et al., 2020; Gommers et al., 2023), and `SCIKIT-LEARN` (Pedregosa et al., 2011; Buitinck et al., 2013; Grisel et al., 2023).

This research has made use of NASA’s Astrophysics Data System. Software citation information has been aggregated using **The Software Citation Station** (Wagg & Broekgaarden, 2024; Wagg et al., 2024).

Chapter 3

Key DESI BAO results

3.1 Introduction

In this chapter, we provide a summary of major Dark Energy Spectroscopic Instrument results that have relied on the semi-analytical covariance matrices (Chapter 2) and many more supporting studies. This review will be far from complete and comprehensive, and is influenced by our personal preferences. E.g., we omitted the neutrino mass problem (see e.g. Elbers et al., 2025) because its fair discussion would include much more than BAO. For the fuller picture, we encourage reading the referenced DESI papers.

Baryon acoustic oscillation (BAO) measurements constrain the expansion of the Universe between now and the emergence of the CMB. Thus, BAO are very important for understanding the Hubble tension between local measurements and CMB inference. The expansion history in late times is also the most informative about dark energy evolution because the fraction of dark energy in the total energy budget of the Universe is expected to become low in the far past.

DESI is the most advanced BAO experiment in scientific operation since 2021. It was designed to measure galaxy redshifts at an unprecedented speed. During its 5-year program, DESI is scheduled to collect more than 30 million galaxy and quasar spectra, an order of magnitude more than the previous state-of-the-art, SDSS.

3.2 Early DESI data (2023)

The first DESI BAO detection (Moon et al., 2023) from the Early DESI data¹ was a crucial part of the validation of DESI’s scientific program (DESI Collaboration et al., 2024d). It confirmed data quality and basic processing choices, abstaining from any cosmological implications. We found that only the first 2 months of the DESI main survey data already give measurement precision and detection level (shown in Figure 3.1) comparable to SDSS-III (DR12) data collected over 6 years (Ross et al., 2017). This confirmed that DESI was on track to reach the planned BAO precision at the end of its full 5-year program.

As we mentioned in Chapter 2, our specific contribution has been covariance matrices for the correlation function measurements, crucial for the resulting precision estimates. For early DESI data, we identified a mismatch in the clustering signal between the simulations and the data, which biased the mock-based covariance. Using more flexible semi-analytic covariances allowed us to fix this discrepancy at a low cost.

3.3 DESI DR1 (2024)

3.3.1 BAO measurements

To ensure solid cosmological interpretation, the DESI DR1 (2024) galaxy and quasar BAO analysis (DESI Collaboration et al., 2025f) involved a thorough re-evaluation of methodology and known systematics. There largely correspond to a suite of supporting papers on: theoretical modeling (Chen et al., 2024b), optimal reconstruction (Paillas et al., 2025; Chen et al., 2024a), optimal combination of overlapping tracers (Valcin et al., 2025), analytical (Alves et al., 2025) and semi-analytical (Chapter 2, Rashkovetskyi et al., 2025b) covariance matrix validation (Forero-Sánchez et al., 2024), fiducial cosmology (Pérez-Fernández et al., 2025) and galaxy-halo connection (Mena-Fernández et al., 2025; Garcia-Quintero et al., 2025) assumptions.

¹The name and some descriptions may suggest that Early DESI data is included in DESI Early Data Release (DESI Collaboration et al., 2024e), but this is not true. The survey validation data comprising the Early Data Release is insufficient for a robust BAO analysis. Instead, the early DESI data represents the first two months of the main survey and is only part of Data Release 1 (DESI Collaboration et al., 2025c).

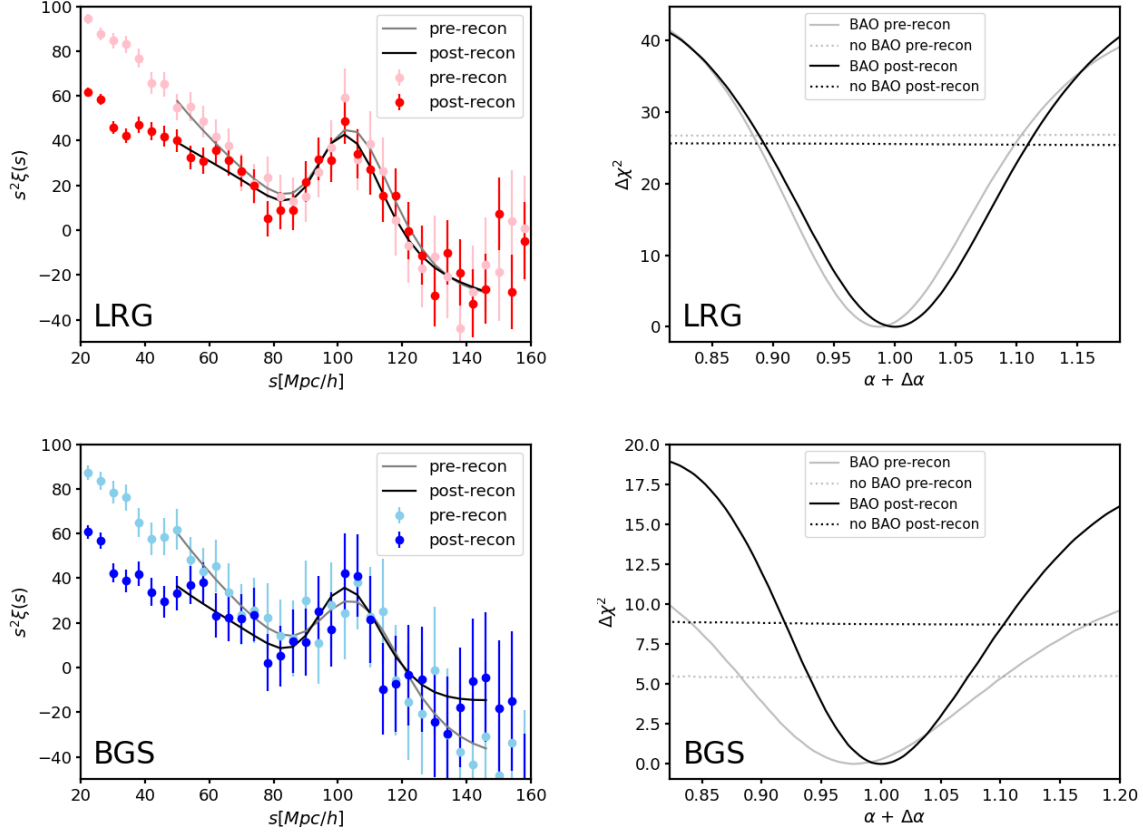


Figure 3.1: BAO in the Early DESI data (from the first two months of the main DESI survey, DESI-M2; figure reproduced from Moon et al., 2023). Left: measurements and best-fit models before and after standard BAO reconstruction (pre- and post-recon). Right: chi-squared differences relative to the best BAO fit in each case. In the DESI-M2 LRG sample ($0.4 < z < 1.1$), the precision of the (isotropic) scale parameter is 1.7% and the detection is $\approx 5\sigma$; in the BGS sample ($0.1 < z < 0.5$) — 2.6% and $\approx 3\sigma$ respectively. Compare LRG to BOSS CMASS ($0.43 < z < 0.7$) precision of 1.6% (DR9, Anderson et al., 2012), 1.0% (DR12, Ross et al., 2017) and the 0.77% aggregate precision for LRG from both BOSS and eBOSS (Alam et al., 2021); and BGS — to BOSS LOWZ ($0.15 < z < 0.43$) precision of 1.7% (DR12, Ross et al., 2017). The scale parameter α was shifted by a small value in the spirit of blinding and to discourage premature cosmological inference before the full one-year analysis.

Many of these supporting studies, as well as the clustering catalog iterations (DESI Collaboration et al., 2024a) and the validation of the blinding technique (Andrade et al., 2025b) intended to prevent confirmation bias, benefited heavily from our fast, flexible and realistic semi-analytic covariance matrices. If mock-based covariance matrices were the only option, fully consistent tests of quick corrections, alternative assumptions or differently selected galaxy samples would cost much more time and effort and thus would likely remain unavailable.

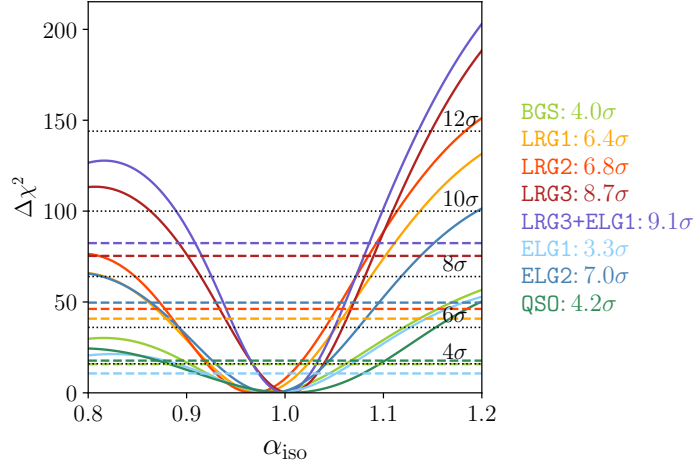


Figure 3.2: Detection significance of the BAO features in various DESI DR1 galaxy tracers (figure reproduced from DESI Collaboration et al., 2025f). Compare with Early DESI Data (Figure 3.1).

We briefly show the impressive detection significances with $\Delta\chi^2$ profiles for DESI DR1 galaxy and quasar BAO in Figure 3.2 before shifting attention to the corresponding distance measurements in Figure 3.3. The latter also include the Lyman- α forest BAO results from DESI Collaboration et al. (2025d). Their aggregated precision surpassed the previous state-of-the-art by a factor of 1.2². At this accuracy level, hints of deviations from the standard model of cosmology emerged.

3.3.2 Cosmology: dynamic dark energy

DESI Collaboration et al. (2025e) presented implications of DESI DR1 BAO measurements

²Although some individual bins (particularly LRG $0.6 < z < 0.8$, LRG2) have a lower precision than optimal BOSS and eBOSS combination (DESI Collaboration et al., 2025f; Alam et al., 2021).

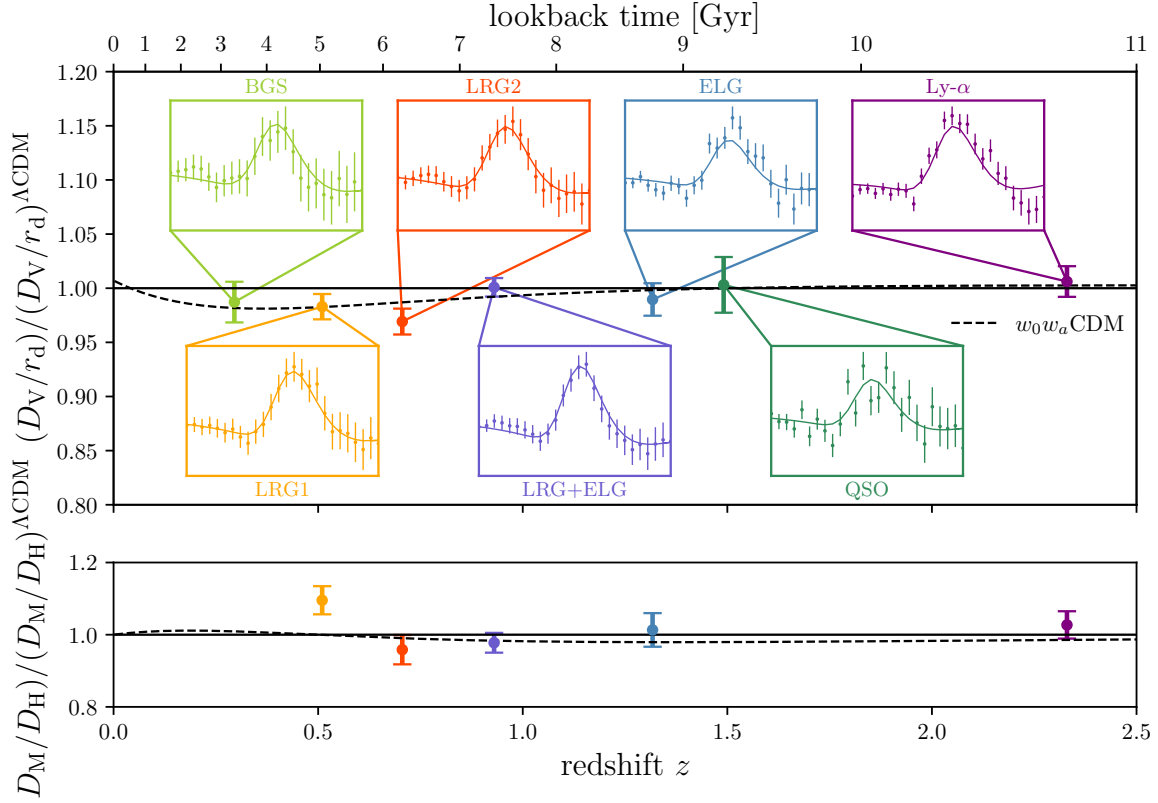


Figure 3.3: Summary of DESI DR1 BAO measurements (figure by Arnaud de Mattia based on DESI Collaboration et al., 2025f,d,e). All of the errorbars on this plot, except for Ly- α , are derived from the RASCALC semi-analytic covariance matrices (Chapter 2). Improved precision allows us to see deviations from the standard model (ΛCDM , solid line). An alternative dark energy model ($w_0w_a\text{CDM}$, dashed line) is preferred.

for selected cosmological models. We choose to highlight the emerging possibility of variable dark energy (Figure 3.4), in a way similar to recent supernova analyses (Brout et al., 2022; Rubin et al., 2023; DES Collaboration et al., 2024). This could become the first confirmed extension to the standard cosmological model, Λ CDM.

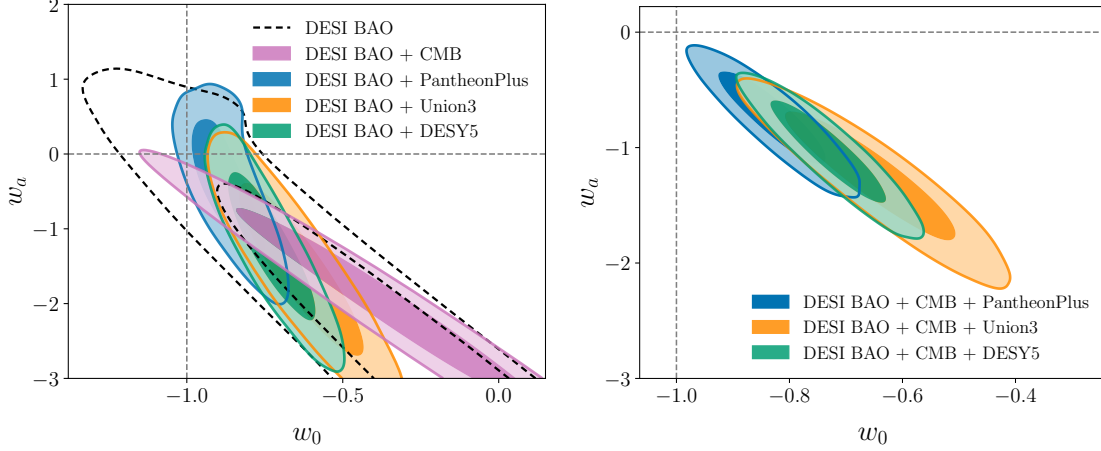


Figure 3.4: $w_0 w_a$ CDM model implications of DESI DR1 data in combination with external datasets (figures reproduced from DESI Collaboration et al., 2025e). External datasets are CMB (a combination of *Planck* PR3 primary temperature and polarization anisotropy power spectra (Planck Collaboration et al., 2020b) and lensing power spectra from *Planck* PR4 (Carron et al., 2022) and ACT DR6 (Qu et al., 2024)) and Type Ia supernovae: PantheonPlus (Brout et al., 2022), Union3 (Rubin et al., 2023) and DES Y5 (DES Collaboration et al., 2024). DESI DR1 BAO with CMB and different Type Ia supernovae datasets indicate a 2.5σ , 3.5σ or 3.9σ deviation from Λ , the constant dark energy ($w_0 = -1, w_a = 0$), whereas removing any of them reduces the significance to $\lesssim 2\sigma$.

This analysis assumes the Chevallier-Polarski-Linder (CPL) parameterization of the dark energy equation of state (Chevallier & Polarski, 2001; Linder, 2003):

$$w_{\text{DE}}(a) = w_0 + w_a \cdot (1 - a). \quad (3.1)$$

The equation of state parameter w drives the density evolution³ via $\rho \propto a^{-3(1+w)} = (1+z)^{3(1+w)}$, a is the scale factor of the Universe and z is cosmological redshift. This parametrization is phenomenological, but it can closely reproduce a wide variety of physically

³Alternatively, w is the ratio of pressure to energy density in a linear equation of state. Assuming no interactions with other components of the Universe, this sets the energy density evolution. But with interactions, the two definitions can differ, then the logarithmic derivative of the density can be called the effective equation of state parameter.

motivated models (Linder, 2003; De Putter & Linder, 2008). The constant dark energy (or cosmological constant Λ) have $w_{\text{DE}} = -1$ and thus $w_0 = -1, w_a = 0$.

Interestingly, DESI DR1 BAO alone and with external datasets (Figure 3.4) align with a line in w_0, w_a space going through zero with a negative inclination. Lodha et al. (2025b) found that it does not correspond to well-known models motivated in fundamental physics, but closely aligns with the “mirage” dark energy predicted by Linder (2007) based on cosmological measurements. Moreover, it implies phantom dark energy ($w_{\text{DE}} < -1$) in earlier cosmic times, which is deemed challenging to model (e.g., in Cort  s & Liddle, 2024) or even unphysical (e.g., in Dinda & Maartens, 2025) due to its violating the null energy condition, $\rho + p \geq 0$. Notably, interactions of dark energy with other components of the Universe (e.g., dark matter in Giar   et al., 2024) may lead to an increase in dark energy density (“effective” $w_{\text{DE}} < -1$) while $\rho + p \geq 0$ holds. However, the exact description and fundamental motivation of such interactions may be challenging. Therefore, the nature of dynamic dark energy seems to remain mysterious.

We do not go into much more detail about DESI DR1 results here because the updated DR2 BAO results are now also available.

3.4 DESI DR2 (2025)

3.4.1 BAO measurements

We have been closely involved in the galaxy and quasar BAO analysis of the DESI Data Release 2 (DR2, DESI Collaboration et al., 2026), comprised of approximately 3 years of the main survey. This latest BAO analysis (DESI Collaboration et al., 2025a; Andrade et al., 2025a) has been largely similar to the DR1 analysis (DESI Collaboration et al., 2025f), reusing the results of its many supporting studies. However, one of the important differences is that the previously performed tests of semi-analytical and analytical covariances (crucially, RASCALC in Chapter 2) allowed us to save considerable time and effort by not waiting for the preparation and calibration of approximate mock suites with a large number of realizations.

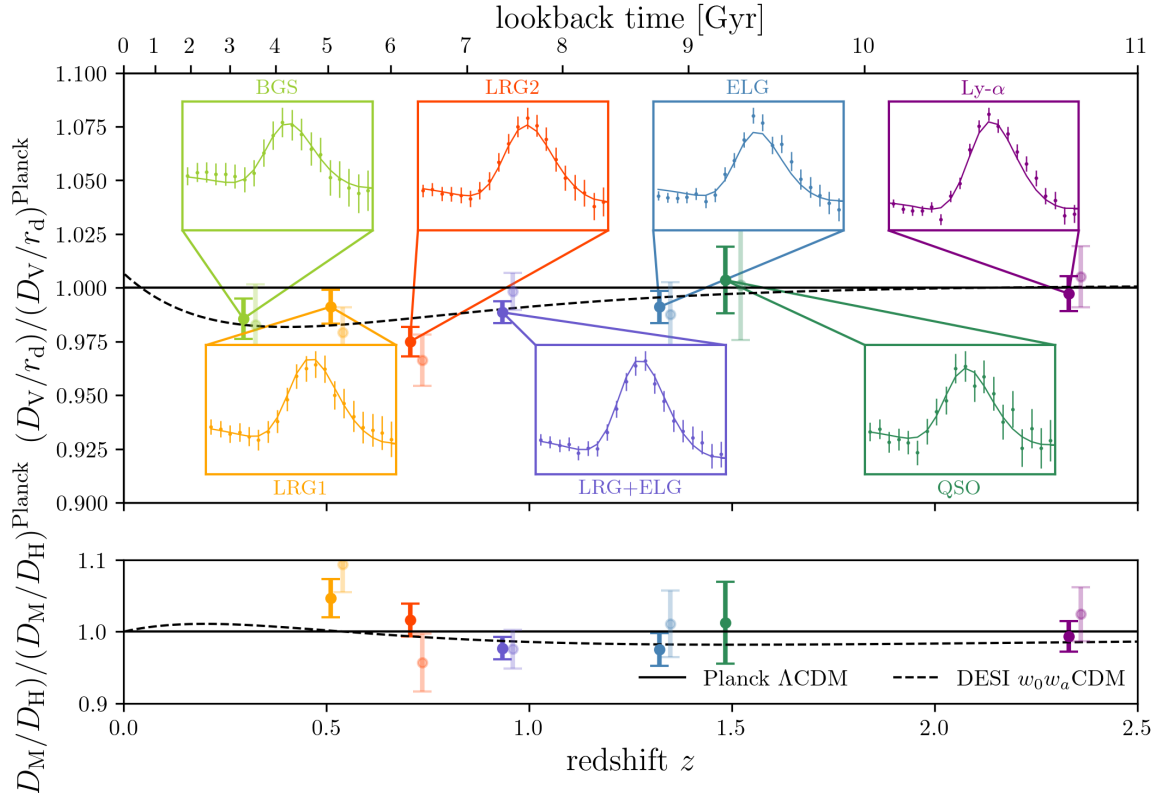


Figure 3.5: Summary of DESI DR2 BAO measurements (figure by Arnaud de Mattia based on DESI Collaboration et al., 2025a,b,f,d); DESI DR1 distance measurements are shown as semi-transparent for comparison. All of the errorbars on this plot, except Ly- α , are also derived from the RASCALC semi-analytic covariance matrices (Chapter 2). An alternative dark energy model (w_0w_a CDM, dashed line) is still preferred to the standard model (Λ CDM, solid line) with *Planck* parameters (Planck Collaboration et al., 2020c).

We show a summary of the BAO features and distance measurements in Figure 3.5. As before, Lyman- α forest BAO results DESI Collaboration et al. (2025b) are also included. The precision of all measurements improved, and anisotropic measurement became possible for the second Emission Line Galaxies (ELG) bin⁴. The precision in all DESI BAO bins now surpasses the BOSS and eBOSS combination (DESI Collaboration et al., 2025a; Alam et al., 2021). The deviations from the standard model did not disappear in the updated measurements, although several bins shifted closer to it.

3.4.2 Cosmological implications

Figure 3.6 presents the DR2 dynamic dark energy (w_0w_a CDM) picture. The constraints based on DESI BAO with both CMB and supernovae improved slightly but shifted closer to the constant dark energy by a similar amount. The increased precision increased the detection without the Type Ia SNe, with DESI BAO and CMB only, although it is still far from the rigorous threshold of 5σ . Lodha et al. (2025a) conducted an extended analysis of alternative dynamic dark energy parametrizations and non-parametric inference methods, to which we refer an interested reader.

Figure 3.7 shows an emerging tension identified between DESI BAO and CMB in Λ CDM. The discrepancy in the matter fraction Ω_m and the Hubble constant H_0 and sound horizon r_d product existed in DR1 at a 1.9σ level. In DR2, constraints became tighter and did not shift, increasing the tension to 2.3σ . After calibrating the sound horizon with the Big Bang Nucleosynthesis (BBN, Schöneberg, 2024) alone or together with the angular sound horizon scale θ_* from the CMB, we infer a higher value of H_0 from DESI DR2 than from the CMB.

Then it is interesting to compare the Hubble constant inferred from DESI BAO with the direct measurement by SH0ES (which is also famously higher than inferred from the CMB, Riess et al., 2022). Figure 3.8 shows that DESI DR2 BAO+BBN H_0 is in a 4.5σ tension with SH0ES. Unfortunately, dynamic dark energy (w_0w_a CDM) does not help to alleviate the Hubble tension, because the inferred Hubble constant value shifts further down (DESI

⁴The first ELG bin is combined with LRG in the main analysis and accordingly this diagram.

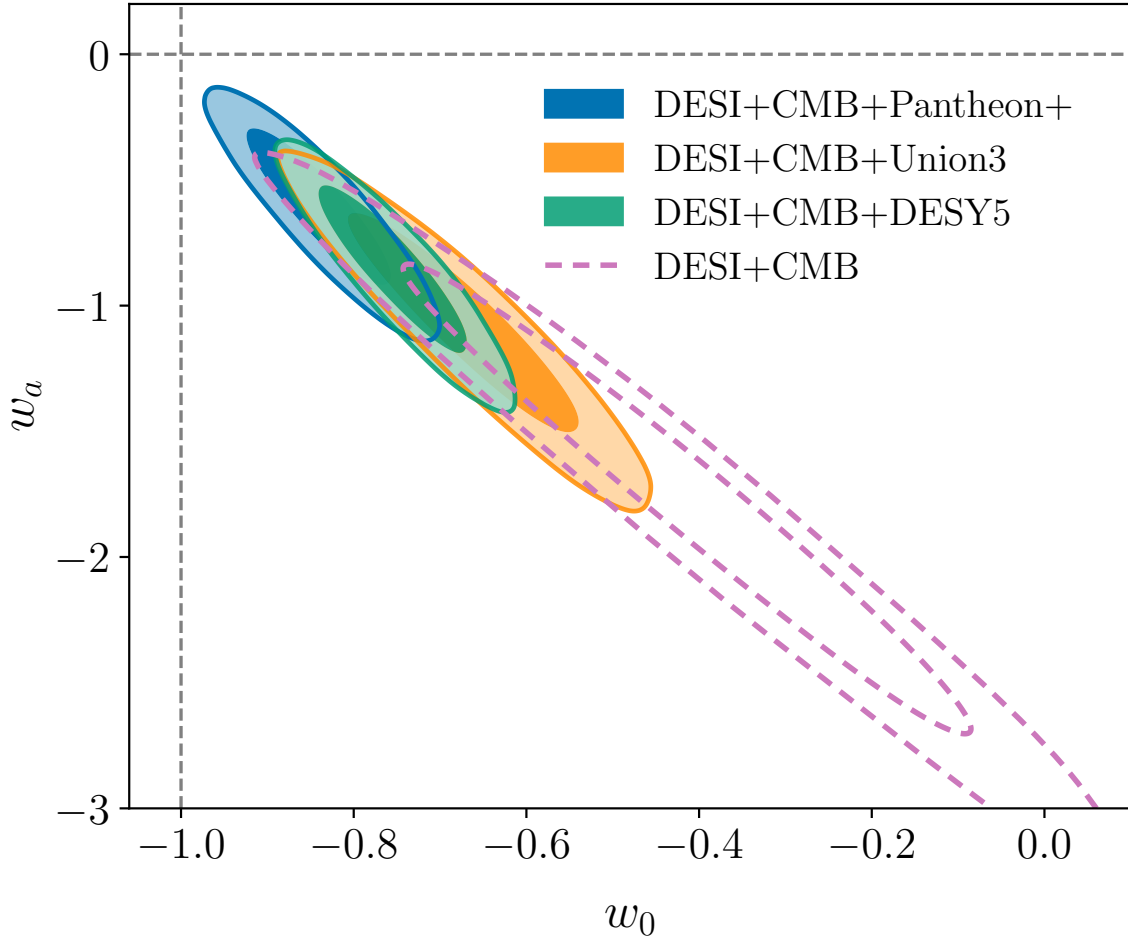


Figure 3.6: w_0w_a CDM model implications of DESI DR2 data in combination with external datasets (figure reproduced from DESI Collaboration et al., 2025a). Compared to DR1 in Figure 3.4, the DESI+CMB+Type Ia SNe contours became tighter but shifted closer to Λ ($w_0 = -1, w_a = 0$), the significance increased slightly to 2.8σ , 3.8σ and 4.2σ respectively. The DESI+CMB combination is now sufficient to indicate the deviation at 3.1σ .

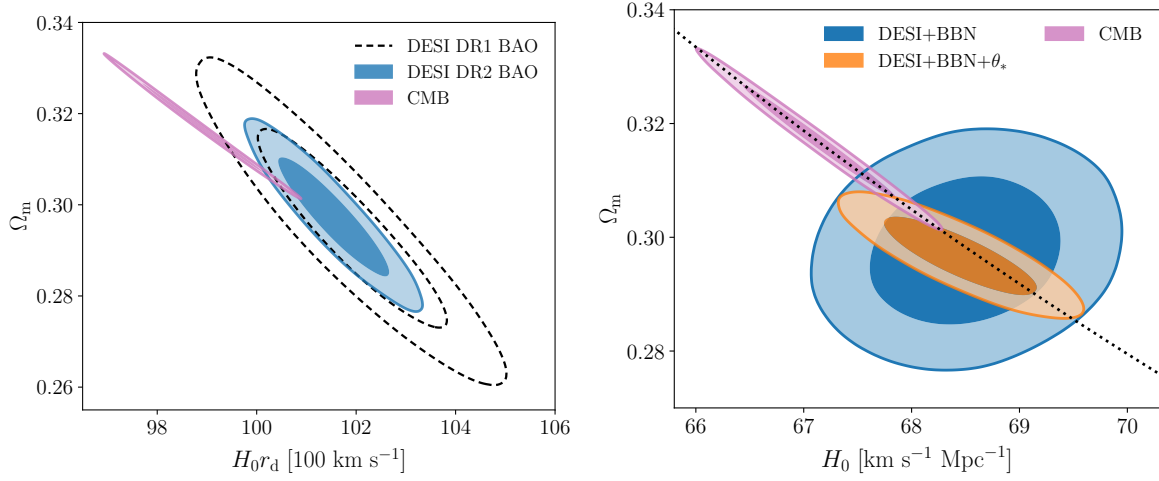


Figure 3.7: Emerging tension identified in Λ CDM between DESI DR2 BAO and CMB (figures reproduced from DESI Collaboration et al., 2025a). The left panel shows uncalibrated BAO, from which we can only infer the product of the Hubble constant and the sound horizon $H_0 r_d$. The DESI DR2 BAO contour is consistent with DR1, but gives tighter constraints thanks to higher measurement precision, which increases the significance of the discrepancy with CMB from 1.9σ to 2.3σ . The right panel shows the results of sound horizon calibration using the Big Bang Nucleosynthesis (BBN, Schöneberg, 2024) alone or together with the angular sound horizon scale θ_* from the CMB. The dotted line shows the degeneracy direction of the CMB.

Collaboration et al., 2025a).

The inference of the Hubble constant from both BAO and CMB crucially depends on sound horizon calibration, which in turn makes certain assumptions about cosmic recombination. The physics of recombination has been studied for a long time, and it may be hard to envision a significant deviation from the standard picture. But Jedamzik & Abel (2011, 2013) suggested that primordial magnetic field (motivated in their own right to e.g. seed the galactic magnetic fields) could induce small-scale inhomogeneities in baryons before recombination, which would alter its pace but would not be directly observable in CMB. This could reduce the sound horizon size and increase the Hubble constant inferred from CMB (Jedamzik & Pogosian, 2020) and BAO. We became interested in this and similar possibilities at an earlier point, which resulted in Chapter 4 of this thesis (originally published as Rashkovetskyi et al., 2021), followed by a summary of more recent developments in Section 4.6.1.

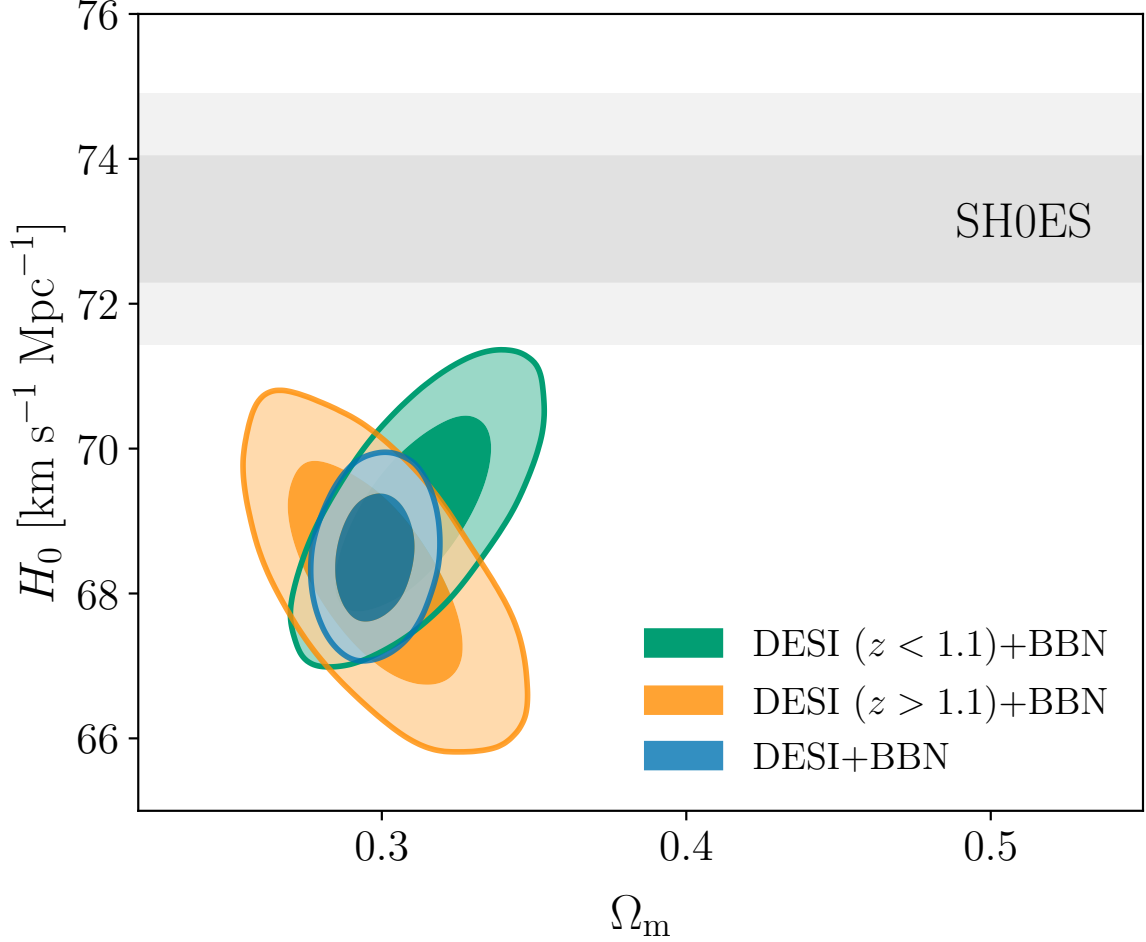


Figure 3.8: Hubble tension implications of DESI DR2 data (figures reproduced from DESI Collaboration et al., 2025a), using Big Bang Nucleosynthesis (BBN, Schöneberg, 2024) for the sound horizon calibration. Without CMB, DESI+BBN is now in a 4.5σ tension with the direct H_0 measurement by SH0ES (Riess et al., 2022). Even after splitting the DESI BAO measurements roughly in two halves in redshift, each has a $> 3\sigma$ discrepancy with SH0ES. However, the BBN calibration of the sound horizon assumes standard recombination.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of High-Energy Physics, under Contract No. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science, Innovation and Universities of Spain (MICIU/AEI/10.13039/501100011033), and by the DESI Member Institutions: <https://www.desi.lbl.gov/collaborating-institutions>. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation, the U. S. Department of Energy, or any of the listed funding agencies.

The authors are honored to be permitted to conduct scientific research on I'oligam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

Chapter 4

Hubble tension and small-scale clumping at recombination¹.

4.1 Introduction

The standard Λ -cold dark matter (Λ CDM) model of cosmology has proven to be remarkably successful in interpreting different measurements consistently and simultaneously. These include the cosmic microwave background (CMB, for example Hinshaw et al. (2013); Planck Collaboration et al. (2020c)), the large-scale structure (LSS, e.g. Eisenstein et al. (2005); Cole et al. (2005); Alam et al. (2017); DES Collaboration et al. (2021)), and probes of the expansion rate of the Universe (such as Perlmutter et al. (1999); Riess et al. (1998, 2021)). However, as the precision of these probes has increased, tensions between them have started to appear.

A notable problem that has emerged within Λ CDM is the Hubble tension—a discrepancy between the cosmic expansion rate today (given by the Hubble parameter H_0) inferred from different data sets. On the one side, the standard CMB analysis of *Planck* 2018 data yields a value of $H_0 = (67.4 \pm 0.5) \text{ km s}^{-1} \text{ Mpc}^{-1}$ (Planck Collaboration et al., 2020c). On the other side, direct H_0 measurements (from type Ia supernovae calibrated with Cepheids (Freedman et al., 2012; Burns et al., 2018; Dhawan et al., 2018; Riess et al., 2021) or surface brightness

¹This chapter was originally published as Rashkovetskyi et al. (2021)

fluctuations (Khetan et al., 2021); from type II supernovae (De Jaeger et al., 2020), strong-lensing time delays (Wong et al., 2020; Birrer et al., 2020; Shajib et al., 2020), gravitational waves standard sirens (Abbott et al., 2017), Tully-Fisher relations (Kourkchi et al., 2020), tip of the red giant branch (Jang & Lee, 2017), Mira variables (Huang et al., 2020) or megamasers (Pesce et al., 2020)) give higher values. In this paper, we will focus on the latest distance-ladder measurement of the *Supernovae, H_0 , for the Equation of State of Dark Energy* (SH0ES) Collaboration $H_0 = (73.2 \pm 1.3)$ km/(s Mpc) (Riess et al., 2021), which is 4.2σ away from *Planck*.

There is another, weaker tension in the values of the matter fraction Ω_m and the amplitude of galaxy clustering σ_8 (on spheres of comoving radius $R = 8/h$ Mpc). Instead of σ_8 , a related parameter $S_8 = \sigma_8 (\Omega_m/0.3)^{0.5}$ is often used, as it is less correlated with Ω_m in LSS data. Planck Collaboration et al. (2020c) report $\Omega_m = 0.315 \pm 0.007$ and $S_8 = 0.831 \pm 0.017$ with *Planck* data, while the Dark Energy Survey Year 1 (DES-Y1, Abbott et al. (2018)) weak-lensing and galaxy-clustering data obtains $\Omega_m = 0.264^{+0.032}_{-0.019}$, $S_8 = 0.783^{+0.021}_{-0.025}$. The new DES Year 3 results are $\Omega_m = 0.339^{+0.032}_{-0.031}$, $S_8 = 0.776 \pm 0.017$ (DES Collaboration et al., 2021), making the tension weaker, but still worth exploring, as other LSS probes disagree with the CMB (Heymans et al., 2021; Krolewski et al., 2021; García-García et al., 2021; Joudaki et al., 2017; Hildebrandt et al., 2020; Ivanov et al., 2020).

Various extensions to Λ CDM have been proposed to solve the H_0 discrepancy. They can be broadly divided into early- and late-type solutions, with the former changing the length of the standard ruler (Poulin et al., 2019; Agrawal et al., 2019; Lin et al., 2019; Sakstein & Trodden, 2020; Kreisch et al., 2020), and the latter the evolution of the expansion rate $H(z)$ at low redshifts (Zhao et al., 2017; Wang et al., 2018; Raveri, 2020; Di Valentino et al., 2020). Only the early-type solutions can be in agreement with low- z standard-ruler measurements of the BAOs, though the most popular model of early dark energy (Hill et al., 2020) worsens the S_8 tension (for a recent review see Knox & Millea (2020)). Here, instead, we study how changing the physics of recombination is an early-type solution to both the H_0 and S_8 tensions, as first proposed in Jedamzik & Pogosian (2020).

The interpretation of CMB data crucially relies on the physics of recombination, so it is natural to ask how well understood, and constrained, this transition is. The process of hydrogen recombination depends crucially on the two-body recombination rate, and thus can be affected by physics at very small scales (Jedamzik & Abel, 2011). In Banerjee & Jedamzik (2004); Jedamzik & Abel (2013); Jedamzik & Saveliev (2019) it was shown that, by creating baryonic clumping at small scales, primordial magnetic fields (PMFs, for a review see Subramanian (2016)) would leave an imprint on the CMB by allowing a more-rapid process of recombination, and shifting the decoupling between photons and baryons to larger z . While such inhomogeneities would take place at very small scales, they enhance the recombination rate when averaged over larger scales. An earlier recombination implies a higher H_0 for a fixed angular sound horizon θ_s in the CMB. Recently, Jedamzik & Pogosian (2020) have shown that such a clumping could relieve both the Hubble and S_8 tensions in current cosmological data. On the other hand, Thiele et al. (2021) argued that the H_0 value inferred from *Planck* and Atacama Cosmology Telescope (ACT) data remains in significant tension with SH0ES.

Here we extend previous analyses by introducing a very generic model of clumping at small scales. This model posits that baryons live in three zones: an average one, and an over/underdense one (see Figure 4.1). This encompasses the models in Jedamzik & Pogosian (2020) and Thiele et al. (2021), as well as other possible origins of small-scale baryonic clumps, such as baryon isocurvature (Dolgov & Silk, 1993).

The key question we tackle is whether a change in recombination that is sufficient to change the sound horizon—and thus explain the H_0 tension—leaves a detectable imprint on the CMB damping tail. The high- ℓ CMB tail has been measured to great success by the *Planck*, ACT (Choi et al., 2020), and South Pole Telescope (SPT, Chown et al. (2018)) collaborations; and upcoming measurements from the *Simons Observatory* (SO, Ade et al. (2019)) and CMB-S4 (Carlstrom et al., 2019) will improve those measurements even further.

This paper is structured as follows. We start in Section 4.2 by defining our generalized clumping model (**M3**), and discussing its physical implications in Section 4.3. In Section 4.4 we analyze the current CMB data from *Planck* 2018 (Planck Collaboration et al., 2020b,d),

exploring clumping- H_0 correlations as well as looking into shifts in S_8 and Ω_m . Then we perform forecasts for future CMB experiments in Section 4.5, to understand how better measurements of the damping tail will test our clumping model more precisely. We conclude in Section 4.6.

4.2 Three-zone model (M3) for recombination

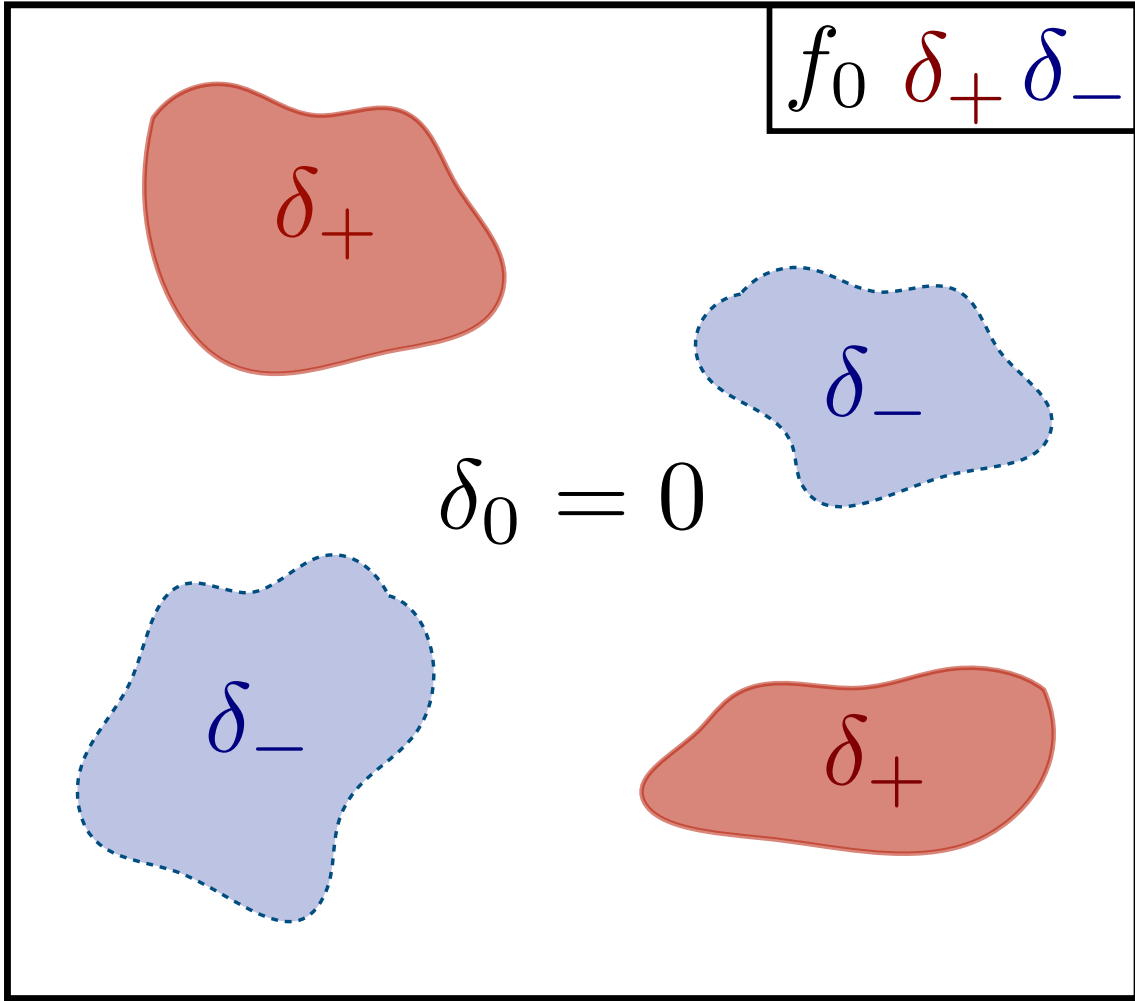


Figure 4.1: M3 spatial structure: regions with average (0), lower (−) and higher (+) density. We take those three effective zones to have constant overdensity $\delta_i = n_{H,i}/\langle n_H \rangle - 1$ each, and the structure is fixed in time. The model is described by six parameters: three δ_i 's and three volume fractions f_i that each zone occupies. There are three constraints: $\sum_i f_i = 1$, $\sum_i f_i \delta_i = 0$ and $\delta_0 = 0$. So we choose f_0 and $\delta_{\pm}$ as input parameters, as highlighted in upper right corner.

We begin by defining our three-zone model (M3) for baryonic clumping. The general idea is that there are fluctuations on very small ($\sim$ kpc for PMFs (Subramanian, 2016)) scales, so that the large-scale behavior of baryons follows the usual assumptions; whereas the recombination rate, which depends on the electron density squared (n_e^2) can be enhanced (for large overdensities) or reduced (for underdensities), with respect to the average.

Our general picture is illustrated in Figure 4.1: there are regions with average (marked by the index 0), lower ($-$) and higher ($+$) density. For simplicity, we take those three effective zones to have constant hydrogen density n_H each, and the structure is constant in time. Then one needs six parameters: three weights (volume fractions) f_i and three densities, which can be parametrized as $\Delta_i = n_{H,i}/\langle n_H \rangle$ or $\delta_i = \Delta_i - 1$ (hereafter $i = \{0, -, +\}$). However, we set three constraints: first, all volume is divided between the three zones ($\sum_i f_i = 1$), second, the total baryonic mass is set by ω_b ($\sum_i f_i \Delta_i = 1$, or equivalently $\sum_i f_i \delta_i = 0$), and finally, one of the zones has average density ($\delta_0 = 0$, which is optional but simplifies the analysis). This leaves three free parameters. For input, we choose the two nonzero relative overdensities δ_- , δ_+ and the volume fraction f_0 of the average-density zone.

This three-parameter model is very flexible, as for example it encompasses the M1 and M2 models presented in Jedamzik & Pogosian (2020) (obtained by fixing $f_0 = 1/3$, and either $\delta_- = -0.9$ for M1 or $\delta_- = -0.7$ for M2). However, our M3 model is only bound by the constraints of volume and mass conservation and the only arbitrary choice is setting one of the three regions to have average density. On the flip side, this flexibility makes the three parameters very degenerate: if either $f_0 \rightarrow 1$, δ_+ or $\delta_- \rightarrow 0$, then the deviation from uniform density becomes negligible.

Therefore, the prior on these parameters ought to be balanced between generality and degeneracy. We choose a log-uniform prior on $|\delta_-|$ ($10^{-5} \leq |\delta_-| \leq 0.955$) and on the ratio $|\delta_+/\delta_-|$ ($0.1 \leq |\delta_+/\delta_-| \leq 10$), and a uniform prior on f_0 ($0 \leq f_0 \leq 1$). The lower bound on $|\delta_-|$ is set to the curvature perturbations $\sqrt{A_s} \sim 10^{-5}$, whereas the higher bound is determined by numerical limitations of the recombination code, more extreme underdensities cause the integration in RECFAST to fail. Bounds on the ratio are chosen so that the under and

overdensity regions are within an order of magnitude from each other, to avoid the unnatural configuration when one δ is negligible and other is significant. Moreover, the constraints force the volume fractions ratio to be $f_+/f_- = -\delta_-/\delta_+$, so very different δ 's will cause one of the volume fractions to be tiny and the whole clumping effect negligible.

An important parameter quantifying the amplitude of the inhomogeneities is the relative variance of densities,

$$b = \frac{\langle (n_H - \langle n_H \rangle)^2 \rangle}{\langle n_H \rangle^2} = -\delta_- \delta_+ (1 - f_0), \quad (4.1)$$

hereafter denoted as the clumping parameter b , following Jedamzik & Pogosian (2020).

Technically, we implement this model in a fork² of the CLASS code³ (Blas et al., 2011). We run the standard recombination code RECFAST (Seager et al., 1999, 2000; Wong et al., 2008) within each zone separately, given its density $n_{H,i}$, producing three recombination histories, in terms of their free electron fractions $x_{e,i}(z) = n_{e,i}/n_{H,i}$. These are then averaged

$$x_e(z) = \sum_i f_i \Delta_i x_{e,i}(z) \quad (4.2)$$

and passed to the rest of the modules in CLASS.

We have also tested our method with the more precise recombination code HyREC (Ali-Haïmoud & Hirata, 2011). While the two codes differ in their predictions of the highest ℓ modes (see e.g. Lee & Ali-Haïmoud (2020)), we find this difference subleading for the purposes of this paper, as shown in Section B.1.1.

4.3 The effects of small-scale gas clumping on the CMB

4.3.1 Recombination

Consider the recombination of an effective three-level hydrogen atom,

$$\frac{dn_e}{dt} + 3Hn_e = -\left(\alpha n_e n_{H^+} - \beta n_{H^0} e^{-E_{21}/kT}\right) C, \quad (4.3)$$

²https://github.com/misharash/class_public

³https://lesgourg.github.io/class_public/class.html

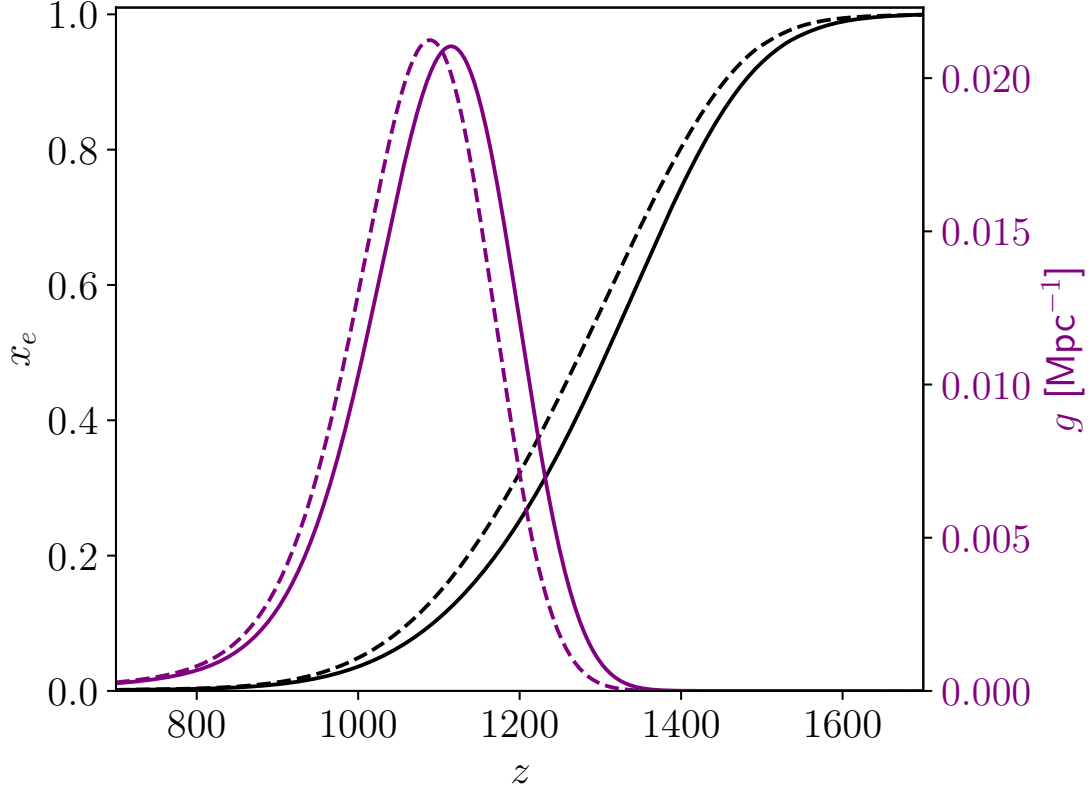


Figure 4.2: Demonstration of the effect of clumping on the recombination history (x_e , black) and visibility function (g , purple) versus redshift z . Standard recombination is shown as dashed curves, and clumping in solid, where we have set $\delta_- = -0.9$, $\delta_+ = 5/3$, and $f_0 = 1/3$, to yield $b = 1$. Cosmology ($\theta_s, \omega_b, \omega_{\text{cdm}}, A_s, n_s, \tau_{\text{reio}}$) is fixed to the *Planck* best fit (Planck Collaboration et al., 2020c).

where α and β are recombination and photoionization rate coefficients, E_{21} is the energy difference between the first excited level and the ground state, k is Boltzmann constant, T is temperature and C is an additional factor taking into account both Lyman- α and two-quantum decays (Peebles, 1968; Zel’dovich et al., 1969). If we take a spatial average, all terms except the first on the right-hand side, will depend on $\langle n_H \rangle$, while that one term will depend on $\langle n_H^2 \rangle$, as it corresponds to recombinations (binding of an electron and an ion). Introducing inhomogeneities enhances the average recombination rate, as $\langle n_H^2 \rangle \geq \langle n_H \rangle^2$ (where equality is only reached for uniform density). This causes n_e to decrease faster, and the Universe to become neutral and transparent to radiation earlier than in the homogeneous case. Given a

recombination history, we define the visibility function

$$g = \dot{\tau} e^{-\tau}$$

as the probability that a CMB photon last scattered per unit conformal time, and thus determines the effective redshift of recombination. It is given in terms of the Thomson optical depth τ and its derivative with respect to conformal time η (namely the inverse of photon’s comoving mean free path),

$$\dot{\tau} = \sigma_T n_{H,\text{now}} x_e (1+z)^2, \quad (4.4)$$

where $n_{H,\text{now}}$ is hydrogen number density today, which is not time- or redshift-dependent, and σ_T is Thomson scattering cross-section.

While Equation (4.3) is a simple approximation—and in our implementation we include the detailed physics of the recombination codes—it serves to illustrate how small scales affect recombination at large scales. As an example, Figure 4.2 shows how clumping affects recombination, where the intuition from Equation (4.3) remains true: clumping shifts recombination and thus the peak of the visibility function to higher redshifts.

To show how recombination evolves in over/under-dense regions, we plot the ionization fraction in each of the three zones of our M3 model in Figure 4.3. The underdense (“-”) zone has a dramatically delayed recombination history, and it presents a sizable low- z tail. The total M3 ionization fraction approaches the “0” zone (namely standard recombination) for lower redshifts.

We note that manually enhancing the average recombination rate (i.e., without keeping track of the overdense and underdense zones) does not capture the entire effect of clumping. To illustrate that, we have also implemented a simple one-parameter clumping model, using only b (Equation (4.1)), where we have assumed a spatially uniform x_e and have therefore just multiplied the rates of recombination and other two-body processes by $\langle n_H^2 \rangle / \langle n_H \rangle^2 = 1 + b$ inside RECFAST. From Figure 4.3 it is clear that this simple model behaves like an overdense zone and does not reproduce the behavior of M3. Such a difference arises because the recombination rate is proportional to $\langle n_e n_{H^+} \rangle \approx \langle x_e^2 n_H^2 \rangle$. Only if one assumes

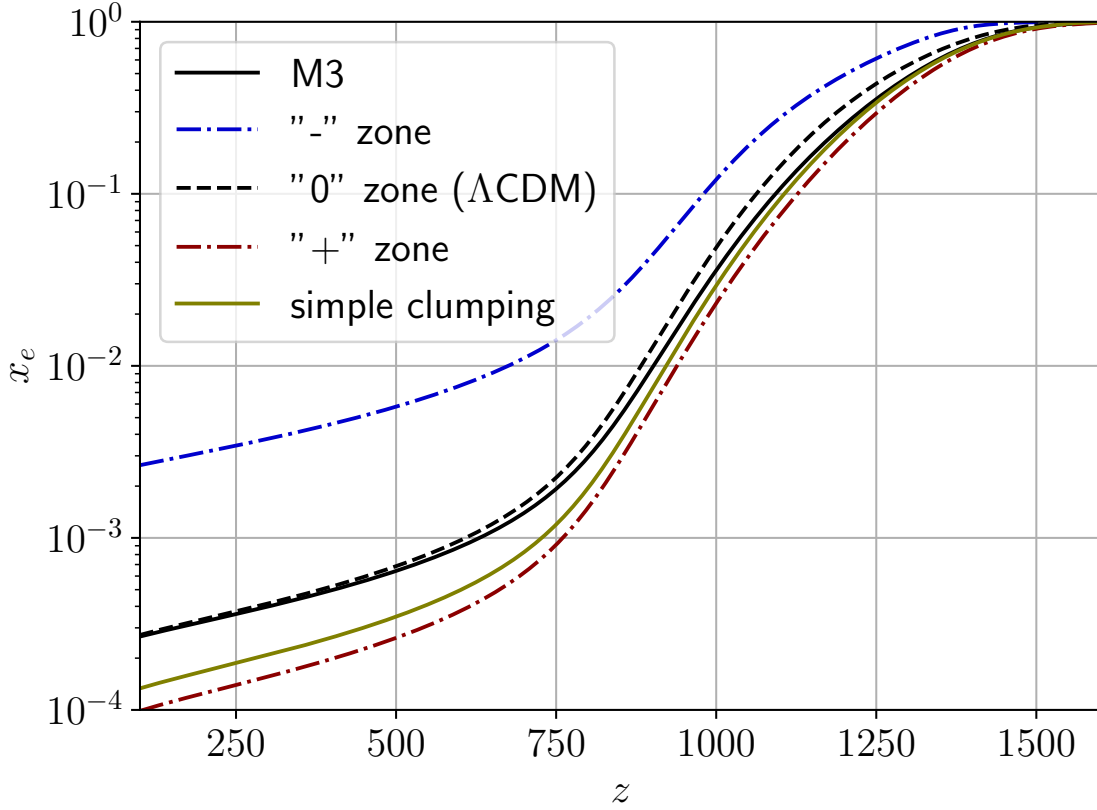


Figure 4.3: Ionization fraction in different zones of the M3 and simple-clumping models. The clumping parameters are $\delta_- = -0.9$, $\delta_+ = 5/3$, and $f_0 = 1/3$, $b = 1$. Cosmology (θ_s , ω_b , ω_{cdm} , A_s , n_s , τ_{reio}) is fixed to the *Planck* best fit (Planck Collaboration et al., 2020c). The “-” zone behaves very different from the others and has a notable low-redshift tail. The total M3 ionization fraction approaches the “0” zone (namely standard recombination), while for simple clumping (when rates of recombination and other two-body processes are just enhanced by a factor of $1 + b$) x_e stays lower. The simple-clumping model produces larger deviations in power spectra than M3 for the same H_0 change and is only shown here for comparison purposes.

constant x_e can one simply put the latter equal to $\langle x_e \rangle^2 \langle n_H^2 \rangle = \langle x_e \rangle^2 \langle n_H \rangle^2 (1 + b)$. But if we assume no electron mixing between the zones, at each density the recombination goes at its own pace. At lower densities the ionization fraction is higher and vice versa. Then $\langle x_e^2 n_H^2 \rangle < \langle x_e \rangle^2 \langle n_H \rangle^2 (1 + b)$, and the actual ratio is time-dependent. As a consequence, this simple clumping model causes significantly higher difference in power spectra C_ℓ for the same change in H_0 than M3 and thus cannot better alleviate the Hubble tension. This implies that matching the low- z tail of recombination is key for consistency with CMB data. Therefore

only M3 and not simple clumping is used further in the paper.

4.3.2 Sound horizon at last scattering

Shifting the epoch of recombination affects the quantities derived from the CMB. For example, distances are inferred from well-measured CMB angular scales, such as the angular scale of the sound horizon at last scattering $\theta_s = r_s/r_*$, where

$$r_* = \int_0^{z_*} \frac{cdz}{H(z)} \quad (4.5)$$

is the comoving distance to the last scattering surface, c is speed of light, and

$$r_s = \int_{z_*}^{\infty} \frac{c_S(z) dz}{H(z)} \quad (4.6)$$

is the comoving distance a sound wave (of speed c_S) could travel before last scattering, called sound horizon. The redshift z_* of last scattering is determined by recombination physics (in particular by the peak of the visibility function) and depends on the radiation, baryon, and matter densities,

$$z_* = z_*(\omega_r, \omega_b, \omega_m); \omega_j = \Omega_j h^2,$$

where $h = H_0/[100 \text{ km}/(\text{s}/\text{Mpc})]$.

The angular sound horizon $\theta_s \sim 1/\ell_{\text{peak}}$ is measured very well from the CMB power spectrum, where ℓ_{peak} is the multipole of the first acoustic peak. This leaves two main avenues to obtain a larger H_0 from the CMB and solve the Hubble tension. Late-type solutions change the $H(z)$ at low redshifts, affecting the distance to last scattering, i.e., r_* in Equation (4.5). Early-type solutions, on the other hand, change the sound horizon r_s in Equation (4.6). This can take the form of an increase in $H(z)$, for instance from early dark energy (Poulin et al., 2019; Sakstein & Trodden, 2020; Agrawal et al., 2019). In our case, however, it is through altering recombination. Clumping changes the $z_*(\omega_r, \omega_b, \omega_m)$ function, which lowers r_s (at fixed ω_i); so to keep the same observed θ_s , the comoving distance r_* must be reduced, yielding higher H_0 .

4.3.3 Silk damping

Another important phenomenon closely related to recombination physics is Silk diffusion damping (Silk, 1968). Photons perform a random walk with nonzero mean free path, which smooths their perturbations, making these decay as time passes. A mode with wave number k is suppressed by a factor $\mathcal{D}(k)$, which can be approximated as

$$\mathcal{D}(k) = \int_0^{\eta_0} d\eta g(\eta) \exp \left\{ - [k/k_D(\eta)]^2 \right\}, \quad (4.7)$$

where the effective diffusion scale is

$$k_D^{-2}(\eta) = \frac{1}{6} \int_0^{\eta} d\eta' \frac{1}{\dot{\tau}} \frac{R^2 + 16(1+R)/15}{(1+R)^2}, \quad (4.8)$$

with $R = 3\rho_b/4\rho_\gamma = (3\omega_b/4\omega_\gamma)(1+z)^{-1}$ (Hu & White, 1997), where ρ_b and ρ_γ are the physical energy densities of baryons and photons respectively.

As the visibility function is peaked around recombination ($\eta = \eta_*$) and normalized ($\int_0^{\eta_0} d\eta g(\eta) = 1$), the damping factor can be approximated by taking out the exponential at peak out of the integral in Equation (4.7), yielding $\mathcal{D}(k) \approx \exp \left\{ - [k/k_D(\eta_*)]^2 \right\}$. This introduces a new length scale into the problem: $r_D = 2\pi/k_D$, which has different parameter dependence from r_S .

In the simplest case of a constant sound speed, r_S scales as $c_S \eta_* = c_S \int_{z_*}^{\infty} dz/H(z)$, which is the distance a sound wave can travel up to recombination, whereas r_D scales roughly as $[\int_0^{\eta_*} d\eta/(1+z)^2]^{1/2} = [\int_{z_*}^{\infty} dz H^{-1}(z)(1+z)^{-2}]^{1/2}$, as the comoving mean free path $\dot{\tau}^{-1}$ scales intrinsically as $(1+z)^{-2}$ (Equation (4.4)). As a consequence, the two scales will react differently to changes in the recombination history. In particular, the damping scale receives a larger contribution from lower redshifts, near z_* , so it is more sensitive to the recombination profile.

Small-scale clumping shifts recombination to earlier times, and with it the peak of g , as we showed in Figure 4.2 for our M3 model. This will alter the damping scale relative to the sound horizon.

To build intuition, we have run a sequence of models with increasing clumping but fixed

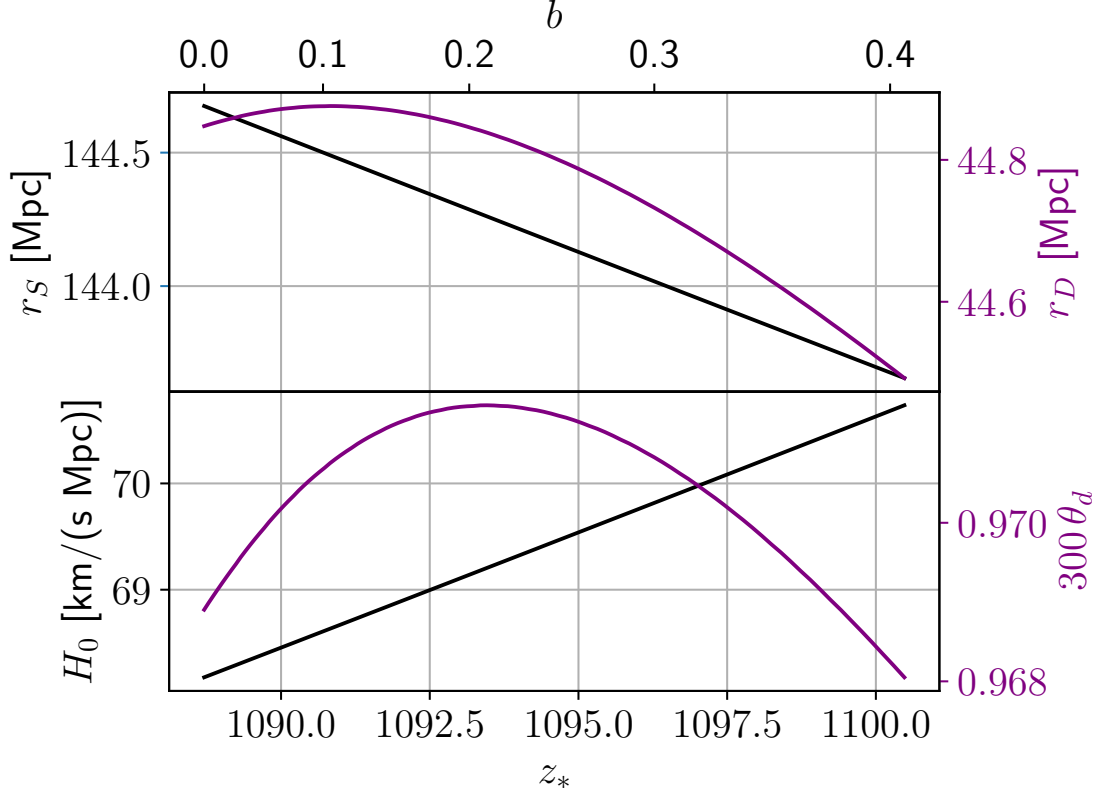


Figure 4.4: Effects of clumping with a fixed cosmology (θ_s , ω_b , ω_{cdm} , A_s , n_s , τ_{reio}), in all cases with $f_0 = 2/3$ and $|\delta_+/\delta_-| = 4/3$. Top panel: the sound horizon r_S (black) decreases with increasing clumping b or recombination redshift z_* , whereas the damping scale r_D (magenta) has a different and more complex behavior. Bottom panel: H_0 increases (as distance to last scattering r_* decreases proportionally to r_s , and other cosmological parameters are fixed); the angular damping scale $\theta_d = r_D/r_*$ first increases and then decreases.

cosmology (in terms of θ_s , which is exquisitely measured by the CMB), keeping $f_0 = 2/3$ and $|\delta_+/\delta_-| = 4/3$ in our M3 model. The effects are shown in Figure 4.4. The comoving damping scale changes differently from the sound horizon, as we reasoned before. Similarly to $\theta_s = r_S/r_*$, we convert it to an angular scale $\theta_d = r_D/r_*$, which first increases with clumping and then decreases.

CMB fluctuations with higher multipoles ℓ are further Silk suppressed, and thus provide a better measurement of the damping scale k_D . Therefore, good precision in the CMB damping tail provides a strong test of clumping.

We illustrate this in Figure 4.5, where we show the relative difference between CMB power spectra, both temperature (TT) and polarization (EE) from M3 and from Λ CDM (with standard recombination), given the same cosmological parameters. In particular, the same sound horizon angular scale θ_s ensures that the acoustic oscillations are in phase with one another, otherwise there would be a large oscillating difference between the power spectra. The gradual deviation on smaller scales ($\ell \gtrsim 1500$ in temperature and $\ell \gtrsim 2000$ in polarization) is caused by the damping scale difference. There are also smaller wiggles in the relative difference, which are caused by the change in duration of last scattering (Hadzhiyska & Spergel, 2019). However, they are less significant than the smoother trend.

We also show the *Planck* measurements and forecasted CMB-S4 binned errors (including both instrumental noise and cosmic variance). Current measurements, from *Planck*, are scattered around zero, so that it is not easy to tell by eye how strongly this particular clumping configuration is disfavored. However, upon evaluating the likelihood we find $\Delta\chi^2_{Planck} = 93$, most of it coming from high- ℓ ($\ell \geq 30$) TT, TE, EE ($\Delta\chi^2_{\text{high } \ell TTTEE} = 91$). The difference is high because in this example we have fixed the rest of the cosmological parameters. As we will show later (Section 4.4.3, Figure 4.9), by shifting the cosmological parameters, M3 is able to fit the $\ell \lesssim 1000$ region very well, though at higher ℓ it diverges due to a difference in the damping scale. With CMB-S4 errors, however, the difference induced by $b \sim 1$ clumping is clearly many sigmas in several dozens of bins both in temperature (TT) and polarization (EE), showing that more precise damping tail measurements will be able to distinguish the presence of significant clumping. For this same example, we find $\Delta\chi^2_{\text{CMB-S4}} \approx 1350$, much higher than for *Planck*.

4.4 Results with *Planck* 2018

In this section we apply the M3 model to *Planck* 2018 data with the key goal of assessing how it alleviates the Hubble tension. We also perform model comparison (Λ CDM with standard recombination versus M3) and discuss the compatibility with LSS measurements.

Our CMB datasets are low- ℓ TT, EE , binned nuisance-marginalized high- ℓ TT, TE, EE

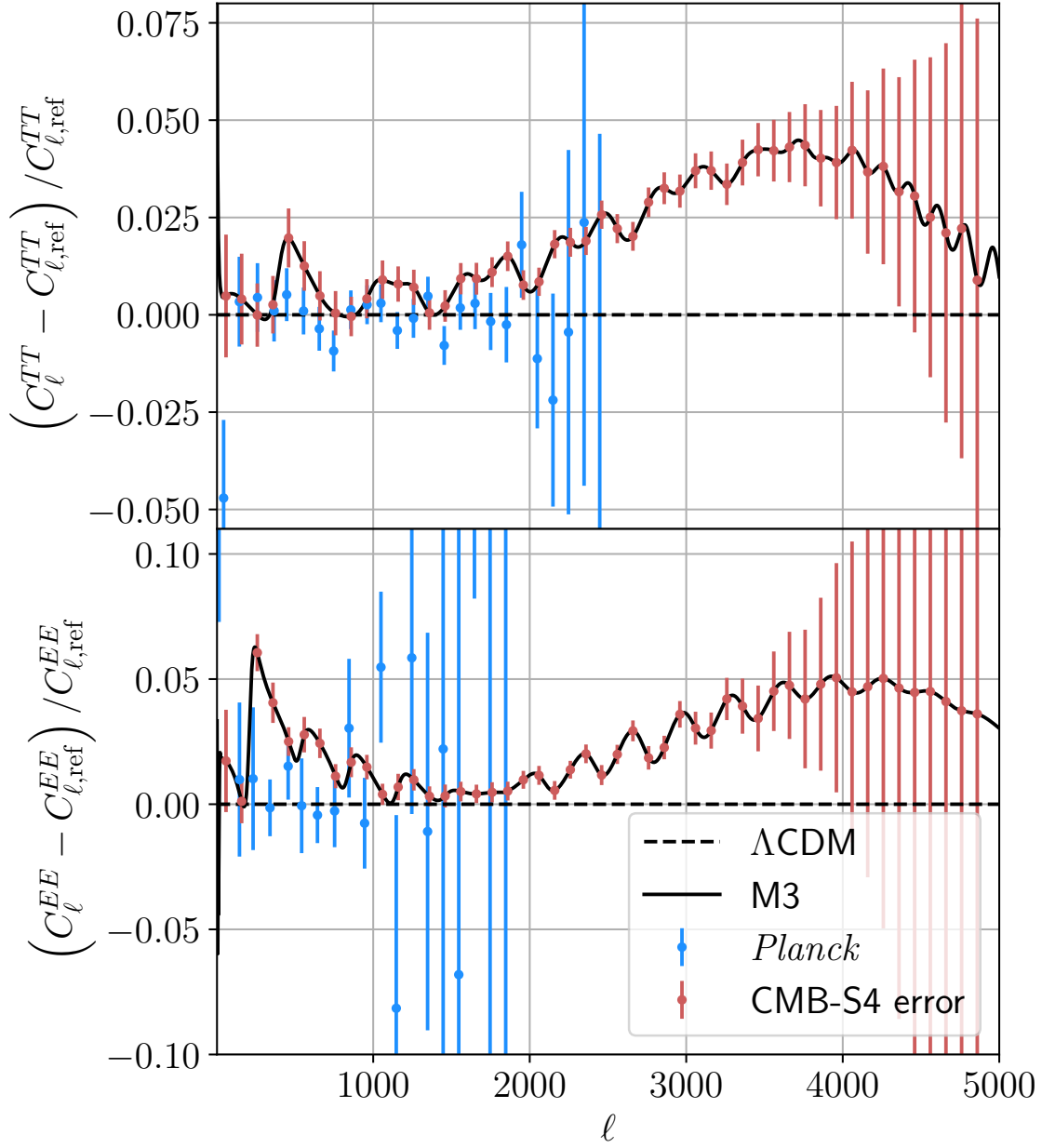


Figure 4.5: Demonstration of the effect of clumping on the CMB damping tail. We define the Λ CDM prediction with standard recombination as the reference $C_{\ell,\text{ref}}$, and compare our clumping model against it. The black dashed line shows a clumping case, with M3 parameters $\delta_- = -0.9$, $\delta_+ = 5/3$, and $f_0 = 1/3$ (giving $b = 1$). The cosmological parameters (θ_s , ω_b , ω_{cdm} , A_s , n_s , τ_{reio}) are fixed to the *Planck* best fit. Cyan points represent the binned *Planck* data, whereas red correspond to CMB-S4 forecasted error bars, both binned with $\Delta\ell = 100$.

(Planck Collaboration et al., 2020b) and lensing (Planck Collaboration et al., 2020d) power spectra. We also consider the Hubble constant measurement from the SH0ES Collaboration: $H_0 = (73.2 \pm 1.3) \text{ km s}^{-1} \text{ Mpc}^{-1}$ (Riess et al., 2021).

We use the Cobaya framework (Torrado & Lewis, 2021) with the Polychord nested sampler (Handley et al., 2015a,b) for evidences (needed to compute the Bayes factors) and posteriors, and the Py-BOBYQA minimizer (Powell, 2009; Cartis et al., 2018a,b) for best-fit determinations. Plots are made with anesthetic (Handley, 2019) and GetDist (Lewis, 2019).

	Prior	Range
$-\delta_-$	Log-uniform	$[10^{-5}, 0.955]$
$ \delta_+/\delta_- $	Log-uniform	$[0.1, 10]$
f_0	Uniform	$[0, 1]$
$\ln(10^{10} A_s)$	Uniform	$[2.55, 3.55]$
n_s	Uniform	$[0.9, 1.05]$
$100\theta_s$	Uniform	$[0.95, 1.15]$
$\Omega_b h^2$	Uniform	$[0.02, 0.025]$
$\Omega_{\text{cdm}} h^2$	Uniform	$[0.1, 0.15]$
τ_{reio}	Uniform	$[0.01, 0.2]$
Ω_K	Fixed	0
$m_\nu \text{ [eV]}$	Fixed	0
$A_{\text{planck}}(y_{\text{cal}})$	Normal	1 ± 0.0025

Table 4.1: Parameters used in this chapter and their priors. $\delta_\pm$ and f_0 are only for M3 model, $A_{\text{planck}}(y_{\text{cal}})$ —for *Planck* likelihoods.

Our parameters and priors are described in Table 4.1. It is important to state that we assumed massless neutrinos throughout the sampling to save computing time, as massive neutrinos slow the Boltzmann solver by a factor of ~ 10 . This assumption shifts upwards the inferred H_0 values, but it does not affect the changes introduced by M3 (relative to Λ CDM with standard recombination) in a meaningful way, as we demonstrate in Section B.1.2.

Full contour plots for the M3 clumping model are presented in Section B.2; here we will focus on particularly important subspaces.

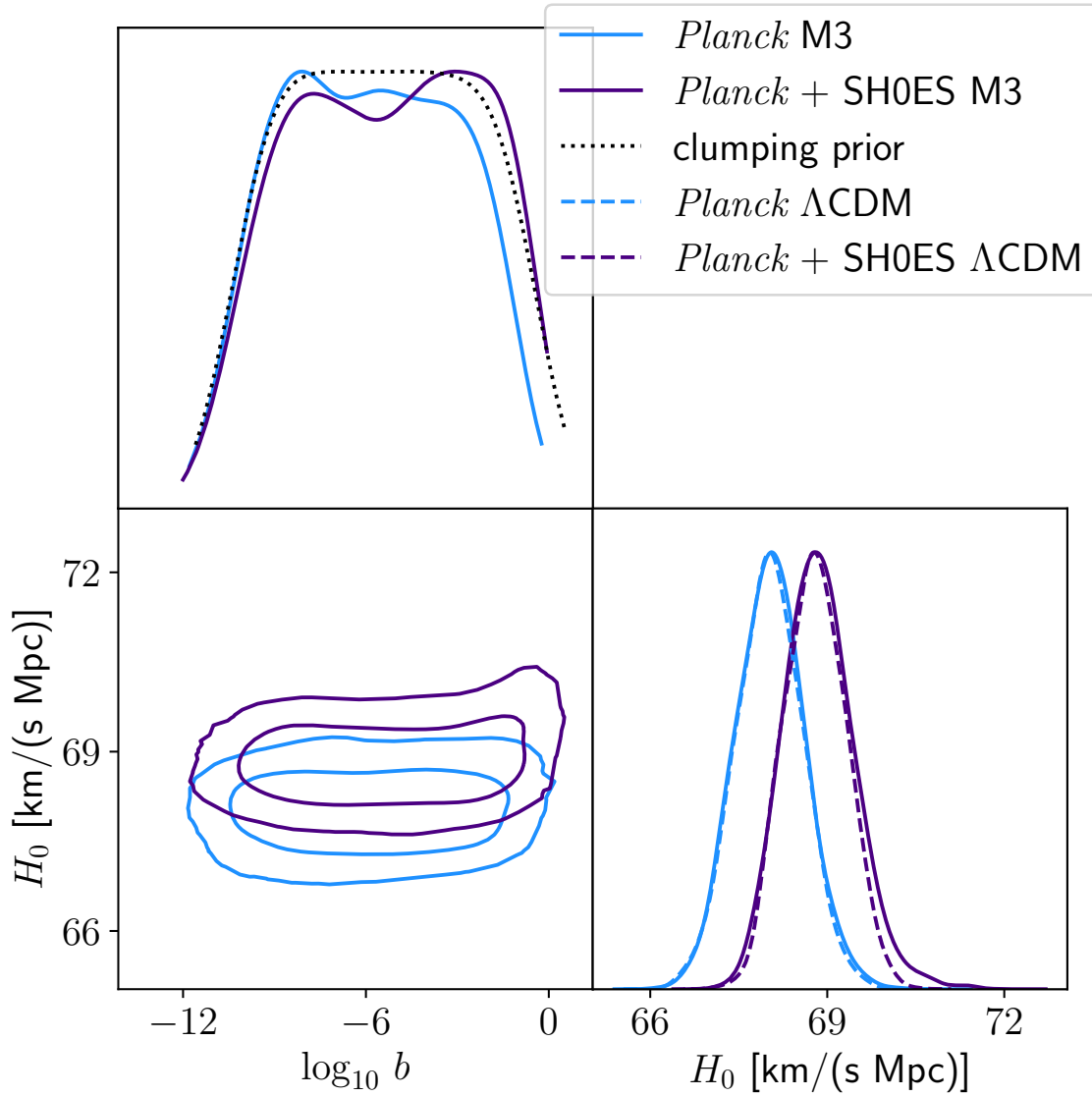


Figure 4.6: Posteriors for H_0 and the clumping parameter b of our M3 model (see Equation (4.1) for its definition). We show 68% and 95% CL ellipses using *Planck* 2018 data alone and in combination with SH0ES, which slightly increases H_0 though does not prefer clumping. We also show the H_0 posteriors from Λ CDM runs with standard recombination, as well as the prior on b for comparison. The 68% C.L. intervals on clumping within M3 are $\log_{10} b = -5.9 \pm 2.7$ for *Planck*, $\log_{10} b = -5.3^{+4.2}_{-3.7}$ for *Planck*+SH0ES (and $\log_{10} b = -5.5 \pm 3.0$ for prior). Neither *Planck* only nor *Planck*+SH0ES data prefer large clumping ($b \sim 1$).

4.4.1 H_0 and model comparison

We show the 2D posterior for the clumping parameter b and H_0 in Figure 4.6. *Planck*-only data show no noticeable change in H_0 compared to Λ CDM with standard recombination, and some preference against high clumping compared to the prior. Adding a direct H_0 measurement, however, creates a weak preference for high clumping and high H_0 . Because most of the posterior weight for the clumping parameter b is below unity, we observe almost no correlation between b and H_0 .

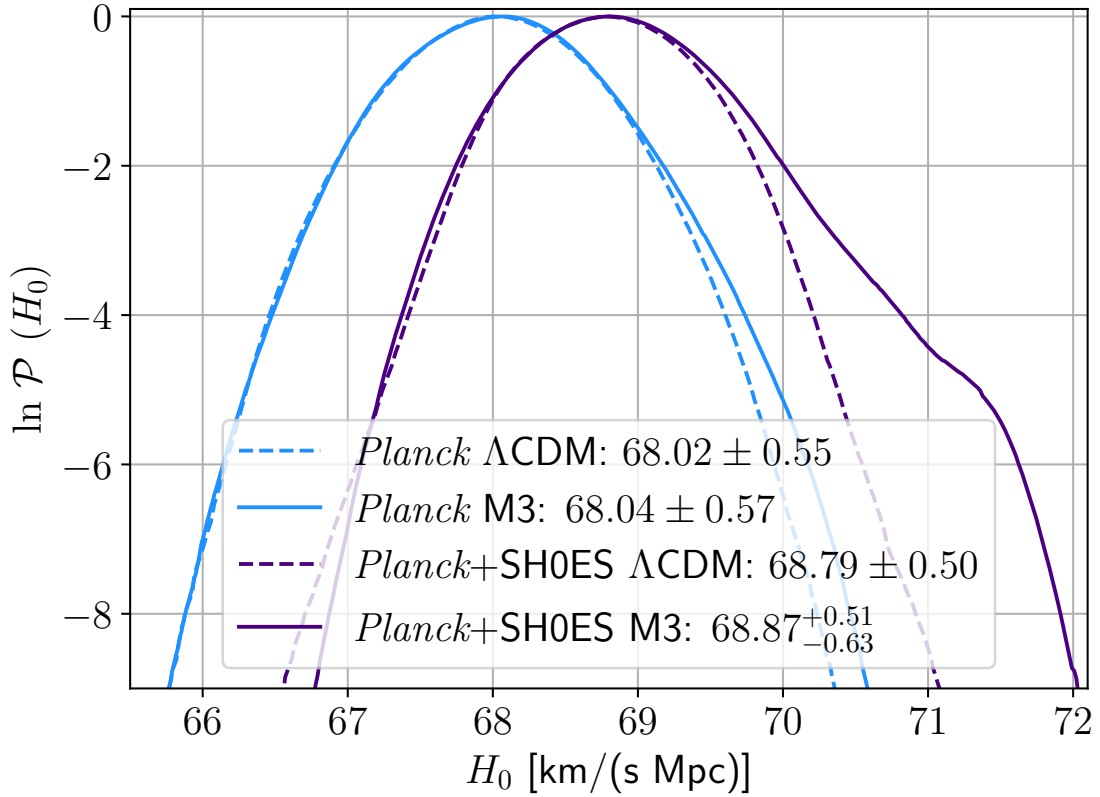


Figure 4.7: Posterior of H_0 inferred from *Planck* data, without and with SH0ES, which shifts the H_0 values upward as expected. The legend shows 68% C.L. intervals.

To explore the tail of the H_0 posterior distribution, we show it in Figure 4.7. With *Planck*-only data, M3 allows for a weak bump towards higher H_0 (compared to standard recombination), whereas the mean is not significantly shifted. For *Planck* data combined with

SH0ES, the bump at higher H_0 for M3 is stronger, given the additional pull from direct H_0 measurement, though the shift in the mean is still not significant [$\Delta H_0 = 0.08$ km/(s Mpc)].

	$\Delta\chi^2_{best}$	$\log_{10} K$
<i>Planck</i>	0	0.04 ± 0.15
<i>Planck</i> +SH0ES	5	0.14 ± 0.19

Table 4.2: Model comparison between M3 and Λ CDM by best-fit χ^2 difference and Bayes factor K between M3 and Λ CDM with *Planck* 2018 data. χ^2 differences rounded to integers because of uncertainty in the minimizer output. M3 does not fit *Planck* data alone better than Λ CDM. With SH0ES, M3 allows for slightly better agreement. Bayes factors $-0.5 \lesssim \log_{10} K \lesssim 0.5$ show no clear preference between models (Kass & Raftery, 1995).

We show the best-fit χ^2 differences and Bayes factors K between M3 and Λ CDM models in Table 4.2. We have not found a better M3 fit to *Planck* 2018 data alone, compared to Λ CDM with standard recombination. M3 is more successful than Λ CDM when considering *Planck*+SH0ES, but $\Delta\chi^2 \approx 5$ can not justify three extra parameters. The Bayes factor K in both cases is consistent with 1 (within $< 1\sigma$), meaning no preference to either model. The Bayes factor is equal to the ratio of marginalized posterior probabilities of the models if they are assumed equally probable *a priori*; more generally, posterior probabilities ratio is given by the Bayes factor times prior probabilities ratio (Kass & Raftery, 1995). The best-fit parameters are presented in Section B.3 and Table B.1.

We conclude that M3 is neither supported by the data nor rejected. This is probably not surprising, since *Planck* data are fit well by Λ CDM with standard recombination and with an $\approx 4\sigma$ tension between *Planck* and SH0ES it is challenging to get a detection in a three-parameter model. One could get more support for M3 by considering additional H_0 data. Here, however, we will limit ourselves to the SH0ES measurement.

4.4.2 Low- ℓ *Planck* data analysis

To build intuition, we now check whether it is the damping tail that prevents *Planck* 2018 data from preferring M3. For that, we perform an analysis with only the $\ell < 1000$ multipoles. This range of scales is chosen to determine the sound horizon angular scale, though not the damping tail.

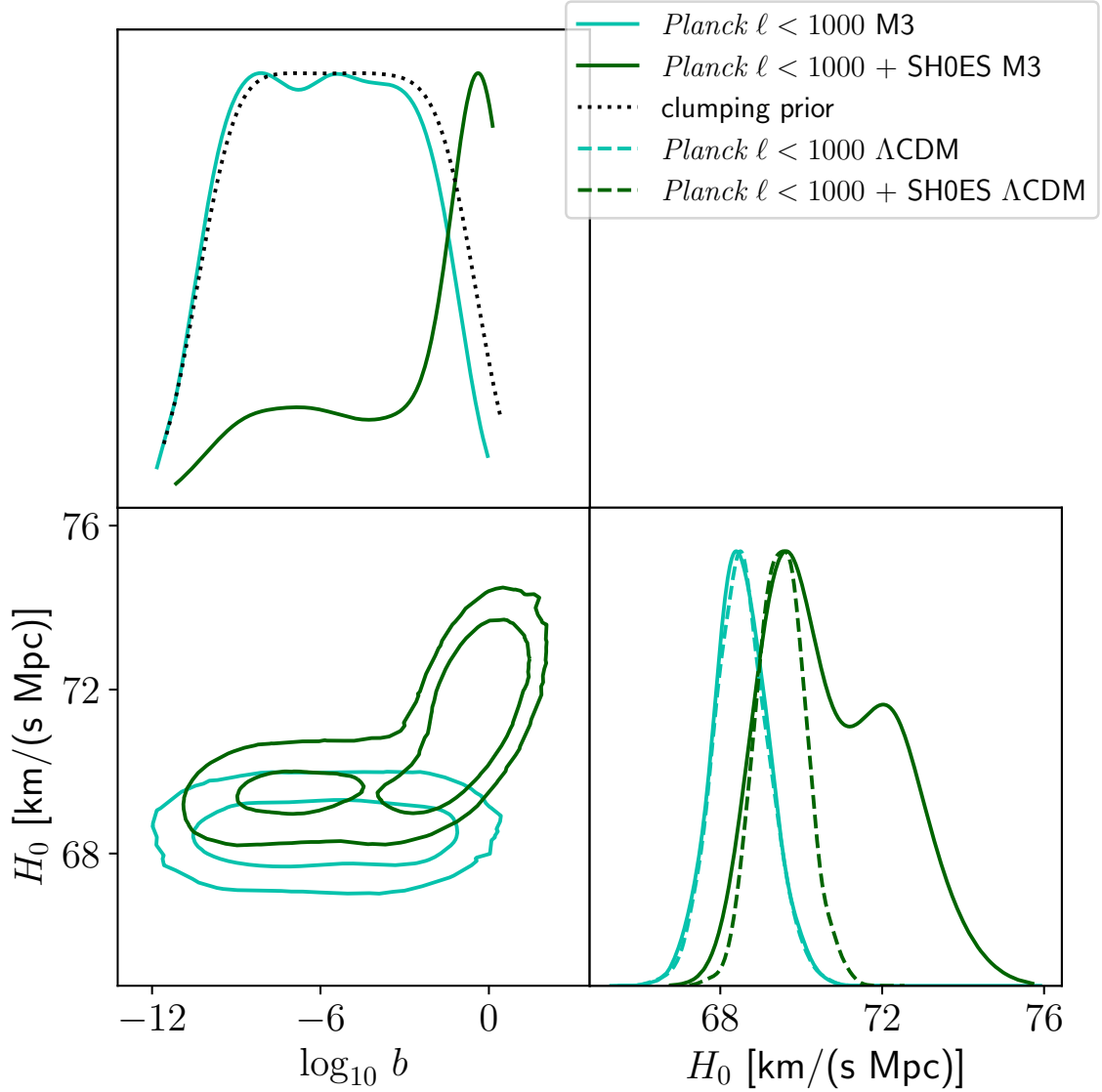


Figure 4.8: Confidence regions for clumping b versus H_0 from *Planck* 2018 $\ell < 1000$ data. The absence of damping tail information from the high- ℓ data allows significant clumping, and thus larger H_0 . The 68% C.L. intervals on clumping within M3 are $\log_{10} b = -5.9 \pm 2.8$ for *Planck* $\ell < 1000$, $\log_{10} b = -2.8^{+2.9}_{-6.3}$ for *Planck* $\ell < 1000$ + SH0ES (and $\log_{10} b = -5.5 \pm 3.0$ for prior). Within M3, $H_0 = (68.51 \pm 0.68)$ km/(s Mpc) for *Planck* $\ell < 1000$, $H_0 = (70.8 \pm 1.5)$ km/(s Mpc) for *Planck* $\ell < 1000$ + SH0ES. The H_0 posteriors from Λ CDM runs are also shown, which do not reach the high H_0 values available to M3.

The corresponding 2D posteriors for b and H_0 are plotted in Figure 4.8. Using *Planck* $\ell < 1000$, we find no significant deviation from standard recombination, as the clumping posterior is only slightly shifted from the prior (towards lower values). However, with the

additional pull from SH0ES, strong clumping $b \sim 1$ is preferred, and we see a significant bump towards higher H_0 . The best-fit parameters are presented in Section B.3 and Table B.2.

	$\Delta\chi^2_{best}$	$\log_{10} K$
<i>Planck</i> $\ell < 1000$	0	-0.09 ± 0.18
<i>Planck</i> $\ell < 1000$ +SH0ES	11	0.21 ± 0.20

Table 4.3: Model comparison with *Planck* $\ell < 1000$ data. χ^2 differences are rounded to integers because of uncertainty in the minimizer output. M3 still does not fit the CMB alone better than Λ CDM, though the joint fit to CMB and SH0ES is improved further than with full *Planck*. The Bayes factors K show no preference between models, as for full *Planck* in Table 4.2.

To determine whether clumping provides a better fit in this case, we perform model comparison by two methods: best-fit χ^2 and Bayes factor K , and show the results in Table 4.3. We still do not find M3 to fit CMB data notably better than the standard model, with the addition of SH0ES improvement is more significant ($\Delta\chi^2_{best} = 11$) than in the case of full *Planck* and SH0ES ($\Delta\chi^2_{best} = 5$). Bayes factors K are still consistent with one, telling no clear preference between the models. This shows that current *Planck* data do not prefer clumping as the solution to the H_0 tension, even without its damping tail.

4.4.3 Damping scale

As we have shown in the previous subsection, considering only $\ell < 1000$ multipoles from the *Planck* data opens more room for clumping than using the full data. We posit that the key source of constraints on clumping is the damping-tail information contained in higher- ℓ multipoles. So now we consider how the damping tail varies within our M3 model.

First, we plot the relative difference in TT, EE power spectra between the best fits in Figure 4.9. Unlike in Figure 4.5, where the cosmology was kept fixed, producing significant differences, here all the models manage to shift the parameters to fit the low- ℓ data. However, they diverge significantly in the damping tail ($\ell \gtrsim 1500$), and CMB-S4 will be able to measure such deviations. More precisely, the χ^2 difference between *Planck* Λ CDM and *Planck*+SH0ES M3 best fits from *Planck* temperature, polarization and lensing is only $\Delta\chi^2_{Planck} \approx 4$, while for CMB-S4 precision in temperature and polarization it can be as high as $\Delta\chi^2_{CMB-S4} \approx 1030$

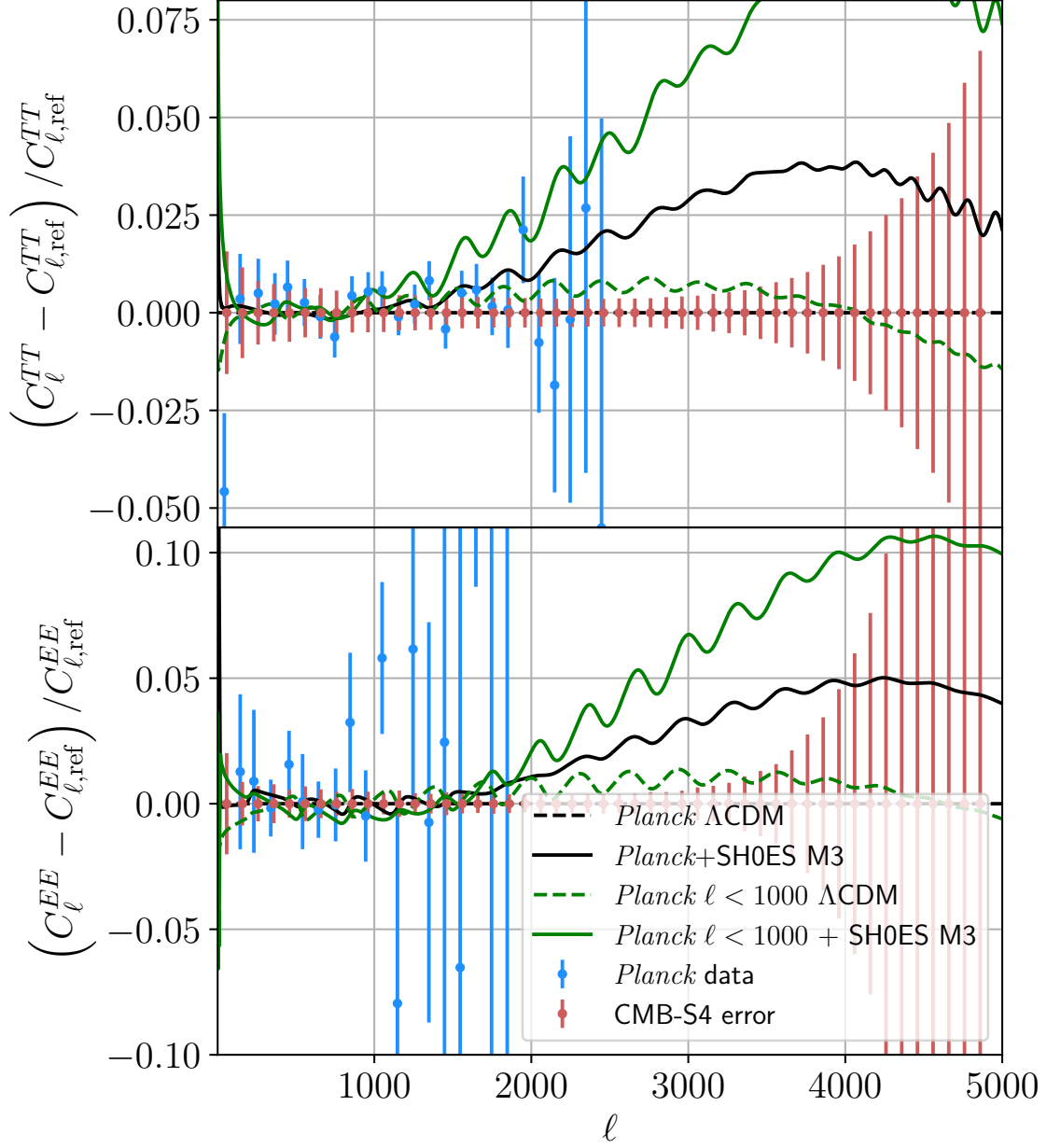


Figure 4.9: Relative difference in CMB temperature TT and polarization EE power spectra between our best fits (Section B.3 and Tables B.1 and B.2), with the Λ CDM best fit to *Planck* taken as reference. Differences are small in the low- ℓ regime, but grow at smaller scales. *Planck* data can barely distinguish them, but CMB-S4 will do so very clearly.

(if one assumes the *Planck* Λ CDM best fit as fiducial).

To build intuition, we formulate the damping-tail constraints in terms of the comoving

damping scale r_D and its angular analog $\theta_d = r_D/r_*$. We postprocessed the *Planck* runs to get this information from CLASS.

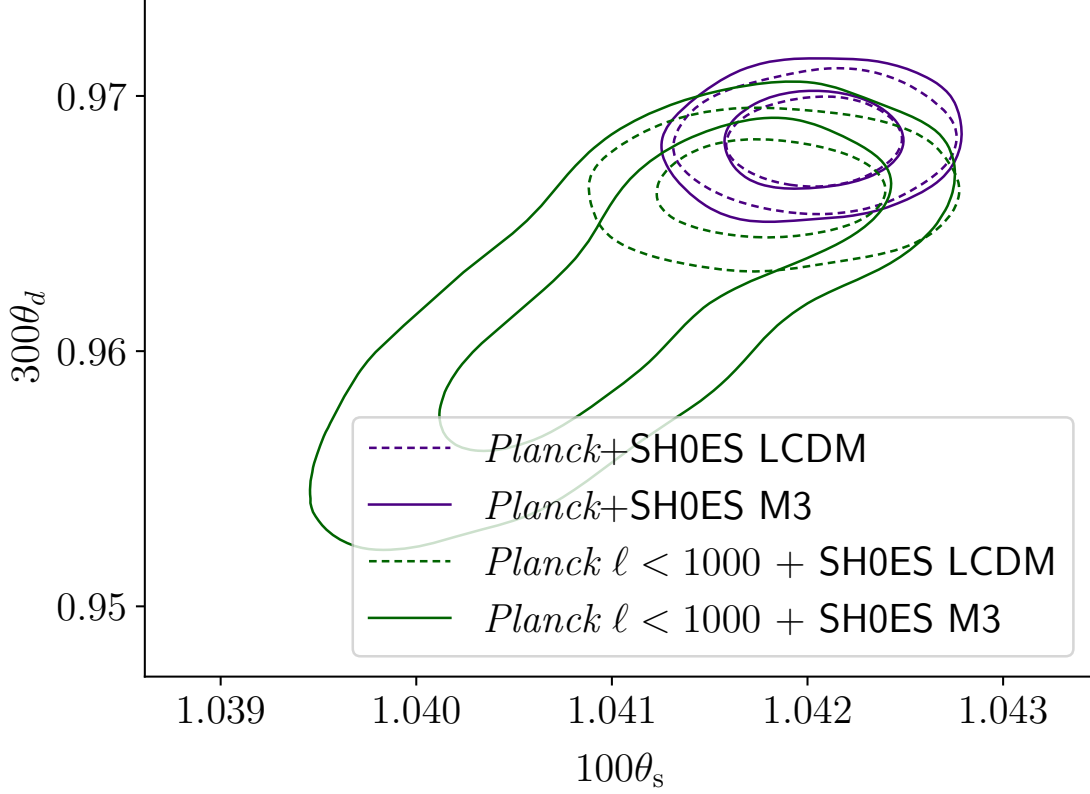


Figure 4.10: Sound-horizon vs. damping angular scales [both multiplied by a prefactor to make them $\mathcal{O}(1)$] for *Planck* 2018 + SH0ES runs. The M3 and Λ CDM contours are quite similar for full *Planck*+SH0ES (purple). On the contrary, for *Planck* $\ell < 1000$ +SH0ES (green), the M3 posterior is significantly extended to lower values of both θ_s and θ_d , compared to Λ CDM.

Figure 4.10 presents the sound-horizon vs. damping angular scales for our *Planck*+SH0ES runs. Within Λ CDM both angular scales are well measured, both for the full *Planck* data as well as with the $\ell < 1000$ modes only. The M3 contours, however, extend to lower values of both θ_d and θ_s compared to Λ CDM. With full *Planck* the shift is not significant, while using only $\ell < 1000$ allows larger deviations to that region. With *Planck* $\ell < 1000$ + SH0ES, the error bar of θ_d within M3 is by a factor of 3 wider than within Λ CDM.

We note that the error bar of θ_s is widened similarly. However, the positions of acoustic

peaks are not determined exactly by sound horizon angular scale alone. For example, stronger damping would shift the power spectrum maxima to slightly lower ℓ . Acoustic peak positions may also be affected by fine changes in visibility function shape introduced by clumping. Such small changes might be important, since the characteristic differences in sound horizon scales in our runs are only $\sim 0.1\%$. To assess this, we calculate the positions of first three peaks in TT power spectrum given by CLASS [by fitting Gaussians to $\mathcal{D}_\ell = \ell(\ell+1)C_\ell/2\pi$, as *Planck* Collaboration (Planck Collaboration et al., 2020a)]. We find that M3 keeps peaks at the same positions as Λ CDM, especially the second one. Therefore we plot second TT peak position instead of θ_s in Figure 4.11.

Figure 4.11 also shows how H_0 changes in the $\ell_{\text{peak2},TT}-\theta_d$ plane, using *Planck* 2018 $\ell < 1000$ and SH0ES data. The main trend is that H_0 increases for smaller values of θ_d . Therefore, lowering the angular damping scale θ_d is necessary to infer a higher H_0 from CMB. But such a change is disfavored by *Planck* damping tail data. This agrees with our previous subsections, where full *Planck* + SH0ES did not show a preference for high clumping, and consequently did not exhibit significant change in H_0 , while without $\ell \geq 1000$ multipoles the data allowed for both.

4.4.4 S_8 tension

We now move to discuss whether clumping is compatible with large-scale structure data, which have not been considered in previous subsections.

A potentially interesting discrepancy between the CMB and LSS is the S_8 tension in the amplitude of matter fluctuations measured from the CMB and the LSS. *Planck* 2018 reported $\Omega_m = 0.315 \pm 0.007$ and $S_8 = 0.831 \pm 0.017$ (Planck Collaboration et al., 2020c), both of which are higher than found in DES-Y1: $\Omega_m = 0.264^{+0.032}_{-0.019}$, $S_8 = 0.783^{+0.021}_{-0.025}$ (Abbott et al., 2018). Earlier recombination (for instance due to small-scale baryon clumping) decreases both the Ω_m and S_8 values inferred from the CMB, and can therefore help to relieve the S_8 tension (Jedamzik & Pogosian, 2020).

However, the new DES-Y3 results ($\Omega_m = 0.339^{+0.032}_{-0.031}$, $S_8 = 0.776 \pm 0.017$ (DES Collabora-

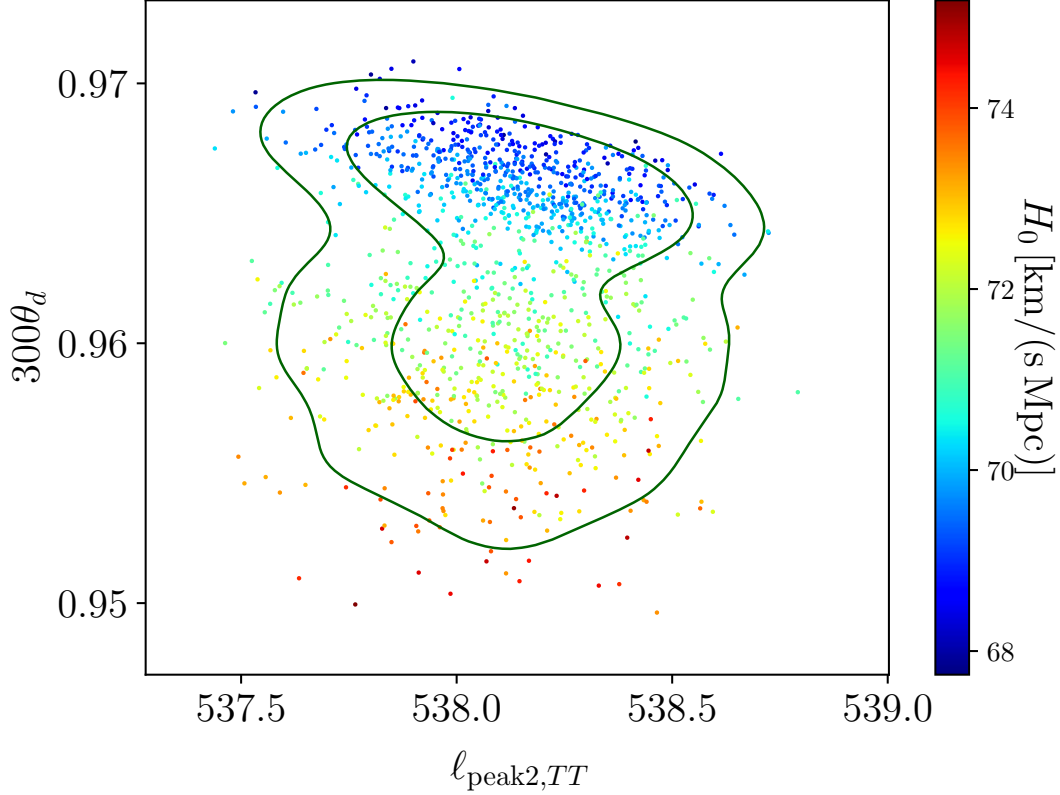


Figure 4.11: Similar to Figure 4.10, but we replace θ_s by position of the second peak in TT power spectrum $\ell_{\text{peak2},TT}$ and color different points by their value of H_0 , for our *Planck* $\ell < 1000$ + SH0ES M3 run. Within the M3 peak positions stay almost same regardless of θ_d . This proves that $\ell < 1000$ multipoles from *Planck* give strong enough constraints on acoustic peaks. In this plane, H_0 increases for lower values of both θ_d and ℓ_{peak} . However, the peak positions are measured with $\sim 0.1\%$ precision and do not allow to vary H_0 significantly with other parameters fixed. This implies that a change in the angular damping scale is necessary to infer higher H_0 from the CMB.

tion et al., 2021)) do not show a preference for lower values of Ω_m , which makes clumping less favorable resolution. A reanalysis of DES-Y1 data according to the DES-Y3 pipeline shifted the parameter estimates to $\Omega_m = 0.303^{+0.034}_{-0.041}$, $S_8 = 0.747^{+0.027}_{-0.025}$ (DES Collaboration et al., 2021), in better agreement with *Planck* on Ω_m but worse on S_8 . Other experiments also find a lower value of S_8 than *Planck* and only a small difference in Ω_m ($< 1\sigma$), including the Kilo-Degree Survey (KiDS-1000, which reported $\Omega_m = 0.305^{+0.010}_{-0.015}$, $S_8 = 0.766^{+0.020}_{-0.014}$ (Heymans et al., 2021)), unWISE galaxies (with *Planck* CMB lensing added, which obtained $\Omega_m = 0.295 \pm 0.017$, $S_8 = 0.776 \pm 0.017$ (Krolewski et al., 2021)), as well as an analy-

sis of the growth of density perturbations from large-scale structure data (which yielded $\Omega_m = 0.311^{+0.021}_{-0.028}$, $S_8 = 0.7769 \pm 0.0095$ (García-García et al., 2021), see also (D’Amico et al., 2020; Colas et al., 2020)).

To study in detail how clumping interfaces with the S_8 tension, we show the one-dimensional posteriors and two-dimensional confidence ellipses for H_0 , Ω_m and S_8 for a few selected runs in Figure 4.12. We also show analogous posteriors and contours on Ω_m and S_8 from DES-Y3 (DES Collaboration et al., 2021), DES-Y1 (Abbott et al., 2018) and KiDS-1000 (Heymans et al., 2021) for comparison. This figure shows that *Planck* $\ell < 1000$ M3 prefers low clumping and therefore is closer to Λ CDM with standard recombination, exhibiting similar correlations between these three parameters. Note that within Λ CDM there is a significant negative correlation between H_0 and Ω_m and a weaker negative correlation between H_0 and S_8 . Therefore increasing H_0 alone (for instance by coadding direct H_0 measurements) decreases both Ω_m and S_8 .

The *Planck* $\ell < 1000$ + SH0ES M3 confidence region exhibits high clumping and explores a different direction, to even lower Ω_m and S_8 . Finally, in full *Planck* with SH0ES the damping-tail data disfavor high clumping, making the contour close to the standard Λ CDM.

We note that a rigorous study of the S_8 tension within M3 would require a reanalysis of the LSS data, as clumping can introduce biases with respect to Λ CDM with standard recombination. Such an analysis is beyond the scope of this work, and given that this tension is weaker than the H_0 one, we tentatively conclude that adding LSS data to our runs would not significantly change whether M3 is preferred.

4.4.5 Baryon drag scale

As advanced above, the addition of clumping changes the length of the sound horizon due to the nonstandard recombination. This is important for the interpretation of baryon acoustic oscillations (BAO) in galaxy surveys at low z , so we now consider how the BAO standard ruler is affected by clumping. The relevant distance is the drag scale r_{drag} —the sound horizon at the drag epoch (when the baryon optical depth is 1). The drag epoch occurs slightly later than last scattering, which makes the drag scale larger than the sound horizon r_s at last scattering,

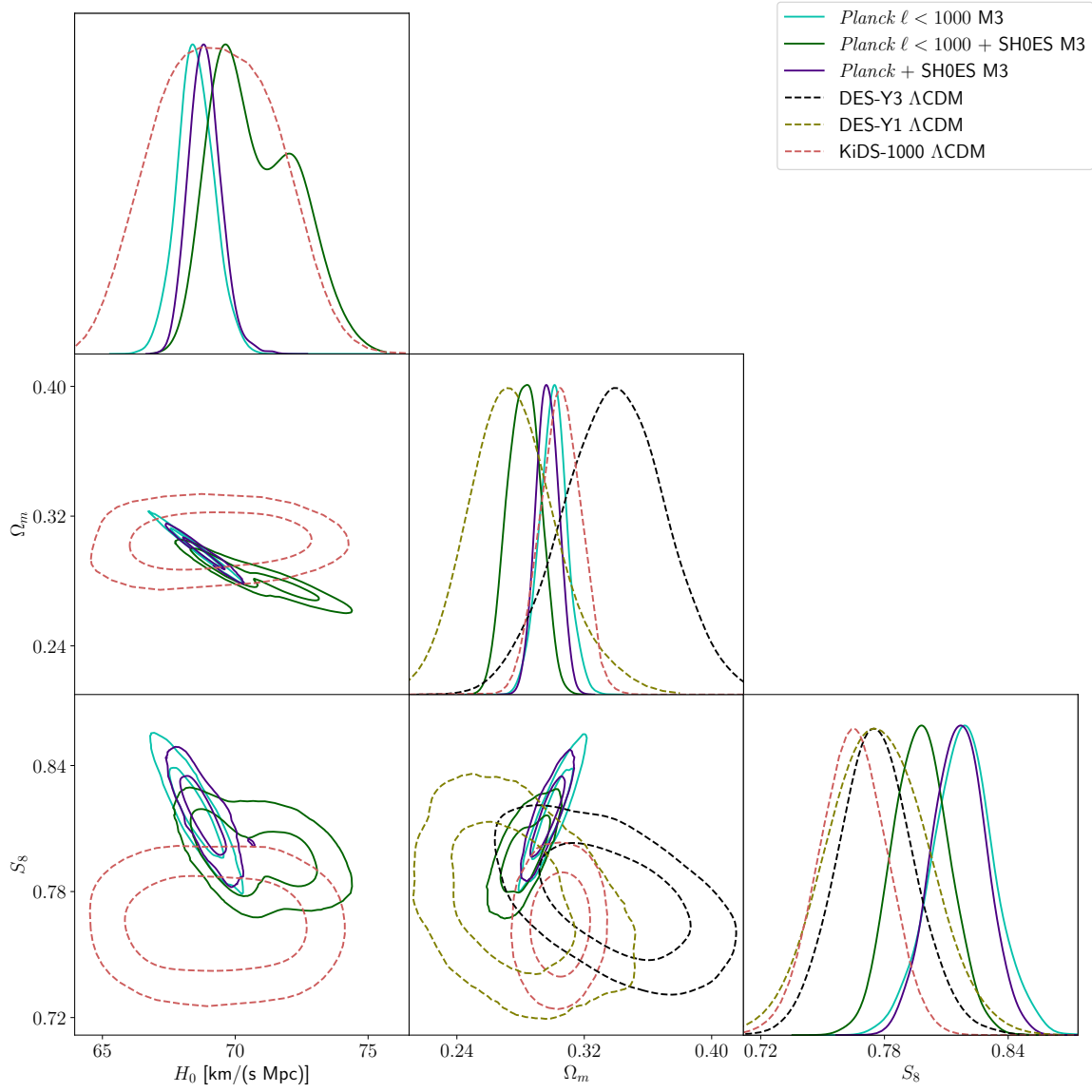


Figure 4.12: Confidence regions and one-dimensional posteriors for H_0 , matter fraction Ω_m and rescaled clustering amplitude S_8 . *Planck* $\ell < 1000$ M3 has low clumping, therefore is close to Λ CDM and exhibits similar negative H_0 – Ω_m and H_0 – S_8 correlations, so increasing H_0 decreases both Ω_m and S_8 . *Planck* $\ell < 1000$ + SH0ES M3 can reach high clumping (see Figure 4.8), giving rise to even lower Ω_m and S_8 . High clumping is disfavored by the damping-tail data of full *Planck*+SH0ES, so the corresponding contours are closer to Λ CDM. We show the confidence regions and posteriors from DES-Y3 (DES Collaboration et al., 2021), DES-Y1 (Abbott et al., 2018) and KiDS-1000 (Heymans et al., 2021) for reference. DES data do not constrain H_0 significantly compared to their prior, which is flat in the range 55 to 91 km/(s Mpc).

albeit only marginally.

Standard-ruler BAO measurements constrain the combinations $d_M(z)/r_{\text{drag}}$ and $H(z)r_{\text{drag}}$ (Alam et al., 2017), where d_M is the comoving angular-diameter distance. In a flat Λ CDM cosmology,

$$d_M = \int_0^z \frac{cdz}{H(z)}, \quad (4.9)$$

and the dominant contributors are the cosmological constant and the nonrelativistic matter, since for low z

$$H(z) \approx H_0 \left(1 + \Omega_m \left[(1+z)^3 - 1 \right] \right)^{1/2}.$$

For a constant Ω_m , both $d_M(z)/r_{\text{drag}}$ and $H(z)r_{\text{drag}}$ depend only on $H_0 r_{\text{drag}}$. On the CMB side, the sound-horizon angular scale θ_s is also proportional to $H_0 r_{\text{drag}}$ for fixed Ω_m , since r_{drag} is very close to r_S [see Equation (4.6)]. The early integrated Sachs-Wolfe effect in the CMB determines the physical matter density $\omega_m = \Omega_m h^2$, which indeed stays roughly constant in our sampling. Therefore, as H_0 increases, Ω_m decreases, which causes changes in $d_M(z)/r_{\text{drag}}$ and $H(z)r_{\text{drag}}$ at low redshift, compared to their high-redshift analog θ_s (effectively fixed by the CMB). As a consequence, clumping models that fit the CMB develop a tension with BAO data (see Jedamzik et al. (2021) for a broader discussion).

As an example, in Figure 4.13 we show the relative difference between these quantities for Λ CDM (best fit to *Planck*) and for M3 (best fit to *Planck*+SH0ES), and overlay the SDSS DR12 measurements (Alam et al., 2017). By eye, we can tell that the *Planck*+SH0ES M3 best fit is mildly disfavored by the transversal $[d_M(z)/r_{\text{drag}}]$ BAO data, while for the radial $[H(z)r_{\text{drag}}]$ BAO the data scatter is too large to tell. Using the full covariance matrix, we find $\Delta\chi_{\text{BAO}}^2 \approx -3.6$, indeed mildly disfavoring M3.

At higher redshifts, the relative difference in $d_M(z)/r_{\text{drag}}$ tends to 0—so as to match $\approx 1/\theta_s$ to the CMB at recombination. The $H(z)r_{\text{drag}}$ relative difference, on the other hand, tends to a negative constant—since ω_m changes weakly, expansion rate at high redshifts is almost the same, so the difference is driven by the change in sound horizon (and thus drag scale), which is $\approx 1\%$ for the *Planck*+SH0ES M3 best fit compared to *Planck* Λ CDM best fit.

While the M3 best fit to *Planck*+SH0ES is in mild tension with the current BAO measure-

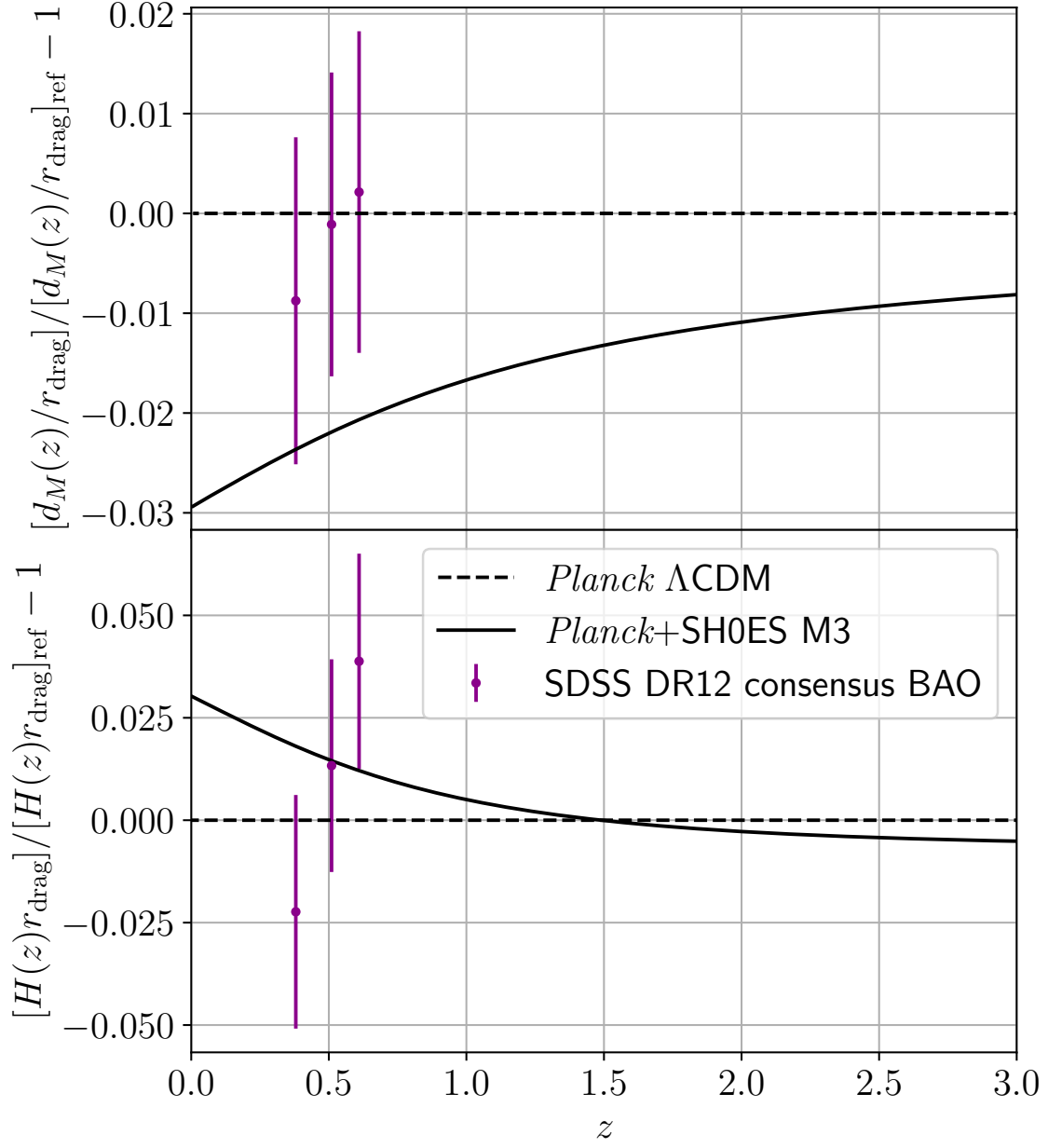


Figure 4.13: Relative difference in the BAO distances— $d_M(z)/r_{\text{drag}}$ and $H(z)r_{\text{drag}}$, for M3 best fit to *Planck*+SH0ES, compared to the Λ CDM best fit to *Planck* (see Section B.3 and Table B.1 for exact parameters) SDSS DR12 measurements (Alam et al., 2017) are overlaid. At higher redshifts, the relative difference in d_M/r_{drag} tends to 0 and in Hr_{drag} —to a negative constant ≈ -0.01 .

ments, that does not necessarily prove that increasing H_0 in M3 is always disfavored by BAO data. There remains a possibility that model parameters can be adjusted to accommodate the datasets and provide a better joint fit. To assess this, we plot $d_M(z = 0.51)/r_s$ versus H_0 for our *Planck* $\ell < 1000$ + SH0ES M3 run in Figure 4.14, overlaying the SDSS DR12 measurement. The upper left dots have low clumping, so they follow the standard Λ CDM degeneracy direction. The right dots, with high clumping, follow a different trend, but still develop more and more tension with SDSS for increasing H_0 , as explained in Jedamzik et al. (2021). We remind the reader that this figure shows only one of three SDSS D12 measurements, and the trend is similar in all, which makes the tension stronger.

Standard-ruler BAO data are being improved, and in the future it will become more decisive for or against the clumping model. A notable example is the *Dark Energy Spectroscopic Instrument* (DESI), which is already operational. DESI is expected to provide subpercent precision measurements of d_M/r_{drag} in seven bins for $0.65 \leq z \leq 1.25$ (DESI Collaboration et al., 2016a), which alone can give $\Delta\chi^2 \approx 20$ between our best-fit models. At higher redshifts, 21-cm data will provide a measurement of $H r_{\text{drag}}$ to percent-level precision using on velocity-induced acoustic oscillations (VAOs) (Muñoz, 2019). Our clumping model predicts only a modest deviation of the radial $H r_{\text{drag}}$, so transverse BAO measurements have more constraining power.

We note that the change in recombination induced by small-scale clumping is likely to affect the shape of BAO fitting templates, and henceforth the distance-scale extraction from observational data. A proper analysis of BAO data within our M3 model should check whether $d_M(z)/r_{\text{drag}}$ and $H(z)r_{\text{drag}}$ are recovered without any biases compared to standard extraction procedures. A quick test with $z = 0$ correlation functions in Figure 4.15 shows that change in hr_{drag} overwhelms the possible bias in drag scale reconstruction. We leave a detailed study of the correlation function in the presence of clumping for future work.

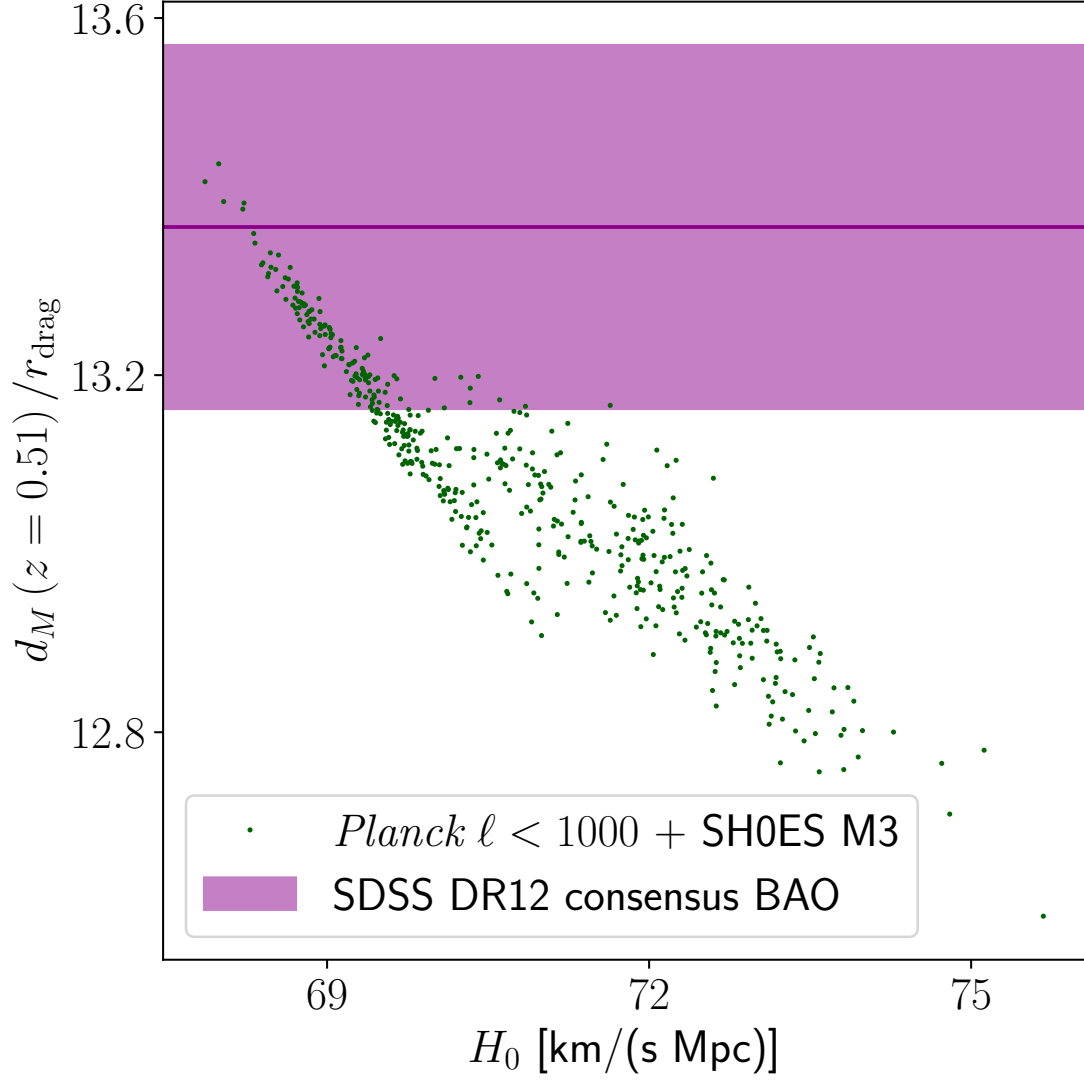


Figure 4.14: Scatter plot of $d_M(z = 0.51)/r_s$ versus H_0 for our $Planck \ell < 1000 + SH0ES M3$ run, which exhibits the highest clumping. We overplot the SDSS DR12 measurements at this redshift in magenta, where the semitransparent band shows its 1σ error bar. Even though high clumping (in the right) produces higher H_0 than inferred from the Λ CDM degeneracy line (in the upper left), larger values of H_0 still cause tension between CMB and the BAO.

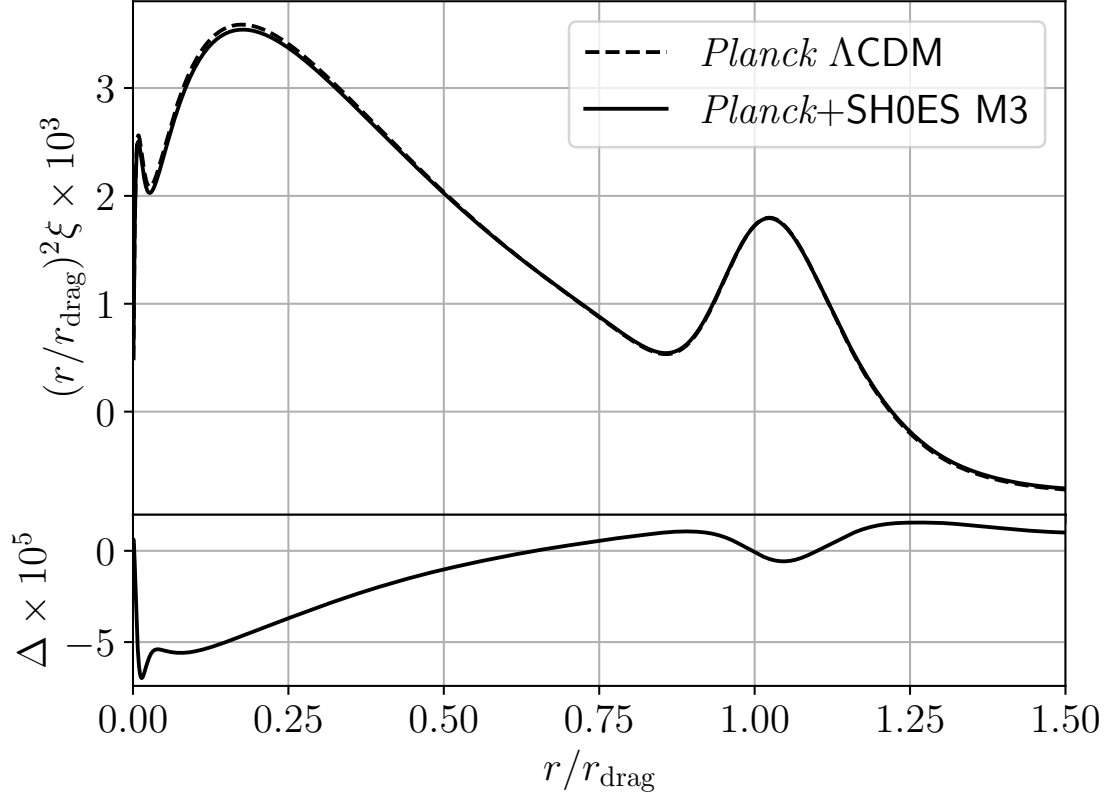


Figure 4.15: Correlation functions at $z = 0$ for ΛCDM (best fit to *Planck*) and for M3 (best fit to *Planck*+SH0ES), rescaled with each baryon drag scale, which gives almost perfect agreement. The lower panel shows the difference between the two lines magnified by an additional factor of 100. The comoving drag scale r_{drag} decreases by 1% between the two cases, but h increases by 4%, so that the drag scale in Mpc/h units hr_{drag} increases by 3%. Such change is clearly noticeable and overwhelms any possible bias arising from template-shape differences.

4.5 Forecasts for future CMB experiments

Having exploited the current data at our disposal, we now perform forecasts for two future CMB experiments: the *Simons Observatory* (SO, Ade et al. (2019)) and CMB-S4 (Abazajian et al., 2019), which will have much better damping-tail precision and therefore will be able to test the clumping model much better.

We have written mock likelihoods for these two experiments within Cobaya⁴ adapting the ones in MontePython (Audren et al., 2013; Brinckmann & Lesgourgues, 2019), and created the models for these two experiments using deproj0 noise curves for temperature and E-mode polarization fluctuations. We focus on primary CMB anisotropies, rather than lensing map.

	Standard	Clumping
best fit to	<i>Planck</i>	+SH0ES
δ_-	n/a (0)	−0.955
δ_+	n/a (0)	1.320
f_0	n/a (1)	0.652
b	n/a (0)	0.439
$10^9 A_s$	2.1094	2.1132
n_s	0.96604	0.96552
$100\theta_s$	1.04192	1.04177
$\Omega_b h^2$	0.022416	0.022714
$\Omega_{\text{cdm}} h^2$	0.11945	0.11999
τ_{reio}	0.0514	0.0542
H_0 [km/(s Mpc)]	68.146	70.916
Ω_K	0	
m_ν [eV]	0	

Table 4.4: Fiducial parameters for our forecasts.

Throughout this section we consider two fiducial power spectra, which bracket our current knowledge on clumping during recombination: for the first we assume standard recombination (Λ CDM) and take CMB + direct H_0 measurement, whereas for the second we choose a model with nonzero clumping and higher H_0 and consider only CMB. Full parameter sets are presented in Table 4.4. For each we perform model comparison between M3 and LCDM, and show posteriors for H_0 and the clumping parameter b . For the H_0 measurement, we assume

⁴https://github.com/misharash/cobaya_mock_cmb

SH0ES, though we have checked that an H_0 precision improvement to 1% will not change our conclusions.

4.5.1 Fiducial with standard recombination

We begin by considering the case that the CMB power spectra of SO/CMB-S4 continue to agree with the standard Λ CDM model [and a low $H_0 \approx 68.1$ km/(s Mpc), where the parameters are taken from our best fit to *Planck* data with massless neutrinos and presented in Table 4.4]. The question we address is whether the degeneracy between clumping and H_0 will still be able to bring future CMB experiments in closer agreement with a direct H_0 measurement in this case. Since our fiducial is Λ CDM with standard recombination, M3 can not fit the data any better, so model comparison on CMB-only data will not be informative. Therefore, in this subsection we consider only future CMB data added to SH0ES.

We show the model comparison between M3 and Λ CDM in Table 4.5. Unlike with *Planck*, SO/CMB-S4+SH0ES have negligible χ^2 improvement. Bayes factors K stay consistent with 1, indicating no clear preference between the models.

	$\Delta\chi^2_{best}$	$\log_{10} K$
<i>Planck</i> +SH0ES	5	0.14 ± 0.19
SO baseline+SH0ES	0	0.15 ± 0.22
CMB-S4+SH0ES	0	-0.02 ± 0.22

Table 4.5: Model comparison forecast with a standard recombination fiducial (see Section 4.5.1). All χ^2 differences are rounded to integers because of uncertainty in the minimizer output. If CMB data continues to be consistent with standard recombination and low H_0 , M3 will be not able to allow CMB to agree with a direct H_0 measurement. Bayes factors $-0.5 \lesssim \log_{10} K \lesssim 0.5$ tell no clear preference, as in Tables 4.2 and 4.3.

Figure 4.16 provides a closer look into the H_0 posterior. If future CMB power spectra continue to agree with Λ CDM, M3 does not allow any significant H_0 shift even with the pull from SH0ES. As expected, the increased CMB precision shifts the H_0 posterior to our input CMB fiducial of $H_0 \approx 68$ km/(s Mpc), shown as the gray dashed vertical line. The posterior for b is very close to its prior, except that high values $b \sim 1$ are disfavored (and thus does not allow any correlations with H_0), so we do not show it here.

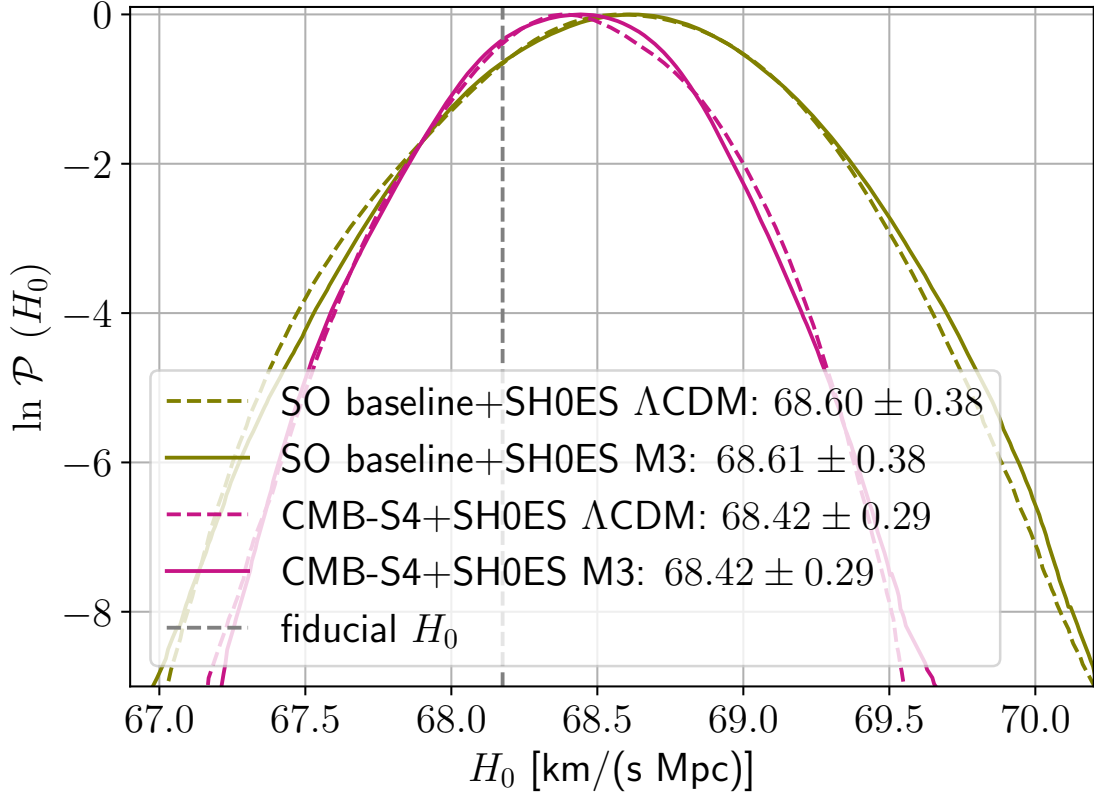


Figure 4.16: Posterior for H_0 in our forecasts with a standard recombination fiducial, in all cases with CMB + SH0ES. The solid lines assume M3, whereas dashed assume Λ CDM with standard recombination. The gray dashed line shows the input fiducial value of H_0 . While the maximum posterior values are shifted to higher H_0 with respect to the input, due to the addition of a direct H_0 measurement, the shift from Λ CDM to M3 is negligible for both experiments. We present 68% C.L. intervals on H_0 in the legend. For clumping (within M3) they are $\log_{10} b = -5.3^{+4.2}_{-3.7}$ for SO+SH0ES, $\log_{10} b = -5.8^{+3.7}_{-3.2}$ for CMB-S4+SH0ES (and $\log_{10} b = -5.5 \pm 3.0$ for prior).

4.5.2 Fiducial with clumping

We next investigate a case that in truth contains substantial small-scale clustering. We generate fiducial power spectra for SO/CMB-S4 following the best fit of M3 to *Planck*+SH0ES, which in particular has $b \approx 0.44$ and $H_0 \approx 70.9$ km/(s Mpc) (full parameter set in Table 4.4). In this case, our aim is to determine how clearly clumping could be discerned by future CMB data, without any direct H_0 measurements.

We present the results of our model comparison in Table 4.6. If there is such clumping in

	$\Delta\chi^2_{\text{best}}$	$\log_{10} K$
<i>Planck</i>	0	0.04 ± 0.15
SO baseline	21	1.57 ± 0.23
CMB-S4	44	5.54 ± 0.23

Table 4.6: Same as Table 4.5, but for a nonzero clumping fiducial (see Section 4.5.2). In this case, both experiments can show a clear preference for M3, given their large $\Delta\chi^2_{\text{best}}$. In terms of the Bayes factor K , the evidence against Λ CDM with standard recombination would be strong for SO ($1 < \log_{10} K < 2$), and decisive for CMB-S4 ($\log_{10} K > 2$) (Kass & Raftery, 1995).

CMB, SO data will show a clear preference for clumping model, and CMB-S4 will be even more decisive. This is the only case when the Bayes factor K is significantly different from 1. With SO data M3 model is deemed ~ 30 times more probable than Λ CDM (strong evidence in favor of M3), and with CMB-S4— $\sim 300,000$ (decisive evidence in favor of M3) (Kass & Raftery, 1995).

We show the posteriors of b and H_0 in Figure 4.17, where the difference in H_0 between standard and clumpy recombination is clear for both SO and CMB-S4. We note that the posteriors for both H_0 and b peak at lower values than our input fiducials, as lower clumping (and therefore lower H_0 for the same θ_s) is favored by the prior. Also note that all results in this subsection are based on CMB data only, without any direct H_0 measurements. It is clear that future CMB data alone will suffice to detect clumping, as the posterior becomes much better constrained (against the prior). For SO the lowest values ($b \lesssim 10^{-2}$) are clearly disfavored by the data, whereas for CMB-S4 the limits become only tighter.

4.6 Conclusions

The Hubble tension poses an increasingly challenging problem to the standard cosmological model. A possible solution is to alter recombination, for instance by adding small-scale baryon clumping, which allows higher H_0 values to be inferred from CMB data. We have studied whether our flexible clumping model M3, having three spatial zones with variable densities and volume fractions, can solve the tension.

We have found that

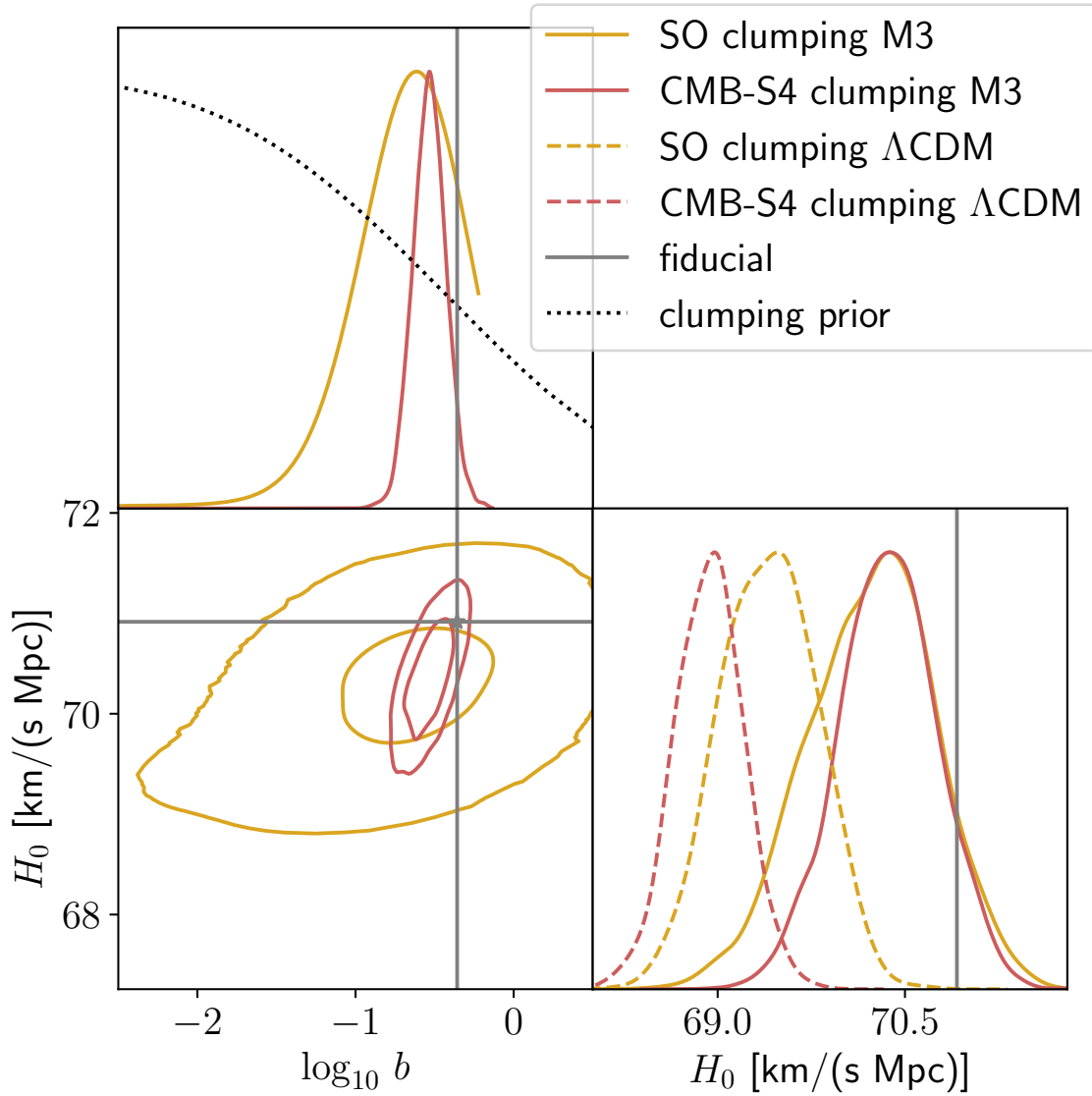


Figure 4.17: Posteriors of b and H_0 for forecasts with a clumping fiducial (best fit to current *Planck*+SH0ES, see Section 4.5.2). As is clear from the right panel, M3 and Λ CDM yield different H_0 values for the same data (both for SO and CMB-S4). In this case, the clumping parameter b will be measured to be nonzero at high significance. The 68% C.L. intervals on clumping within M3 are $\log_{10} b = -0.93^{+0.51}_{+0.16}$ for SO, $\log_{10} b = -0.53 \pm 0.11$ for CMB-S4 (and $\log_{10} b = -5.5 \pm 3.0$ for prior). Within M3, $H_0 = 70.23^{+0.54}_{-0.47}$ km/(s Mpc) for SO and $H_0 = (70.36 \pm 0.39)$ km/(s Mpc) for CMB-S4.

- (i) Current *Planck* data does not prefer clumping, even when adding the local H_0 measurement from the SH0ES Collaboration.
- (ii) Including only $\ell < 1000$ multipoles, *Planck* data allow for a larger shift to higher values

of H_0 , as the damping tail is more weakly constrained.

- (iii) Increasing H_0 within Λ CDM decreases both Ω_m and S_8 . The clumping model M3 follows the same trend, which relieves the potential S_8 tension with weak-lensing data.
- (iv) However, the same change of Ω_m is in tension with BAO standard-ruler measurements at low z . We showed that the BAO template is largely unaltered in the presence of clumping.

We have made forecasts for two future CMB experiments—*Simons Observatory* and CMB-S4—which will better measure the damping tail. First, we have found that if the power spectra stay consistent with Λ CDM (i.e., with standard recombination), increasing H_0 via clumping is strongly disfavored. Second, we have shown that the current best-fit model to *Planck*+SH0ES with clumping ($b \approx 0.4$) can be detected at high significance based solely on future CMB data. Therefore, future CMB experiments will provide considerable diagnostic power to investigate small-scale clumping at the epoch of recombination and shed light onto possible solutions to the H_0 tension.

4.6.1 Subsequent developments

After we published this study (Rashkovetskyi et al., 2021), there have been further interesting developments on the topic, which we summarize here. Galli et al. (2022) tested one of the Jedamzik & Pogosian (2020) models with data from the Atacama Cosmology Telescope (ACT, earlier considered by Thiele et al., 2021) and South Pole Telescope (SPT), finding no preference for clumping for CMB and some preference for CMB+SH0ES combination (yielding a middle value of H_0). They also made promising forecasts for the full SPT-3G survey, Simons Observatory, and CMB-S4. Lynch et al. (2024a) considered more general parametrized variations in the recombination history and found they prefer a higher H_0 from CMB+BAO. Lynch et al. (2024b) used the DESI DR1 BAO measurements and found that preferred recombination histories resemble effects of small-scale clumping and they reduce the Hubble tension between CMB+BAO and SH0ES to 2σ . Mirpoorian et al. (2025b) suggested

a 4-parametric data-driven model for modified recombination, reported the Hubble tension reduced to below 2σ , improved fits to *Planck* CMB + DESI DR1 BAO and relief of S_8 tension. Jedamzik et al. (2025) report mild to moderate preference for the accurate recombination modifications in the primordial magnetic fields scenario. Finally, Mirpoorian et al. (2025a) very recently argued that the 4-parametric modified recombination model resolves the emerging tensions between DESI DR2 BAO, CMB and Type Ia supernovae better than the dynamic dark energy.

Acknowledgments

Julian B. Muñoz was funded through a Clay fellowship at the Smithsonian Astrophysical Observatory. DJE was partially supported by U.S. Department of Energy Grant No. DE-SC0013718 and as a Simons Foundation Investigator. Cora Dvorkin was partially supported by the Department of Energy (DOE) Grant No. DE-SC0020223.

Chapter 5

Galaxy selection using the thermal Sunyaev-Zeldovich effect (a way beyond 2-point)¹

5.1 Introduction

Modern cosmology is a high-precision science thanks to a rich variety of data collected. The large-scale structure of the Universe contributes a major portion of this data. Most important techniques include baryon acoustic oscillations distance measurements (e.g., DESI Collaboration et al., 2025b,a) and full-shape clustering analyses (DESI Collaboration et al., 2024b,c), both based on galaxy redshift surveys.

Looking into the future, it is important to determine priorities for the next spectroscopic surveys. The expansion towards higher redshifts is well motivated by the primordial Universe studies. However, observations pose new challenges as spectral features of galaxies redshift out of the near-visible range accessible to ground-based telescopes (Ferraro et al., 2022).

In contrast, measuring more galaxy spectra at lower redshifts is easier technically. We already have a dense sample in the DESI Bright Galaxy Survey (Hahn et al. (2023)). Only a

¹This chapter is work in progress, to be published as Rashkovetskyi et al. (2025a)

small portion of it is used for the BAO and full-shape analyses (DESI Collaboration et al., 2025f, 2024b, 2025a). In the future DESI upgrade, DESI-II, the density of the luminous red galaxy (LRG) sample will also increase greatly (Schlegel et al., 2022). However, the scientific case for such higher-density samples is not so clear and specific.

There are promising ideas for leveraging information from different galaxy environments that can benefit from dense samples. One is density-marked clustering, which assigns additional weight to galaxies based on a local density estimate at their location. It was originally introduced in White (2016) to enhance modified gravity tests, and was later tuned to give tighter constraints on neutrino mass (Massara et al., 2021) and other cosmological parameters (Massara et al., 2023, 2024). Another is density-split clustering, which splits galaxies into subsamples based on the local density estimate. It was introduced for better modeling of redshift-space distortions in Paillas et al. (2021). Later works predict tighter constraints on standard cosmological parameters and neutrino masses (Paillas et al., 2023), or primordial non-Gaussianity (Morawetz et al., 2024). A simulation-based model of density-split clustering was built (Cuesta-Lazaro et al., 2024) and applied to BOSS CMASS data (Paillas et al., 2024), surpassing the standard analysis.

A number of works demonstrated the benefit of combining galaxy redshift surveys with data on secondary cosmic microwave background (CMB) anisotropies, particularly lensing. For example, Kim et al. (2024); Sailer et al. (2024) cross-correlate lensing with spectroscopic galaxies to better constrain the growth of cosmic structure, and Krolewski et al. (2024); Bermejo-Clement et al. (2024) improve measurements of primordial non-Gaussianity. This combination of probes is very promising, because DESI is taking galaxy spectra at an unprecedented rate, and the next CMB experiments like Simons Observatory (Ade et al., 2019) and CMB-S4 (Abazajian et al., 2019; Carlstrom et al., 2019) have great prospects for secondary CMB anisotropies, including lensing and Sunyaev-Zeldovich effects.

We decided to use the thermal Sunyaev-Zeldovich (tSZ) effect. It is inverse Compton scattering of cosmic microwave background (CMB) photons on free thermal electrons moving randomly (whereas bulk motions give rise to the kinetic or kinematic Sunyaev-Zeldovich effect).

This process results in a net increase in energy of scattered photons and creates a distinct frequency-dependent distortion in the CMB spectrum. The relative change in photon energy is approximately equal to the Compton y parameter, which is proportional to the integral of free electron density n_e and temperature T_e along the path (with length element dr ; Planck Collaboration et al., 2016a; Sunyaev & Zeldovich, 1972):

$$y = \int \frac{k_B T_e}{m_e c^2} n_e \sigma_T dr, \quad (5.1)$$

other quantities are constants: electron rest mass energy $m_e c^2$, Boltzmann’s constant T_e and Thomson scattering cross-section σ_T . Accordingly, the effect is strongest in ionized, hot and dense gas in or around galaxy clusters (as suggested by Sunyaev & Zeldovich, 1970, 1980). The clusters represent a distinct environment.

The tSZ data comes from CMB telescopes independent of DESI, whereas density estimates depend on the galaxy observation strategies. Using different datasets would allow us to check their consistency, check for systematics and potentially discover new fundamental tensions. Specific improvements could include testing the conformity of different galaxy subtypes (e.g., Patej & Eisenstein, 2016), improving the redshift-space distortion modeling for full-shape clustering analyses by removing a small fraction of the strongest Fingers of God (Baleato Lizancos et al., 2025), obtaining multiple tracers with better-constrained biases for the primordial non-Gaussianity measurements (similarly to Sullivan et al., 2023) or constraining environmental effects on galaxy formation or galaxy-halo connection (e.g., Yuan et al., 2023).

We need to remark that Sunyaev-Zeldovich maps have been used for cluster studies: their detection, mass determination, and more (Planck Collaboration et al., 2016b; Hilton et al., 2021; Robertson et al., 2024; Bocquet et al., 2023; Kornoelje et al., 2025). But rigorously detected and confirmed cluster candidates are rare. We aim to extract more information from the lower signal-to-noise parts, which comprise a much bigger fraction of the map.

This work is not going to provide the final answers and methodology, but rather motivate further developments. It is structured as follows: Section 5.2 introduces the data and simulations we use, Section 5.3 details our processing of real data and provides its results,

Section 5.4 describes our simulation-based toy model, compares it with data and shows new insights, Section 5.5 concludes with a summary and a future outlook.

5.2 Data

5.2.1 Galaxy catalog: DESI DR1 LRG

The Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2016a, 2022) conducts a 5-year galaxy redshift survey. Its scientific program has been successfully validated (DESI Collaboration et al., 2024d) alongside the Early Data Release (DESI Collaboration et al., 2024e). The key results based on Data Release 1 (DESI Collaboration et al., 2025c, recently made public) include clustering catalogs and two-point statistics measurements (DESI Collaboration et al., 2024a), baryon acoustic oscillation (BAO) distance measurements from galaxies, quasars (DESI Collaboration et al., 2025f) and Lyman- α forest (DESI Collaboration et al., 2025d) along with their detailed implications for cosmological models (DESI Collaboration et al., 2025e); and full-shape clustering of galaxies and quasars (DESI Collaboration et al., 2024b) with their cosmological analysis (DESI Collaboration et al., 2024c). Furthermore, updated BAO measurements from Lyman- α (DESI Collaboration et al., 2025b), galaxies and quasars with accompanying cosmological interpretations (DESI Collaboration et al., 2025a) are also available, based on Data Release 2 (DESI Collaboration et al., 2026).

We use Luminous Red Galaxies (LRG; Zhou et al., 2023) from the DESI DR1 clustering catalog (DESI Collaboration et al., 2024a, 2025c). We chose LRG because their redshift range ($0.4 < z < 1.1$) covers the peak in redshift distribution of Sunyaev-Zeldovich clusters (Hilton et al., 2021). We discard galaxies at $z > 0.85$ because then both the LRG density (DESI Collaboration et al., 2024a) and the SZ cluster abundance (Hilton et al., 2021) drop.

5.2.2 Sunyaev-Zeldovich map: ACT DR6

The Atacama Cosmology Telescope (ACT; Fowler et al., 2007; Choi et al., 2020) was a ground-based cosmic microwave background (CMB) experiment. Compared to the space-

based *Planck* mission (Planck Collaboration et al., 2020a), it has a smaller sky fraction, but reaches small angular scales thanks to higher resolution, and measures the CMB polarization with higher precision due to lower instrumental noise. Both of these are very important for Sunyaev-Zeldovich analyses.

We use the thermal Sunyaev-Zeldovich Compton y parameter map from ACT DR6 (Coulton et al., 2024). We rely on the accompanying noise simulations (Atkins et al., 2023) to estimate the signal-to-noise ratio in our analysis.

5.2.3 Simulations: ABACUSSUMMIT halo catalogs

We also use a $z = 0.8$ snapshot of a $(2h^{-1}\text{Gpc})^3$ cubic box (for the fiducial cosmology) from the ABACUSSUMMIT suite of N -body simulations (Maksimova et al., 2021) produced with ABACUS code (Garrison et al., 2021). The halos have been identified with the COMPASO halo finder (Hadzhiyska et al., 2022). We use a galaxy catalog produced within the halo occupation distribution (HOD) galaxy-halo connection framework, efficiently implemented in ABACUSHOD (Yuan et al., 2022) with parameters based on Yuan et al. (2023). This galaxy catalog was one of the base cubic boxes for **Abacus-2** cut-sky mocks described in DESI Collaboration et al. (2025f).

5.3 Measurements

5.3.1 Methodology and challenges

We divide the LRGs into subsamples (“SNR bins”) according to the signal-to-noise ratio at their location in the ACT DR6 + *Planck* y map (Coulton et al., 2024). First, we match the positions of the LRGs from the DESI clustering catalog to the pixels in the map. Second, we compute the pixel-level standard deviations in y using the corresponding 304 Gaussian noise simulation (Atkins et al., 2023). Finally, we compute the signal-to-noise ratio in each pixel by dividing the y value by its standard deviation.

Such external selection of galaxies imposes complex geometry changes that affect the

estimation of the clustering statistics. E.g., the Landy-Szalay estimator (Landy & Szalay, 1993) for the correlation function between samples 1 and 2 (which may be identical or different):

$$\hat{\xi}_{12}^{\text{LS}} = \frac{D_1 D_2 - D_1 R_2 - R_1 D_2}{R_1 R_2} + 1, \quad (5.2)$$

where $D_1 D_2$ are the (binned) pair counts between data (galaxies) from sample 1 and data 2, $D_1 R_2$ are the pair counts between data 1 and random points (reflecting the survey geometry and selection etc) for sample 2, $R_1 D_2$ are between random points 1 and data 2, and $R_1 R_2$ are between random points 1 and 2.

The key to the problem is that the full-survey random catalogs are not representative of our subsamples. One might think that imposing the same on-sky position filter on random points would be a solution, however, it is not perfect. The resulting randoms would be (narrow) columns extended along the line of sight, whereas a significant portion of selected galaxies would be truly clumped around the cluster position in 3 dimensions.

It is possible to leave the unusually-looking clustering measurement and model it consistently. But we have decided to keep the correlation functions more intuitive. To achieve this, we avoid the issue with the subsample geometry by employing an asymmetric Davis-Peebles estimator (Davis & Peebles, 1983) for the correlation function:

$$\hat{\xi}_{12}^{\text{DP}} = \frac{D_1 D_2}{D_1 R_2} - 1. \quad (5.3)$$

The $D_1 D_2$ and $D_1 R_2$ are the same pair counts as in the Landy-Szalay estimator (Equation (5.2)). However, we note that the randoms representing sample 1 are not required: neither $R_1 D_2$ nor $R_1 R_2$ are involved. The randoms are only necessary for sample 2. Therefore, with this estimator, we can compute the correlation function between the tSZ subsample (as 1) and the full LRG sample (as 2).

We also apply filters to the y map for two reasons. First, a filter matched to a cluster SZ profile would optimally detect them and recover their positions (Turin, 1960). Second, the environmental influence of a cluster is likely to extend further than its Compton parameter profile. Our fiducial filter is not thoroughly matched, but has a Gaussian shape with a 2.4

arcmin full width at the half-maximum for simplicity. The single filter in the ACT DR5 catalog paper has the same scale (Hilton et al., 2021). We apply the same filter to all the noise simulations to obtain the corresponding pixel-wise standard deviation map.

5.3.2 Larger-scale clustering and galaxy bias

First, we compute the larger-scale cross-correlation functions between the different SNR bins and the full LRG sample. We use the radial (s) and angular (μ) binning. We disregard the pairs located too closely on the sky² for two reasons. The first is to avoid using pairs of galaxies belonging to the same SZ pixel, or associated with the same line of sight smeared by the beam and the filter. The second is to mitigate the DESI fiber assignment incompleteness effects (Pinon et al., 2025). We only count pairs with the angular separation above 0.1 degrees (6 arcmin)³.

We show the resulting isotropic cross-correlation functions in Figure 5.1 with covariances estimated using the jackknife technique. We see a significant clustering enhancement in the monopole as the tSZ detection level increases, even though we stop at 4σ , which was the threshold for cluster candidates in Hilton et al. (2021)⁴. This almost certainly corresponds to an increase in galaxy bias. 0 to 1 σ subsample appears very similar to the full LRG sample (as will continue to hold in many other respects). For reference, Table 5.1 provides the number of galaxies in each SNR bin.

We should remark on the negative signal we obtain, contradicting the theoretical definition of the tSZ Compton y parameter (Equation (5.1)). Instrumental (or atmospheric) noise plays a certain part. Kinematic Sunyaev-Zeldovich effect, resulting from the bulk motions of free electrons, can have different signs and can become partially confused with “negative” thermal effect. Other components, like the cosmic infrared background, may also leak into the thermal

²This is technically known as the theta (θ) cut.

³This leaves $\mu \approx 1$ bins with no pair counts and undefined correlation functions. We discard these bins and average over the remaining bins to obtain the correlation function monopole.

⁴However, bear in mind the simplicity of our filter, the matched filters of Hilton et al. (2021) should be more optimal.

Sunyaev-Zeldovich reconstruction, potentially adding negative signal (Coulton et al., 2024).

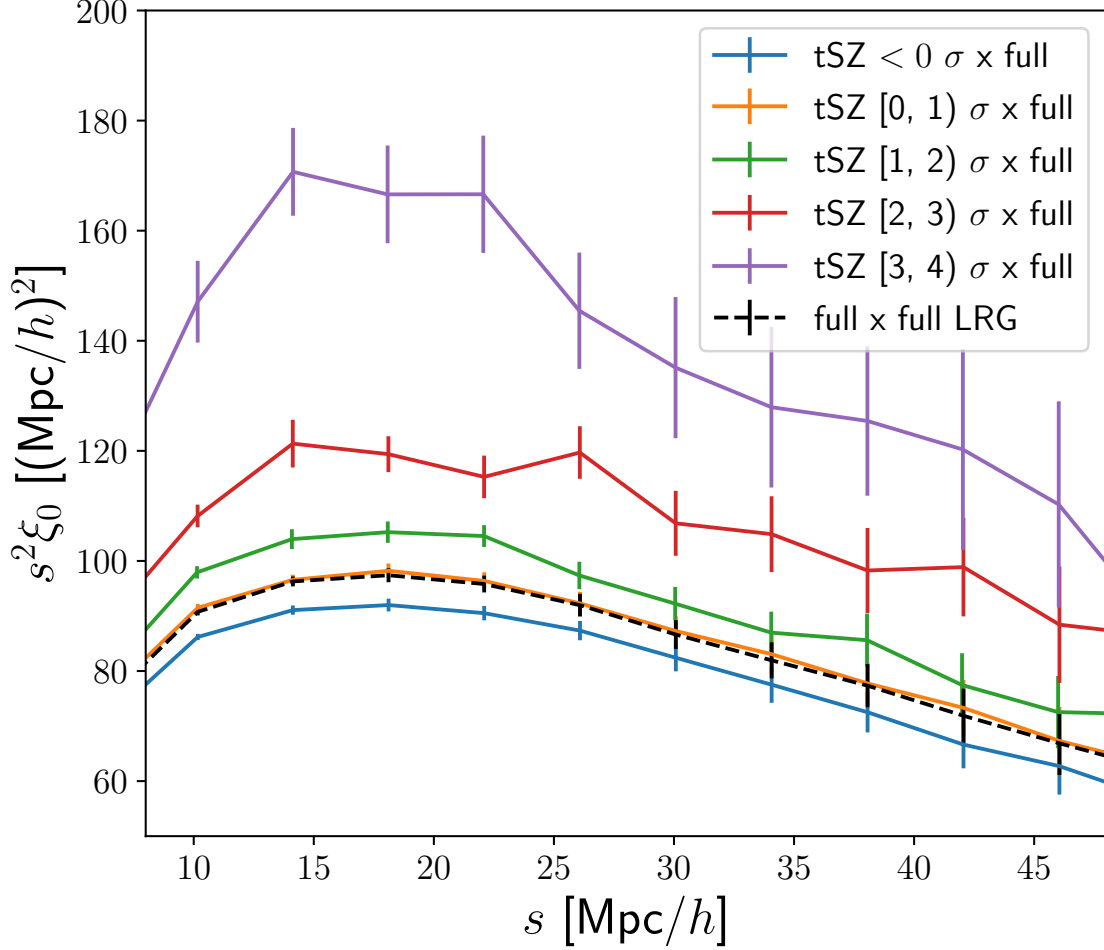


Figure 5.1: Large-scale isotropic (monopole) cross-correlation functions of different SNR bins with the full LRG sample. (Hereafter, we apply our fiducial Gaussian filter to the tSZ y parameter map, and use the jackknife technique to estimate errorbars.) There is a significant clustering enhancement with increasing tSZ detection level even below the threshold for cluster candidates (4σ in Hilton et al., 2021).

For clarity, we seek to summarize each line in Figure 5.1 with a single number. A simple way is to find the scaling for the full LRG correlation function to best match the cross-correlation function with any given signal-to-noise bin. This multiplier should be similar to the ratio of the linear galaxy biases between the SNR subsample and the full LRG sample. Accordingly, we also refer to this scaling as relative galaxy bias.

SNR bin	N_{LRG}
$(-\infty, -2) \sigma$	25101
$[-2, -1) \sigma$	130331
$[-1, 0) \sigma$	308498
$[0, 1) \sigma$	305819
$[1, 2) \sigma$	127875
$[2, 3) \sigma$	25469
$[3, 4) \sigma$	3577
$[4, 5) \sigma$	981
$[5, 6) \sigma$	484
$[6, \infty) \sigma$	676
Total	928811

Table 5.1: Number of DESI DR1 LRGs ($0.4 < z < 0.85$ and in the overlap with ACT footprint) in different tSZ SNR bins.

The jackknife covariance estimates for each correlation function allow us to estimate the precision of the scaling. However, we have not estimated the covariances between different cross-correlation functions shown in Figure 5.1 consistently. This makes our relative bias errorbar estimates imperfect and approximate.

We show the resulting galaxy biases of our subsamples (relative to the full LRG sample) in Figure 5.2. The picture is consistent with the conclusions we have drawn from Figure 5.1: the increase of galaxy bias with the tSZ detection level, and the 0 to 1 σ bin being close to average over all LRGs. In addition, we can inspect more categories at negative signal-to-noise, different from the full LRG but not so significantly deviating from each other.

5.3.3 Small-scale line-of-sight clustering and velocity dispersions

The cluster (supercluster) peculiar velocities are most apparent in configuration space at small scales near the line of sight. Here we switch from s and μ to the line-of-sight (π) and the perpendicular (on-sky, r_p) separations between the galaxy pair members.

Accordingly, we plot the line-of-sight small-scale correlation functions for our SNR bins in Figure 5.3. We notice an increase in correlation functions with the tSZ detection level in each bin, similar to the trends in larger-scale correlation functions (Figure 5.1). More interestingly, the slopes of the curves are also noticeably different, although it is hard to judge

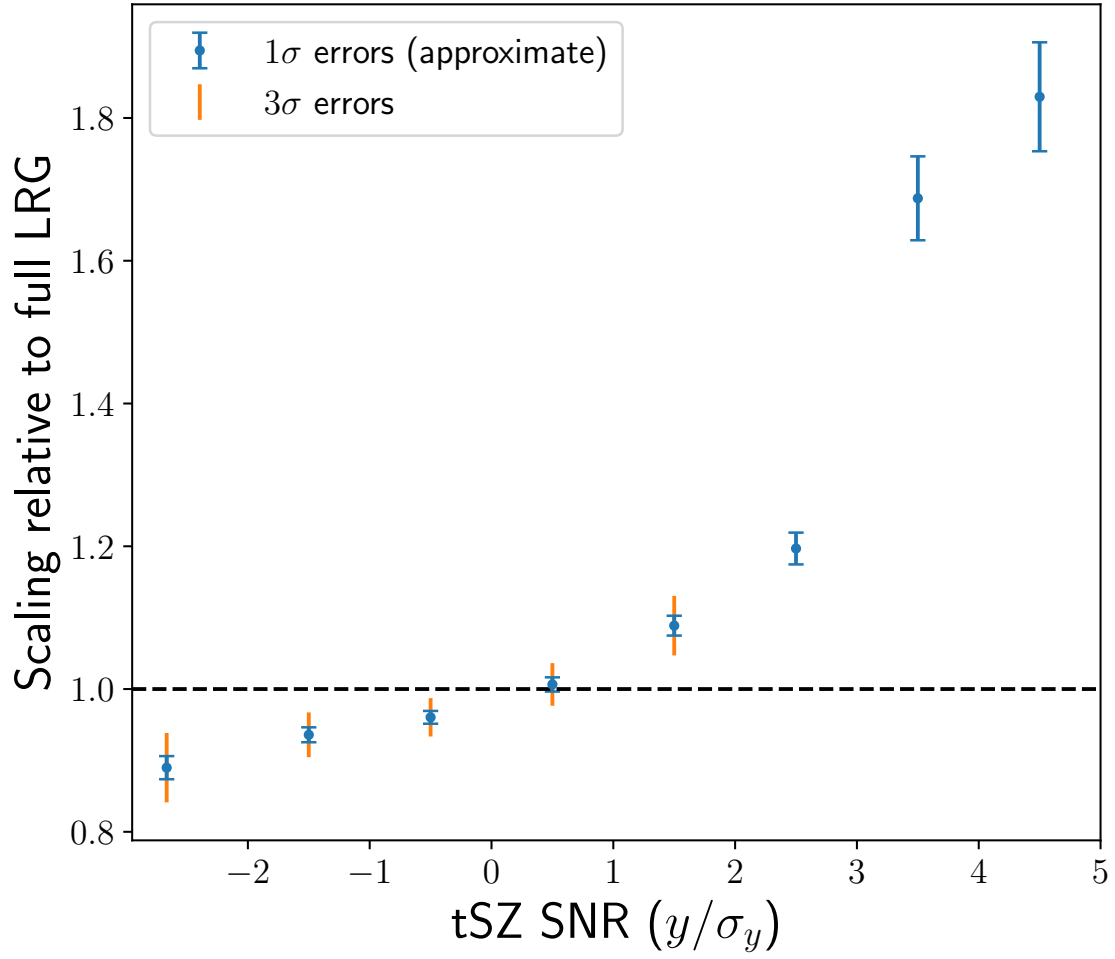


Figure 5.2: Galaxy bias of different SNR bins relative to the full LRG sample. The bias increase (clustering enhancement) consistent with Figure 5.1 can be seen more concisely.

the significance of these differences visually.

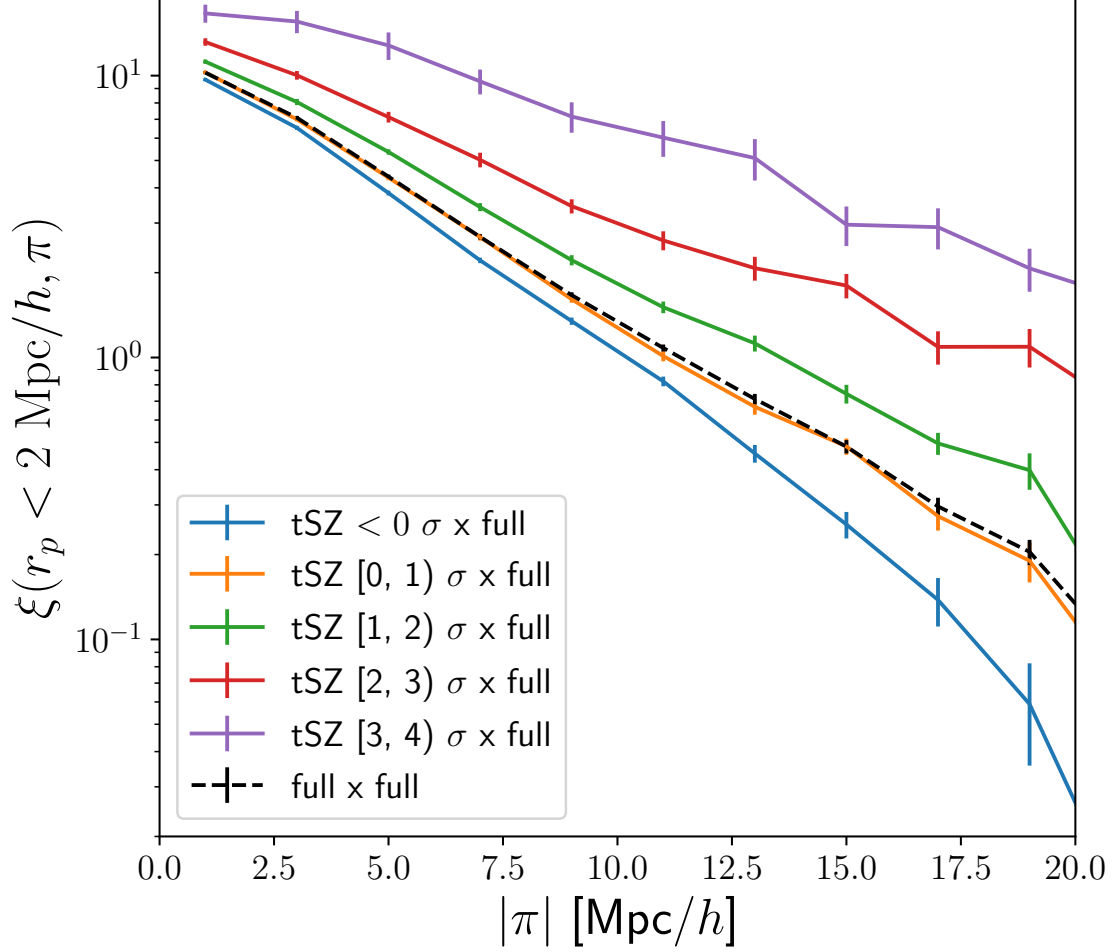


Figure 5.3: Small-scale line-of-sight cross-correlation functions of different SNR bins with the full LRG sample. There is not only an increase in amplitude, but also a flattening of the slopes of the curves with increasing tSZ SNR, indicating regions with higher tSZ signal have hotter small-scale velocity dispersions.

To better quantify the slopes, we fit exponentials to the correlation functions:

$$\xi(\pi) \approx C \exp\left(-\sqrt{2}\frac{|\pi|}{\sigma_\pi}\right). \quad (5.4)$$

This is motivated by the exponential fit to the distributions of peculiar radial velocities (Peebles, 1976) and a physical Press–Schechter (Gaussian mixture) model (Sheth, 1996). σ_π gives the one-dimensional dispersion (standard deviation).

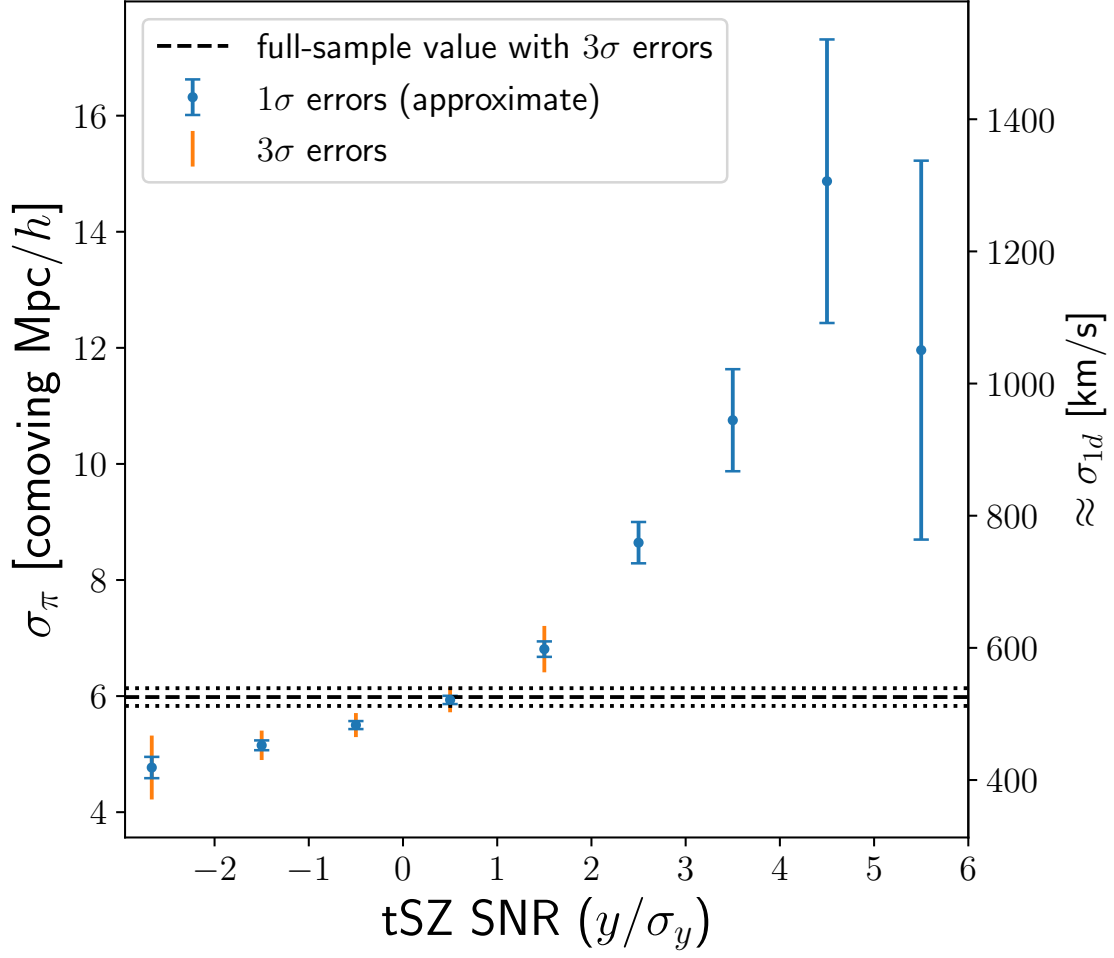


Figure 5.4: One-dimensional comoving coordinate dispersions σ_π from small-scale line-of-sight clustering. Approximate conversion to one-dimensional (line-of-sight) velocity dispersions σ_{1d} is provided on the right. The dispersion values increase with the signal-to-noise ratio, except the 5 to 6 σ_y bin.

We plot the fit results for each tSZ SNR bin in Figure 5.4. We measure them in comoving line-of-sight separations, but also provide an approximate conversion to line-of-sight peculiar velocities using the average redshift. The dispersions increase with the tSZ detection level, more significantly for positive values. The last, 5 to 6 σ_y bin, is an exception, although it has a large errorbar. The 0 to 1 σ_y bin is again very close to the average LRG.

Another natural way to summarize Figure 5.3 would be by integrating the correlation functions. This is directly related to the average number of any LRGs in narrow cylinders with axes along the line of sight centered on LRGs belonging in a particular SNR bin. We provide this summary in Figure 5.5. The average number of neighbors increases, in accordance with the expectation that LRGs with higher tSZ SNR tend to live in larger clusters.

We have found even more striking differences in the abundance of galaxies with many neighbors in different SNR bins. However, it is hard to visualize and interpret more meaningfully. We provide it in Figure C.1 and Section C.1 for posterity.

5.4 Simulation-based toy model

We aim to simulate both galaxy catalogs and y map consistently. Galaxy catalogs are often made from N -body simulations using a halo occupation distribution. The process has been highly optimized in ABACUSHOD (Yuan et al., 2022). Therefore we choose to make a Compton parameter map from a halo catalog.

We are aware of the advanced models for SZ-halo connection informed by hydrodynamic simulations or observations (e.g., Stein et al., 2020; Osato & Nagai, 2023; Liu et al., 2025). However, in this proof-of-concept work, we choose to develop a simpler model that is easier to control.

5.4.1 Simple y signal map from an N -body simulation

The Compton y parameter depends on the electron temperature $T_e(\mathbf{r})$ and number density $n_e(\mathbf{r})$ as functions of spatial position $\mathbf{r}$, whereas an N -body simulation provides masses and positions of discrete dark matter particles grouped in halos. Let us recall the theoretical

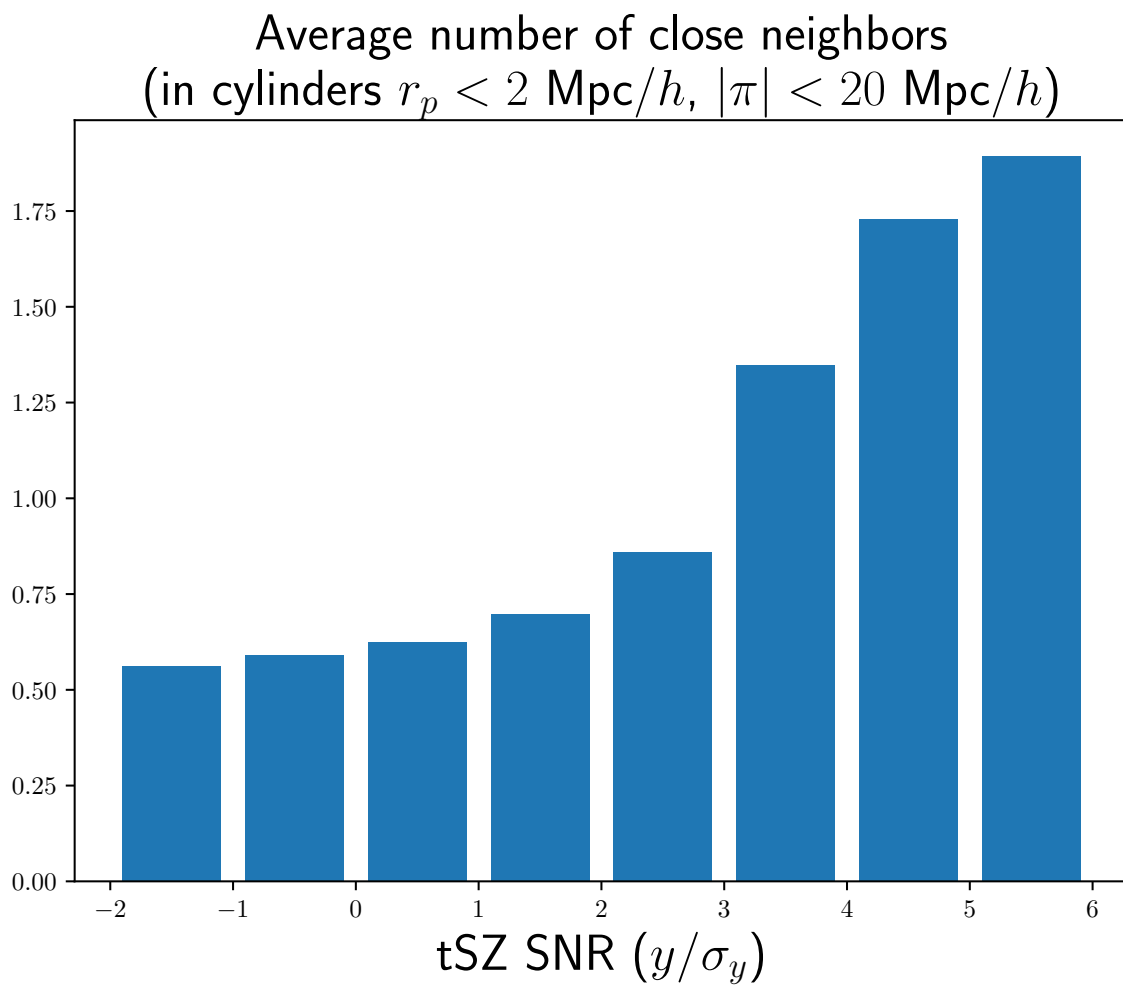


Figure 5.5: Average number of close neighbors in different tSZ SNR bins. It increases, confirming that stronger tSZ signatures are associated with clusters that have more galaxies.

expression for y as the integral along the line of sight (Planck Collaboration et al., 2016a; Sunyaev & Zeldovich, 1972):

$$y(\boldsymbol{\theta}) = \int_0^{r_{\text{ls}}} \frac{k_B T_e(r, \boldsymbol{\theta})}{m_e c^2} n_e(r, \boldsymbol{\theta}) \sigma_T dr. \quad (5.5)$$

The integral above is taken along the radial coordinate r to the last scattering surface r_{ls} , where CMB emerges. k_B is the Boltzmann's constant, $m_e c^2$ is the electron's rest energy and σ_T is the Thomson cross-section.

We estimate electron temperatures T_e using the motions of dark matter particles in halos. The root-mean-square thermal velocity is analogous to the random velocity dispersion of particles. The velocity dispersion of dark matter particles belonging to each halo is available in the ABACUSSUMMIT COMPASO halo catalogs. The remaining non-trivial step is connecting dark matter to hot gas. We assume complete ionization, primordial composition⁵, and thermal equilibrium between ions and electrons in the gas. We find adequate hot gas temperatures⁶ by taking the same kinetic energy of random motions per unit mass as for dark matter particles.

Then, we approximately compute the number of electrons N_e from the dark matter mass. First, we obtain the baryonic matter mass using the global baryon-to-dark matter ratio given by the cosmic density parameters $\Omega_b/\Omega_{\text{cdm}}$. Second, we estimate the hot gas mass by assuming it constitutes a constant fraction of all baryons $f_{\text{hot}} \approx 0.85$. Finally, we compute the electron number N_e assuming complete ionization of the hot gas and primordial composition as previously.

For simplicity, we use the flat-sky approximation valid for small angular scales. Accordingly, we switch from $\boldsymbol{r} = (r, \boldsymbol{\theta})$ decomposition to $(r_{\parallel}, \boldsymbol{r}_{\perp})$:

$$y(\boldsymbol{r}_{\perp}) \approx \frac{k_B \sigma_T}{m_e c^2} \int_0^{r_{\text{max}}} T_e(r_{\parallel}, \boldsymbol{r}_{\perp}) n_e(r_{\parallel}, \boldsymbol{r}_{\perp}) dr_{\parallel}. \quad (5.6)$$

Next, we need to account for particle discreteness. As the first rough approximation, we

⁵I.e., mix of hydrogen-1 and helium-4 with the mass fraction of the latter $Y_{\text{He}} = 0.2454$ (Planck Collaboration et al., 2020c).

⁶The 3-dimensional velocity dispersion $\sigma_{3d} \sim 200 \text{ km s}^{-1}$ yields $T \sim 10^6 \text{ K}$.

take infinitely narrow profiles of $T_e(\mathbf{r})n_e(\mathbf{r})$ for each dark matter particle with the 3-dimensional Dirac delta function $\delta^{(3)}$:

$$y(\mathbf{r}_\perp) \approx \frac{k_B \sigma_T}{m_e c^2} \sum_j \int_0^{r_{\max}} T_{e,j} N_{e,j} \delta^{(3)}(\mathbf{r} - \mathbf{r}_j) dr_\parallel. \quad (5.7)$$

Here j indexes dark matter particles with positions $\mathbf{r}_j$ and the corresponding electron temperature and number estimates. We can evaluate the integral over $r_\parallel$ by decomposing the delta function in $r_\parallel$ and $r_\perp$ ⁷:

$$y(\mathbf{r}_\perp) \approx \frac{k_B \sigma_T}{m_e c^2} \sum_j T_{e,j} N_{e,j} \delta^{(2)}(\mathbf{r}_\perp - \mathbf{r}_{\perp,j}). \quad (5.8)$$

In the following step, we need to discretize the y map into pixels. We can do this by averaging over pixel number k with area A_k :

$$y_k = \frac{1}{A_k} \int_{A_k} d\mathbf{r}_\perp y(\mathbf{r}_\perp). \quad (5.9)$$

Substituting Equation (5.8) into the integral above gives counts of dark matter particles in pixels weighted by their $T_e N_e$ along with constant factors:

$$y_k = \frac{1}{A_k} \frac{k_B \sigma_T}{m_e c^2} \sum_{j: \mathbf{r}_{\perp,j} \in A_k} T_{e,j} N_{e,j}. \quad (5.10)$$

Subsequently, we need to convolve the pixelated map with the point spread function (the CMB instrument beam). We use fine pixels at this stage. The common beam for the ACT DR6 + *Planck* map has a Gaussian shape with 1.6 arcmin full width at half-maximum (Coulton et al., 2024). This translates to $\approx 0.9h^{-1}\text{Mpc}$ at $z = 0.8$ — close to the typical halo size. Thus, the exact width of the electron temperature-density profile is not so important.

Finally, we downsample the map to larger pixels, approximately matching the larger pixel side of 0.5 arcmin in the ACT DR6 + *Planck* y map. Afterwards, we also apply a 2.4 arcmin FWHM Gaussian filter analogously to our data processing.

⁷Namely, $\delta^{(3)}(\mathbf{r} - \mathbf{r}_j) = \delta^{(1)}(r_\parallel - r_{\parallel,j}) \delta^{(2)}(\mathbf{r}_\perp - \mathbf{r}_{\perp,j})$. The integral of the former over $dr_\parallel$ gives 1 as long as $r_{\parallel,j}$ is not out of integration bounds.

5.4.2 Simple noise model

First, we have tried to add noise independently for each galaxy. We have divided the pure y signal produced in the previous section for every galaxy from a galaxy catalog by the y standard deviation. We take that as the center of the total signal-to-noise ratio distribution. We assume that the distribution is normal with a standard deviation of 1 (because the value is already normalized by the y standard deviation). Then, we have computed the contribution of each galaxy to each SNR bin by integrating each Gaussian within bin boundaries (this gives a difference of error functions, which can be evaluated quickly in common libraries). However, we have found that in this model, the clustering increase with SNR is much stronger than in the data.

Therefore, we implemented realistic spatial correlations in the noise map. We have done this by generating the noise (normalized by its pixel-wise standard deviations) according to the power spectrum of a noise simulation provided with the ACT DR6 products (accordingly normalized, and filtered⁸). We also produce a periodic square y standard deviation map, in which the same values occupy the same fractions of the area as for ACT DR6 + *Planck*. To obtain the total signal-to-noise ratio, we add the normalized noise to the ratio of the pure signal to the standard deviation, all taken from the pixel to which a mock galaxy falls.

By construction, our final mock noise map is independent of the mock signal. This might not be true in actual data processing (deprojection of different foregrounds) and may account for some of the differences we notice later.

5.4.3 Calibration

To refine our arbitrary coefficients in the y signal making described above, we rescale the signal to match the 3-dimensional density of $> 4\sigma$ detections (cluster candidates) in the data ($0.4 < z < 0.85$). This should be good as long as there is rarely more than one significant SZ cluster per pixel. Otherwise, we would need to match depth with the real Universe, which

⁸We applied the filter after normalization by pixel-wise standard deviations (computed previously from 304 realizations of noise simulations). Filtering before normalization corresponds to the data processing more exactly, but it creates small-scale artifacts at the boundaries of regions observed by ACT to different depths.

would be challenging with cut-sky mocks as well, because it would require bigger simulation boxes.

We match the 5711 peaks ($> 4\sigma$) detected in the ACT DR6 + *Planck* tSZ y map after our fiducial filtering to the ACT DR5 SZ cluster catalog (Hilton et al., 2021) containing 4195 confirmed objects. Because of the differences in data and cluster candidate detection methods, the matching is not 1-to-1. We restrict good matches to 2285 unique clusters within 2.4 arcmin of our peaks, among which 1179 belong to our redshift range ($0.4 < z < 0.85$). Given the sky area of the ACT DR5 cluster search mask, their average density is $\approx 40(h^{-1}\text{Gpc})^{-3}$. From this, we extrapolate the density of our $> 4\sigma$ peaks as $\approx 100(h^{-1}\text{Gpc})^{-3}$. We rescale the mock y signal to match this density of $> 4\sigma$ peaks in our first map (given that our simulation box is $(2h^{-1}\text{Gpc})^3$). The resulting signal rescaling factor is ≈ 1.14 , which suggests that our rough assumptions were not far off.

It is important to take noise into account in this calibration: without noise, we find the rescaling factor ≈ 1.59 . We do not assume that the noise distribution is asymmetric, but higher signals are rarer⁹. Therefore, at a given final signal-to-noise¹⁰, more objects were shifted “up” by noise (have a smaller true signal) than “down” (from a larger true signal).

5.4.4 Mock clustering

Bringing together all the ingredients, we can finally compare the clustering of SNR bins in mocks to data. We focus on the larger scales, showing full cross-correlation functions in Figure 5.6 and relative galaxy bias in Figure 5.7. We see a clustering/bias enhancement trend generally similar to the data, although the increase appears to be slower near zero signal and faster at high signal. As we mentioned before, one of the reasons for the difference may be the independence of noise and signal in our mocks.

Our next step will be to investigate how the differences in the clustering of SNR bins react

⁹That said, true negative signals are not possible in our simplistic model, and are probably hard to achieve in reality.

¹⁰At least a positive value.

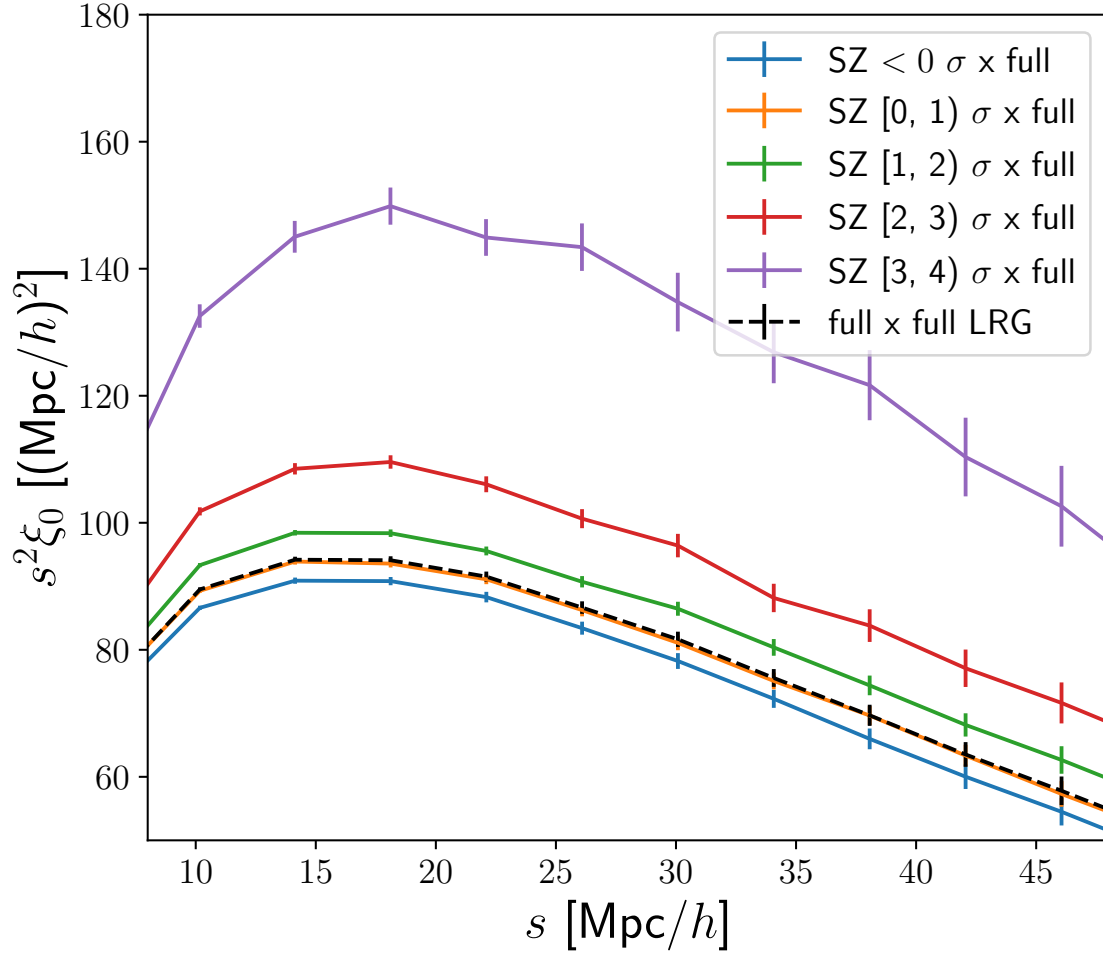


Figure 5.6: Mock large-scale cross-correlation function of different SNR bins with the full LRG sample. Similarly to the data (Figure 5.1), the clustering is enhanced at higher SNR, but more slowly. Also note that the $[0, 1)$ σ bin remains very similar to the full sample.

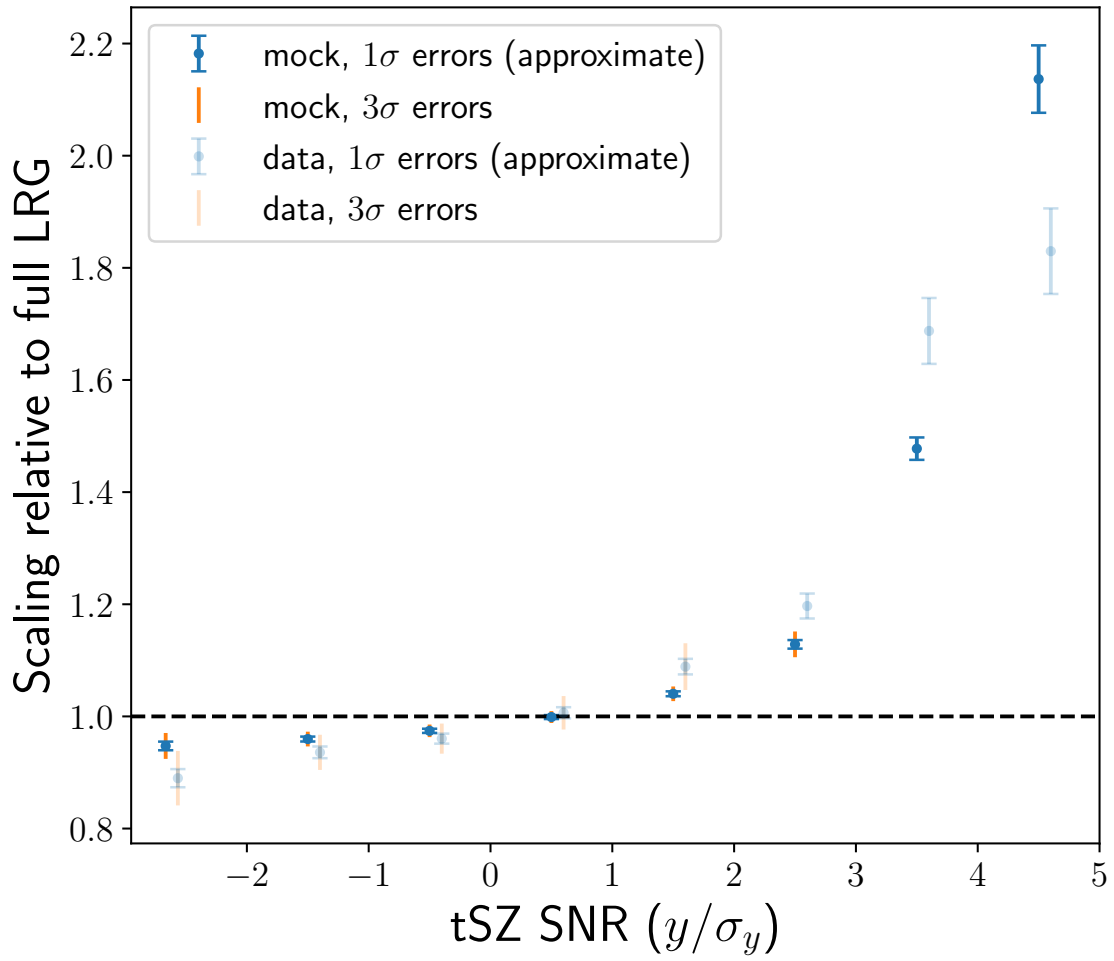


Figure 5.7: Mock galaxy bias of different SNR bins relative to the full LRG sample. Similarly to data (shaded and shifted slightly to the right for clarity), the relative bias increases with the SNR, but less rapidly near 0 and then more steeply at higher values.

to changes in halo occupation distribution parameters. Doing this with all cross-correlation functions (Figure 5.6) is hardly viable, because it would add even more lines to the plot, making it hard to interpret. The relative bias plot (Figure 5.7) provides a cleaner way forward.

Currently, we do not focus on the small-scale line-of-sight clustering (analogs of Figures 5.3 and 5.4). One of the reasons is that we do not model fiber assignment in our cubic mocks for simplicity. On large scales, the fiber assignment incompleteness effects can be mitigated quite well with weights (Bianchi et al., 2024; DESI Collaboration et al., 2024a), but on small scales it is a much more difficult problem.

5.4.5 Relating halo mass to thermal Sunyaev-Zeldovich signatures

Simulations also allow us to access the true halo properties, unlike the data. With this, we can explore connections with SZ properties (bearing in mind they are produced approximately).

We present mock correlations between halo mass and the noiseless tSZ y signal (left) or the final tSZ signal-to-noise ratio (right; both in their central pixels) in Figure 5.8. We count halo contributions by their mass¹¹ to minimize assumptions about the galaxy-halo connection. When we have better constraints on the halo occupation distribution, we can count the number of galaxies, which is non-linear in halo mass. We see a wide range of tSZ observables for smaller halo masses, but the most massive halos can be identified with the highest signal or signal-to-noise. Furthermore, the noiseless signal might have a power-law relation with halo mass, whereas the smearing towards smaller halo masses might be caused by such halos appearing in the same lines of sight as larger halos (within the point spread function).

A well-known or calibrated halo mass to tSZ observables relation can be used to make theoretical predictions about clustering, particularly on larger scales. Furthermore, if this relation is less dependent on cosmology than halo mass-density relations, it would enable more robust cosmological inference beyond the 2-point function.

Alternatively, it is possible to build a simulation-based model after investing in a large number of high-quality mocks (like for density-split clustering in Cuesta-Lazaro et al., 2024).

¹¹In numbers, small halos strongly dominate, making such plots less informative in our opinion.

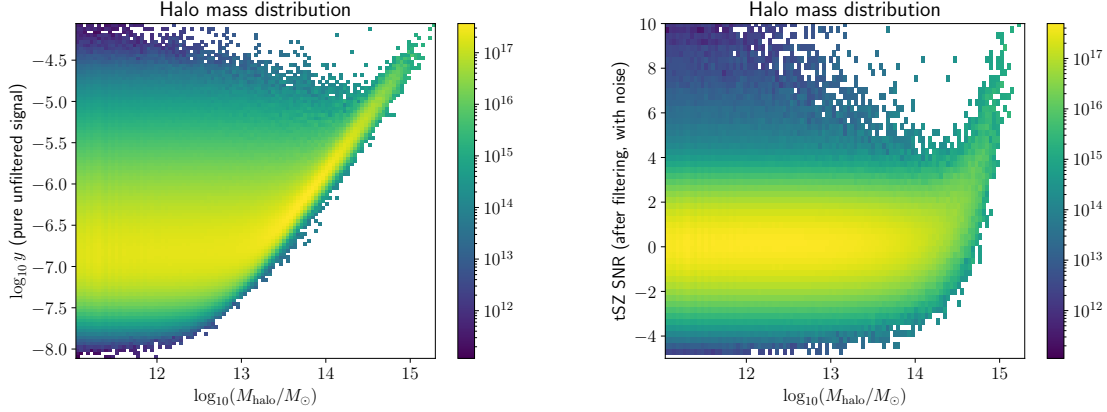


Figure 5.8: Histograms displaying correlations between halo mass and tSZ quantities in the corresponding pixels in our mock map. Left: noiseless y parameter. Right: SNR including our simplistic noise. Histograms are weighted by halo mass to minimize dependence on galaxy-halo connection assumptions.

This is a long way from our toy model. More generally, we are concerned about robustness, whether one can simulate SZ from first principles well enough.

5.5 Conclusions

We demonstrate that there is valuable cosmological information in the thermal Sunyaev-Zeldovich (tSZ) map in the “noise” beyond the individual cluster candidates. Soon, we aim to quantify the sensitivity of our effects to the galaxy-halo connection using the halo simulations. In further work, we are going to investigate constraining the growth of structure.

From large-scale clustering, we find that a higher tSZ signal-to-noise ratio corresponds to galaxies with a higher bias (Figure 5.2). This can allow us to select luminous red galaxy sub-samples and test their structural consistency (e.g., following Patej & Eisenstein, 2016), or better understand their biases to better constrain primordial non-Gaussianity (in the spirit of Sullivan et al., 2023).

From small-scale line-of-sight clustering, we find that the velocity dispersion increases considerably with tSZ SNR (Figure 5.4). This is likely a clean indication of strong non-perturbative non-linearities (Fingers of God) which can be removed to improve the theoretical

modeling (Baleato Lizancos et al., 2025).

We also see that galaxies with higher SNR bins have higher numbers of close neighbors (Figure 5.5). This could help inform the galaxy multiplet studies (e.g. Lamman et al., 2024).

We have built a simple mock tSZ map that allowed us to reproduce general trends from the data (Figures 5.6 and 5.7). However, we note differences, and investigating their causes will be our next priority.

Eventually, we hope to use a relatively small number of simulations to calibrate the relation between the halo mass and the Sunyaev-Zeldovich observables and build a semi-analytical model of galaxy subsample clustering. Alternatively, one can invest in a large number of accurate simulations with different cosmological and galaxy-halo connection parameters to build a simulation-based model or inference pipeline like for density-split (Cuesta-Lazaro et al., 2024) or density-marked clustering (Massara et al., 2024).

Combining 3D galaxy surveys and Sunyaev-Zeldovich data is very promising. As well as the number of spectra with DESI, the SZ measurements are improving rapidly with the current and next CMB experiments like Simons Observatory (Ade et al., 2019) and CMB-S4 (Abazajian et al., 2019; Carlstrom et al., 2019). Joint analysis of different measurements has high potential to improve our measurements and uncover new tensions.

Acknowledgments

MR and DJE are supported by U.S. Department of Energy grant DE-SC0013718 and as a Simons Foundation Investigator.

This material is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of High-Energy Physics, under Contract No. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore

Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: <https://www.desi.lbl.gov/collaborating-institutions>. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U. S. National Science Foundation, the U. S. Department of Energy, or any of the listed funding agencies.

The authors are honored to be permitted to conduct scientific research on Iolkam Du’ag (Kitt Peak), a mountain with particular significance to the Tohono O’odham Nation.

This work has used the following software packages: `ASTROPY` (Astropy Collaboration et al., 2013, 2018, 2022), `JUPYTER` (Perez & Granger, 2007; Kluyver et al., 2016), `MATPLOTLIB` (Hunter, 2007), `NUMPY` (Harris et al., 2020), `PYCORR` (De Mattia et al., 2024; Sinha & Garrison, 2020, 2019), `PYTHON` (Van Rossum & Drake, 2009), `SCIPY` (Virtanen et al., 2020; Gommers et al., 2023), and `SCIKIT-LEARN` (Pedregosa et al., 2011; Buitinck et al., 2013; Grisel et al., 2023).

This research has made use of NASA’s Astrophysics Data System. Software citation information has been aggregated using **The Software Citation Station** (Wagg & Broekgaarden, 2024; Wagg et al., 2024).

Chapter 6

Conclusions and outlook

Galaxy redshift surveys will remain a powerful cosmological probe in the near future. Among the key analyses of DESI DR2 (DESI Collaboration et al., 2026), only the BAO (DESI Collaboration et al., 2025b,a) have been finished, many others are ongoing. The products of the originally planned 5 years of observations could become available to the DESI collaboration next year. The plans beyond that include the upgraded DESI-2 and the next-generation Stage V instrument (Schlegel et al., 2022). Besides DESI, the Euclid space telescope (Euclid Collaboration et al., 2024) is already collecting data, and we hope to see quality spectroscopic products soon. The Roman Space Telescope (also known as the Wide-Field Infrared Survey Telescope and the Joint Dark Energy Mission Akeson et al., 2019) should be launched into space soon.

A major part of this thesis (Chapter 2) was dedicated to the development of fast and reliable semi-analytical covariance matrices for 2-point correlation functions of cosmological point tracers. This approach is much cheaper (both in computing and human time resources) and more flexible than relying on sample variations in large suites of approximate simulations. Accordingly, it enabled or enhanced many tests of the galaxy and quasar BAO methodology for DESI DR1 (DESI Collaboration et al., 2025f), and then considerably streamlined and sped up the DR2 BAO analysis (Andrade et al., 2025a; DESI Collaboration et al., 2025a). We are looking to further applications of the particular and more generic experience, besides the

reiteration for the upcoming DESI Y3 BAO.

Our first natural interest is the adjacent full-shape galaxy clustering analysis (e.g, DESI Collaboration et al., 2024b). It considers more aspects than only the scale of the BAO feature, including redshift-space distortions and signals from the matter-radiation equality scale. Accordingly, this analysis informs us much better about the growth of cosmic structure, regulated by the nature of dark matter and gravity.

The advantages come with challenges, such as higher-dimensional data vectors and accordingly covariance matrices (harder to estimate), and the need for more precise non-linear models with more parameters and more degeneracies between them. The future analysis of DR2 will likely include correlation functions or higher-point statistics to help break such degeneracies. It may be feasible to extend the covariance framework for these cases (similarly to Philcox & Eisenstein, 2019). We have already suggested a way to correct the simulation-based covariance for power spectra for the 2024 full-shape analysis (DESI Collaboration et al., 2024b,a) using the RASCALC semi-analytical covariance matrices.

The official DESI DR1 full-shape analysis (DESI Collaboration et al., 2024b) has been conducted in Fourier space (power spectra instead of correlation functions). We understand it was motivated by concerns that the precision of the modeling, naturally conducted in Fourier space (Maus et al., 2025a), may be partially lost in the Fourier transformation. However, Ramirez-Solano et al. (2025) reported a high quality of configuration-space (correlation function) modeling and analysis.

DESI Collaboration et al. (2024b) also combined the DR1 BAO post-reconstruction results (DESI Collaboration et al., 2025f) with full-shape power spectra, estimating their covariance via mocks. Analytical or semi-analytical covariance between the reconstructed correlation function and pre-reconstruction power spectrum would be beneficial, but it is very challenging. It might be feasible to compute the covariance between pre- and post-reconstruction correlation functions similarly to Section 2.2.3. Additionally, we may improve the early implementation of the RASCALC covariance between modified¹ power spectra and correlation functions (Philcox

¹The difference from the conventional power spectra should be small at small scales.

& Eisenstein, 2020b).

Our results also include a generic covariance matrix validation strategy, not specific to the numeric methods. In Sections 2.3.1 and 2.3.3 we have developed a brief, quantitative, and statistically interpretable framework for comparison with simulation-based covariances. These methods can be applied to a wide variety of observables besides correlation functions, from galaxy power spectra to CMB instrumental noise.

Higher-point statistics could also be used in future DESI measurements of primordial non-Gaussianity (building upon the completed analysis of DR1 by Chaussidon et al., 2024). The dimensionality of the data vector grows fast with the number of points, and simulation-based covariances then require proportionally more mock realizations. The semi-analytical covariances can help to maintain high accuracy at low computation cost.

Another promising direction is the joint analysis of multiple distinct galaxy types — different tracers of the large-scale structure. In DESI, several target types overlap in volume, most notably Luminous Red Galaxies (LRG, Zhou et al., 2023) with Emission Line Galaxies (ELG, Raichoor et al., 2023) in the redshift range $0.8 < z < 1.1$, and ELG with quasars (Chaussidon et al., 2023) in $1.1 < z < 1.6$. Additionally, the Bright Galaxy Survey (Hahn et al., 2023) includes a wide variety of objects and can be separated into multiple sub-types, e.g., red and blue. Cross-correlations of different galaxies can break parameter degeneracies and effectively increase the measurement precision. They can also serve as a consistency check to ensure the different samples reflect the same underlying structure (e.g., following Patej & Eisenstein, 2016), with the potential to uncover new cosmological tensions.

If the overlapping tracers are analyzed completely separately, valuable information from their cross-correlation remains unused. Moreover, the two samples are not independent but correlated to some degree, which needs to be quantified. The full analysis requires high-dimensional covariance matrices between all the auto-correlation and cross-correlation functions, and our semi-analytical framework provides them (in Section A.1). We have contributed a study of the optimal combination of DESI DR1 LRG and ELG into a single combined tracer (Valcin et al., 2025), which has been used for the 2024 BAO measurements

(DESI Collaboration et al., 2025f).

Our work on the DESI DR1 and DR2 BAO measurements empowered very intriguing cosmological findings (DESI Collaboration et al., 2025e,a). We summarized some of them in Chapter 3, focusing on dynamic dark energy and Hubble tension (which is not resolved by the dynamic dark energy preferred by DESI). This motivates us to become more involved in building and testing cosmological models capable of explaining either or both.

We summarized our earlier exploration of a promising Hubble tension relief in Chapter 4. It was a follow-up on the idea of small-scale baryon clumping at recombination (Jedamzik & Pogosian, 2020), possibly induced by primordial magnetic fields (Jedamzik & Abel, 2011, 2013; Jedamzik & Saveliev, 2019), which gives a higher value of the Hubble constant from CMB even without new physics, and reduces the S_8 tension. Whereas alternative recombination does not require fundamentally new physics, quantifying the uncertainty in known effects is also crucially important. We conducted a detailed follow-up study with a more flexible and general model for inhomogeneous recombination. We determined that with data available at the time, the preference for the effect depends significantly on the assumed a priori distributions for clumping parameters, the amount of external Hubble constant measurements, and the inclusion of additional probes like BAO (a conclusion similar to Jedamzik et al., 2021). We found that the strongest constraints on alternative recombination come from high multipoles of the CMB power spectrum, the “tail” affected by Silk damping. We showed that the future CMB experiments, Simons Observatory (Ade et al., 2019) and CMB-S4 (Abazajian et al., 2019; Carlstrom et al., 2019), will give decisive evidence for or against clumpy recombination. It may be interesting to revisit the idea with the newest measurements, although some other groups have been working on alternative recombination more actively (e.g., Lynch et al., 2024a).

Synergies between advanced CMB and galaxy redshift samples are very promising. For example, joint analysis of CMB lensing with DESI galaxies has given tighter constraints on the growth of cosmic structure (Kim et al., 2024; Sailer et al., 2024) and primordial non-Gaussianity (Krolewski et al., 2024; Bermejo-Climent et al., 2024). As well as the number of spectra with DESI, the measurements of secondary CMB anisotropies are improving rapidly with the

current and next CMB experiments (Qu et al., 2024; Coulton et al., 2024; Ade et al., 2019; Abazajian et al., 2019; Carlstrom et al., 2019). Joint analysis of different measurements has high potential to improve precision, provide valuable information on standing tensions, and discover new discrepancies.

In Chapter 5 we are investigating new techniques using thermal Sunyaev-Zeldovich (tSZ) data instead. This work is partially inspired by known ideas for leveraging information from different environments: density-split (e.g. Paillas et al., 2023) and density-marked (e.g. White, 2016) clustering, which deliver tighter cosmological constraints than the conventional 2-point correlation functions or power spectra. Strong tSZ signals are associated with distinct environments — large galaxy clusters. We are grouping DESI luminous red galaxies (Zhou et al., 2023) by the signal-to-noise ratio in the tSZ y parameter map. We find that with increasing tSZ signal-to-noise, there are significant increases in large-scale galaxy bias (allowing multi-tracer benefits discussed above), the velocity dispersion (allowing Finger-of-God mitigation for more precise and robust modeling, e.g. Baleato Lizancos et al., 2025) and the number of neighbors (which could help inform the galaxy multiplet studies, e.g. Lamman et al., 2024).

Thus, we demonstrate that there is valuable cosmological information in the tSZ map beyond the individual clusters. We hope our preliminary findings are sufficient to draw attention to the topic. We are going to continue developing the adjacent analysis techniques and encourage others to join us.

References

- Abazajian K., et al., 2019, arXiv e-prints, p. arXiv:1907.04473
- Abbott B. P., et al., 2017, *Nature*, 551, 85
- Abbott T. M. C., et al., 2018, *Phys. Rev. D*, 98, 043526
- Ade P., et al., 2019, *J. Cosmology Astropart. Phys.*, 2019, 056
- Agrawal P., Cyr-Racine F.-Y., Pinner D., Randall L., 2019, arXiv e-prints, p. arXiv:1904.01016
- Akeson R., et al., 2019, arXiv e-prints, p. arXiv:1902.05569
- Alam S., et al., 2017, *MNRAS*, 470, 2617
- Alam S., et al., 2021, *Phys. Rev. D*, 103, 083533
- Ali-Haïmoud Y., Hirata C. M., 2011, *Phys. Rev. D*, 83, 043513
- Alves O., et al., 2025, in preparation
- d’Amico G., Gleyzes J., Kokron N., Markovic K., Senatore L., Zhang P., Beutler F., Gil-Marín H., 2020, *J. Cosmology Astropart. Phys.*, 2020, 005
- Anderson L., et al., 2012, *MNRAS*, 427, 3435
- Andrade U., et al., 2025a, arXiv e-prints, p. arXiv:2503.14742
- Andrade U., et al., 2025b, *J. Cosmology Astropart. Phys.*, 2025, 128
- Astropy Collaboration et al., 2013, *A&A*, 558, A33
- Astropy Collaboration et al., 2018, *AJ*, 156, 123
- Astropy Collaboration et al., 2022, *ApJ*, 935, 167
- Ata M., et al., 2018, *MNRAS*, 473, 4773
- Atkins Z., et al., 2023, *J. Cosmology Astropart. Phys.*, 2023, 073
- Audren B., Lesgourgues J., Benabed K., Prunet S., 2013, *J. Cosmology Astropart. Phys.*, 2013, 001

Baleato Lizancos A., Seljak U., Karamanis M., Bonici M., Ferraro S., 2025, arXiv e-prints, p. arXiv:2501.10587

Banerjee R., Jedamzik K., 2004, Phys. Rev. D, 70, 123003

Bermejo-Climent J. R., et al., 2024, arXiv e-prints, p. arXiv:2412.10279

Bianchi D., et al., 2024, arXiv e-prints, p. arXiv:2411.12025

Birrer S., et al., 2020, A&A, 643, A165

Blas D., Lesgourgues J., Tram T., 2011, J. Cosmology Astropart. Phys., 7, 034

Bocquet S., et al., 2023, arXiv e-prints, p. arXiv:2310.12213

Brieden S., Gil-Marín H., Verde L., 2021, J. Cosmology Astropart. Phys., 2021, 054

Brinckmann T., Lesgourgues J., 2019, Physics of the Dark Universe, 24, 100260

Brout D., et al., 2022, ApJ, 938, 110

Buitinck L., et al., 2013, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp 108–122

Burden A., Percival W. J., Howlett C., 2015, MNRAS, 453, 456

Burns C. R., et al., 2018, ApJ, 869, 56

Carlstrom J., et al., 2019, in Bulletin of the American Astronomical Society. p. 209 (arXiv:1908.01062)

Carron J., Mirmelstein M., Lewis A., 2022, J. Cosmology Astropart. Phys., 2022, 039

Cartis C., Fiala J., Marteau B., Roberts L., 2018a, arXiv e-prints, p. arXiv:1804.00154

Cartis C., Roberts L., Sheridan-Methven O., 2018b, arXiv e-prints, p. arXiv:1812.11343

Chaussidon E., et al., 2023, ApJ, 944, 107

Chaussidon E., et al., 2024, arXiv e-prints, p. arXiv:2411.17623

Chen S.-F., Vlah Z., White M., 2020, J. Cosmology Astropart. Phys., 2020, 062

Chen S.-F., Vlah Z., Castorina E., White M., 2021, J. Cosmology Astropart. Phys., 2021, 100

Chen X., et al., 2024a, arXiv e-prints, p. arXiv:2411.19738

Chen S. F., et al., 2024b, MNRAS, 534, 544

Chevallier M., Polarski D., 2001, International Journal of Modern Physics D, 10, 213

Choi S. K., et al., 2020, J. Cosmology Astropart. Phys., 2020, 045

Chown R., et al., 2018, ApJS, 239, 10

Chuang C.-H., Wang Y., 2013, MNRAS, 431, 2634

Chuang C.-H., Kitaura F.-S., Prada F., Zhao C., Yepes G., 2015, MNRAS, 446, 2621

Colas T., d’Amico G., Senatore L., Zhang P., Beutler F., 2020, J. Cosmology Astropart. Phys., 2020, 001

Cole S., et al., 2005, MNRAS, 362, 505

Cort  s M., Liddle A. R., 2024, J. Cosmology Astropart. Phys., 2024, 007

Coulton W., et al., 2024, Phys. Rev. D, 109, 063530

Cuesta-Lazaro C., et al., 2024, MNRAS, 531, 3336

DES Collaboration et al., 2021, arXiv e-prints, p. arXiv:2105.13549

DES Collaboration et al., 2024, arXiv e-prints, p. arXiv:2401.02929

DESI Collaboration et al., 2016a, arXiv e-prints, p. arXiv:1611.00036

DESI Collaboration et al., 2016b, arXiv e-prints, p. arXiv:1611.00037

DESI Collaboration et al., 2022, AJ, 164, 207

DESI Collaboration et al., 2024a, arXiv e-prints, p. arXiv:2411.12020

DESI Collaboration et al., 2024b, arXiv e-prints, p. arXiv:2411.12021

DESI Collaboration et al., 2024c, arXiv e-prints, p. arXiv:2411.12022

DESI Collaboration et al., 2024d, AJ, 167, 62

DESI Collaboration et al., 2024e, AJ, 168, 58

DESI Collaboration et al., 2025a, arXiv e-prints, p. arXiv:2503.14738

DESI Collaboration et al., 2025b, arXiv e-prints, p. arXiv:2503.14739

DESI Collaboration et al., 2025c, arXiv e-prints, p. arXiv:2503.14745

DESI Collaboration et al., 2025d, J. Cosmology Astropart. Phys., 2025, 124

DESI Collaboration et al., 2025e, J. Cosmology Astropart. Phys., 2025, 021

DESI Collaboration et al., 2025f, J. Cosmology Astropart. Phys., 2025, 012

DESI Collaboration et al., 2026, in preparation

Davis M., Peebles P. J. E., 1983, ApJ, 267, 465

Dey A., et al., 2019, AJ, 157, 168

Dey B., Newman J. A., Andrews B. H., Izbicki R., Lee A. B., Zhao D., Rau M. M., Malz A. I., 2021, arXiv e-prints, p. arXiv:2110.15209

- Dey B., Andrews B. H., Newman J. A., Mao Y.-Y., Rau M. M., Zhou R., 2022, MNRAS, 515, 5285
- Dhawan S., Jha S. W., Leibundgut B., 2018, A&A, 609, A72
- Di Valentino E., Melchiorri A., Mena O., Vagnozzi S., 2020, Phys. Rev. D, 101, 063502
- Dinda B. R., Maartens R., 2025, arXiv e-prints, p. arXiv:2504.15190
- Dolgov A., Silk J., 1993, Phys. Rev. D, 47, 4244
- Eisenstein D. J., et al., 2005, ApJ, 633, 560
- Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007, ApJ, 664, 675
- Elbers W., et al., 2025, arXiv e-prints, p. arXiv:2503.14744
- Ereza J., et al., 2023, arXiv e-prints, p. arXiv:2311.14456
- Euclid Collaboration et al., 2024, arXiv e-prints, p. arXiv:2405.13491
- Fang X., Eifler T., Krause E., 2020, MNRAS, 497, 2699
- Feldman H. A., Kaiser N., Peacock J. A., 1994, Astrophys. J., 426, 23
- Ferraro S., Sailer N., Slosar A., White M., 2022, arXiv e-prints, p. arXiv:2203.07506
- Ferreira T., Marra V., 2022, MNRAS, 513, 5438
- Forero-Sánchez D., et al., 2024, arXiv e-prints, p. arXiv:2411.12027
- Fowler J. W., et al., 2007, Appl. Opt., 46, 3444
- Freedman W. L., Madore B. F., Scowcroft V., Burns C., Monson A., Persson S. E., Seibert M., Rigby J., 2012, ApJ, 758, 24
- Freedman W. L., Madore B. F., Jang I. S., Hoyt T. J., Lee A. J., Owens K. A., 2024, arXiv e-prints, p. arXiv:2408.06153
- Galli S., Pogosian L., Jedamzik K., Balkenhol L., 2022, Phys. Rev. D, 105, 023513
- García-García C., Ruiz-Zapatero J., Alonso D., Bellini E., Ferreira P. G., Mueller E.-M., Nicola A., Ruiz-Lapuente P., 2021, J. Cosmology Astropart. Phys., 2021, 030
- Garcia-Quintero C., et al., 2025, J. Cosmology Astropart. Phys., 2025, 132
- Garrison L. H., Eisenstein D. J., Ferrer D., Maksimova N. A., Pinto P. A., 2021, MNRAS, 508, 575
- Giarè W., Sabogal M. A., Nunes R. C., Di Valentino E., 2024, Phys. Rev. Lett., 133, 251003
- Gommers R., et al., 2023, scipy/scipy: SciPy 1.11.1, doi:10.5281/zenodo.8092679, <https://doi.org/10.5281/zenodo.8092679>

Grieb J. N., Sánchez A. G., Salazar-Albornoz S., Dalla Vecchia C., 2016, MNRAS, 457, 1577

Grisel O., et al., 2023, scikit-learn/scikit-learn: Scikit-learn 1.3.2, doi:10.5281/zenodo.10034229, <https://doi.org/10.5281/zenodo.10034229>

Hadzhiyska B., Spergel D., 2019, Phys. Rev. D, 99, 043537

Hadzhiyska B., Eisenstein D., Bose S., Garrison L. H., Maksimova N., 2022, MNRAS, 509, 501

Hahn C., et al., 2023, AJ, 165, 253

Hand N., Li Y., Slepian Z., Seljak U., 2017, J. Cosmology Astropart. Phys., 2017, 002

Handley W., 2019, The Journal of Open Source Software, 4, 1414

Handley W. J., Hobson M. P., Lasenby A. N., 2015a, MNRAS, 450, L61

Handley W. J., Hobson M. P., Lasenby A. N., 2015b, MNRAS, 453, 4384

Hanif M. M. S., et al., 2025, in preparation

Harris C. R., et al., 2020, Nature, 585, 357

Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399

Heymans C., et al., 2021, A&A, 646, A140

Hildebrandt H., et al., 2020, A&A, 633, A69

Hill J. C., McDonough E., Toomey M. W., Alexander S., 2020, Phys. Rev. D, 102, 043507

Hilton M., et al., 2021, ApJS, 253, 3

Hinshaw G., et al., 2013, ApJS, 208, 19

Hu W., White M., 1997, ApJ, 479, 568

Huang C. D., et al., 2020, ApJ, 889, 5

Hubble E. P., 1926, ApJ, 64, 321

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168

Hunter J. D., 2007, Computing in Science & Engineering, 9, 90

Isserlis L., 1918, Biometrika, 12, 134

Ivanov M. M., Simonović M., Zaldarriaga M., 2020, J. Cosmology Astropart. Phys., 2020, 042

de Jaeger T., Stahl B. E., Zheng W., Filippenko A. V., Riess A. G., Galbany L., 2020, MNRAS, 496, 3402

Jang I. S., Lee M. G., 2017, ApJ, 836, 74

Jedamzik K., Abel T., 2011, arXiv e-prints, p. arXiv:1108.2517

Jedamzik K., Abel T., 2013, *J. Cosmology Astropart. Phys.*, 2013, 050

Jedamzik K., Pogosian L., 2020, *Phys. Rev. Lett.*, 125, 181302

Jedamzik K., Saveliev A., 2019, *Phys. Rev. Lett.*, 123, 021301

Jedamzik K., Pogosian L., Zhao G.-B., 2021, *Communications Physics*, 4, 123

Jedamzik K., Pogosian L., Abel T., 2025, arXiv e-prints, p. arXiv:2503.09599

Joudaki S., et al., 2017, *MNRAS*, 465, 2033

Kass R. E., Raftery A. E., 1995, *Journal of the American Statistical Association*, 90, 773

Keihänen E., et al., 2019, *A&A*, 631, A73

Kerscher M., Szapudi I., Szalay A. S., 2000, *ApJ*, 535, L13

Khetan N., et al., 2021, *A&A*, 647, A72

Kim J., et al., 2024, *J. Cosmology Astropart. Phys.*, 2024, 022

Kluyver T., et al., 2016, in *ELPUB*. pp 87–90

Knox L., Millea M., 2020, *Phys. Rev. D*, 101, 043533

Kobayashi Y., 2023, *Phys. Rev. D*, 108, 103512

Kornoelje K., et al., 2025, arXiv e-prints, p. arXiv:2503.17271

Kourkchi E., Tully R. B., Anand G. S., Courtois H. M., Dupuy A., Neill J. D., Rizzi L., Seibert M., 2020, *ApJ*, 896, 3

Kreisch C. D., Cyr-Racine F.-Y., Doré O., 2020, *Phys. Rev. D*, 101, 123505

Krolewski A., Ferraro S., White M., 2021, arXiv e-prints, p. arXiv:2105.03421

Krolewski A., et al., 2024, *J. Cosmology Astropart. Phys.*, 2024, 021

Lai Y., et al., 2025, *J. Cosmology Astropart. Phys.*, 2025, 139

Lamman C., et al., 2024, *MNRAS*, 534, 3540

Landy S. D., Szalay A. S., 1993, *ApJ*, 412, 64

Lasker J., et al., 2025, *J. Cosmology Astropart. Phys.*, 2025, 127

Leavitt H. S., Pickering E. C., 1912, *Harvard College Observatory Circular*, 173, 1

Lee N., Ali-Haïmoud Y., 2020, *Phys. Rev. D*, 102, 083517

Lewis A., 2019, arXiv e-prints, p. arXiv:1910.13970

Lin M.-X., Benevento G., Hu W., Raveri M., 2019, *Phys. Rev. D*, 100, 063542

Linder E. V., 2003, *Phys. Rev. Lett.*, 90, 091301

Linder E. V., 2007, arXiv e-prints, p. arXiv:0708.0024

Liu R. H., Hadzhiyska B., Ferraro S., Bose S., Hernández-Aguayo C., 2025, arXiv e-prints, p. arXiv:2504.11794

Lodha K., et al., 2025a, arXiv e-prints, p. arXiv:2503.14743

Lodha K., et al., 2025b, *Phys. Rev. D*, 111, 023532

Lynch G. P., Knox L., Chluba J., 2024a, *Phys. Rev. D*, 110, 063518

Lynch G. P., Knox L., Chluba J., 2024b, *Phys. Rev. D*, 110, 083538

Maksimova N. A., Garrison L. H., Eisenstein D. J., Hadzhiyska B., Bose S., Satterthwaite T. P., 2021, *MNRAS*, 508, 4017

Massara E., Villaescusa-Navarro F., Ho S., Dalal N., Spergel D. N., 2021, *Phys. Rev. Lett.*, 126, 011301

Massara E., et al., 2023, *ApJ*, 951, 70

Massara E., et al., 2024, arXiv e-prints, p. arXiv:2404.04228

de Mattia A., Rashkovetskyi M., Sinha M., Garrison L. H., 2024, pycorr: Two-point correlation function estimation, *Astrophysics Source Code Library*, record ascl:2403.009

Maus M., et al., 2025a, *J. Cosmology Astropart. Phys.*, 2025, 134

Maus M., et al., 2025b, *J. Cosmology Astropart. Phys.*, 2025, 138

Mena-Fernández J., et al., 2025, *J. Cosmology Astropart. Phys.*, 2025, 133

Mirpoorian S. H., Jedamzik K., Pogosian L., 2025a, arXiv e-prints, p. arXiv:2504.15274

Mirpoorian S. H., Jedamzik K., Pogosian L., 2025b, *Phys. Rev. D*, 111, 083519

Mohammad F. G., Percival W. J., 2022, *MNRAS*, 514, 1289

Moon J., et al., 2023, *MNRAS*, 525, 5406

Morawetz J., Paillas E., Percival W. J., 2024, arXiv e-prints, p. arXiv:2409.13583

Muñoz J. B., 2019, *Phys. Rev. Lett.*, 123, 131301

Myles J., et al., 2021, *MNRAS*, 505, 4249

Noriega H. E., et al., 2025, *J. Cosmology Astropart. Phys.*, 2025, 136

O’Connell R., Eisenstein D. J., 2019, *MNRAS*, 487, 2701

O’Connell R., Eisenstein D., Vargas M., Ho S., Padmanabhan N., 2016, *MNRAS*, 462, 2681

Osato K., Nagai D., 2023, MNRAS, 519, 2069

Paillas E., Cai Y.-C., Padilla N., Sánchez A. G., 2021, MNRAS, 505, 5731

Paillas E., et al., 2023, MNRAS, 522, 606

Paillas E., et al., 2024, MNRAS, 531, 898

Paillas E., et al., 2025, J. Cosmology Astropart. Phys., 2025, 142

Patej A., Eisenstein D., 2016, MNRAS, 460, 1310

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peebles P. J. E., 1968, ApJ, 153, 1

Peebles P. J. E., 1976, Ap&SS, 45, 3

Peebles P. J. E., Groth E. J., 1975, ApJ, 196, 1

Peebles P. J. E., Yu J. T., 1970, ApJ, 162, 815

Percival W. J., et al., 2014, MNRAS, 439, 2531

Percival W. J., Friedrich O., Sellentin E., Heavens A., 2022, MNRAS, 510, 3207

Perez F., Granger B. E., 2007, Computing in Science and Engineering, 9, 21

Pérez-Fernández A., et al., 2025, J. Cosmology Astropart. Phys., 2025, 144

Perlmutter S., et al., 1999, ApJ, 517, 565

Pesce D. W., et al., 2020, ApJ, 891, L1

Philcox O. H. E., Eisenstein D. J., 2019, MNRAS, 490, 5931

Philcox O. H. E., Eisenstein D. J., 2020a, MNRAS, 492, 1214

Philcox O. H. E., Eisenstein D. J., 2020b, MNRAS, 492, 1214

Philcox O. H. E., Eisenstein D. J., O’Connell R., Wiegand A., 2020, MNRAS, 491, 3290

Philcox O. H. E., Ivanov M. M., Zaldarriaga M., Simonović M., Schmittfull M., 2021, Phys. Rev. D, 103, 043508

Pinon M., et al., 2025, J. Cosmology Astropart. Phys., 2025, 131

Planck Collaboration et al., 2016a, A&A, 594, A22

Planck Collaboration et al., 2016b, A&A, 594, A27

Planck Collaboration et al., 2020a, A&A, 641, A1

Planck Collaboration et al., 2020b, A&A, 641, A5

Planck Collaboration et al., 2020c, *A&A*, 641, A6

Planck Collaboration et al., 2020d, *A&A*, 641, A8

Planck Collaboration et al., 2020e, *A&A*, 641, A9

Pons-Bordería M.-J., Martínez V. J., Stoyan D., Stoyan H., Saar E., 1999, *ApJ*, 523, 480

Poulin V., Smith T. L., Karwal T., Kamionkowski M., 2019, *Phys. Rev. Lett.*, 122, 221301

Powell M. J. D., 2009, Technical Report, Department of Applied Mathematics and Theoretical Physics

de Putter R., Linder E. V., 2008, *J. Cosmology Astropart. Phys.*, 2008, 042

Qu F. J., et al., 2024, *ApJ*, 962, 112

Raichoor A., et al., 2023, *AJ*, 165, 126

Ramirez-Solano S., et al., 2025, *J. Cosmology Astropart. Phys.*, 2025, 129

Rashkovetskyi M., Muñoz J. B., Eisenstein D. J., Dvorkin C., 2021, *Phys. Rev. D*, 104, 103517

Rashkovetskyi M., et al., 2023, *MNRAS*, 524, 3894

Rashkovetskyi M., Eisenstein D. J., et al., 2025a, in preparation

Rashkovetskyi M., et al., 2025b, *J. Cosmology Astropart. Phys.*, 2025, 145

Raveri M., 2020, *Phys. Rev. D*, 101, 083524

Rezaie M., et al., 2024, *MNRAS*, 532, 1902

Riess A. G., et al., 1998, *AJ*, 116, 1009

Riess A. G., Casertano S., Yuan W., Bowers J. B., Macri L., Zinn J. C., Scolnic D., 2021, *ApJ*, 908, L6

Riess A. G., et al., 2022, *ApJ*, 934, L7

Robertson N. C., et al., 2024, *A&A*, 681, A87

Ross A. J., et al., 2017, *MNRAS*, 464, 1168

Ross A. J., et al., 2025, *J. Cosmology Astropart. Phys.*, 2025, 125

van Rossum G., Drake F. L., 2009, *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA

Rubin D., et al., 2023, arXiv e-prints, p. arXiv:2311.12098

Sailer N., et al., 2024, arXiv e-prints, p. arXiv:2407.04607

Sakstein J., Trodden M., 2020, *Phys. Rev. Lett.*, 124, 161301

Schlaflly E. F., et al., 2023, *AJ*, 166, 259

Schlegel D. J., et al., 2022, arXiv e-prints, p. arXiv:2209.03585

Schöneberg N., 2024, *J. Cosmology Astropart. Phys.*, 2024, 006

Seager S., Sasselov D. D., Scott D., 1999, *ApJ*, 523, L1

Seager S., Sasselov D. D., Scott D., 2000, *ApJS*, 128, 407

Shajib A. J., et al., 2020, *MNRAS*, 494, 6072

Shapley H., 1933, *Proceedings of the National Academy of Science*, 19, 389

Sheth R. K., 1996, *MNRAS*, 279, 1310

Silber J. H., et al., 2023, *AJ*, 165, 9

Silk J., 1968, *ApJ*, 151, 459

Sinha M., Garrison L. H., 2019, in *Software Challenges to Exascale Computing. Second Workshop*. pp 3–20 (arXiv:1911.08275), doi:10.1007/978-981-13-7729-7_1

Sinha M., Garrison L. H., 2020, *MNRAS*, 491, 3022

Slipher V. M., 1917, *Proceedings of the American Philosophical Society*, 56, 403

Stein G., Alvarez M. A., Bond J. R., van Engelen A., Battaglia N., 2020, *J. Cosmology Astropart. Phys.*, 2020, 012

Subramanian K., 2016, *Reports on Progress in Physics*, 79, 076901

Sullivan J. M., Prijon T., Seljak U., 2023, *J. Cosmology Astropart. Phys.*, 2023, 004

Sunyaev R. A., Zeldovich Y. B., 1970, *Ap&SS*, 7, 3

Sunyaev R. A., Zeldovich Y. B., 1972, *Comments on Astrophysics and Space Physics*, 4, 173

Sunyaev R. A., Zeldovich I. B., 1980, *ARA&A*, 18, 537

Takada M., Jain B., 2003, *MNRAS*, 340, 580

Taylor A., Joachimi B., Kitching T., 2013, *MNRAS*, 432, 1928

Thiele L., Guan Y., Hill J. C., Kosowsky A., Spergel D. N., 2021, *Phys. Rev. D*, 104, 063535

Torrado J., Lewis A., 2021, *J. Cosmology Astropart. Phys.*, 2021, 057

Trusov S., et al., 2024, *MNRAS*, 527, 9048

Turin G., 1960, *IRE Transactions on Information Theory*, 6, 311

Valcin D., Rashkovetskyi M., Seo H., Beutler F., McDonald P., de Mattia A., et al., 2025, in preparation

Vargas-Magaña M., et al., 2018, MNRAS, 477, 1153

Variu A., et al., 2024, MNRAS, 527, 11539

Virtanen P., et al., 2020, Nature Methods, 17, 261

Wadekar D., Scoccimarro R., 2020, Phys. Rev. D, 102, 123517

Wagg T., Broekgaarden F. S., 2024, arXiv e-prints, p. arXiv:2406.04405

Wagg T., Broekgaarden F., Gültekin K., 2024, TomWagg/software-citation-station: v1.2, doi:10.5281/zenodo.13225824, <https://doi.org/10.5281/zenodo.13225824>

Wang Y., Pogosian L., Zhao G.-B., Zucca A., 2018, ApJ, 869, L8

White M., 2016, J. Cosmology Astropart. Phys., 2016, 057

Wick G. C., 1950, Physical Review, 80, 268

Wong W. Y., Moss A., Scott D., 2008, MNRAS, 386, 1023

Wong K. C., et al., 2020, MNRAS, 498, 1420

Xu X., et al., 2010, ApJ, 718, 1224

Yuan S., Garrison L. H., Hadzhiyska B., Bose S., Eisenstein D. J., 2022, MNRAS, 510, 3301

Yuan S., et al., 2023, arXiv e-prints, p. arXiv:2306.06314

Zel'dovich Y. B., Kurt V. G., Syunyaev R. A., 1969, Soviet Journal of Experimental and Theoretical Physics, 28, 146

Zhao G.-B., et al., 2017, Nature Astronomy, 1, 627

Zhao C., et al., 2021, MNRAS, 503, 1149

Zhao C., et al., 2025, in preparation

Zhou R., et al., 2023, AJ, 165, 58

Appendix A

Appendix to the covariance matrices (Chapter 2)

A.1 Estimators for multi-tracer covariance matrices

In this section, we provide the generalized multi-tracer expressions for reference. We have not endeavored to validate them in this work because in the **EZmocks** suite, different tracers have mostly been made from boxes at different redshifts. We expect that the shift of the galaxies' positions with time will not allow self-consistent cross-correlations in such circumstances. DESI has primarily focused on the LRG and ELG overlap in $z = 0.8 - 1.1$ bin (Valcin et al., 2025), but **EZmocks** for these tracers have been made from box snapshots at redshifts 1.1 and 0.95, respectively.

We add one capital Latin superscript to all the overdensities to signify the tracer involved (e.g. δ_i^X). For counts, it is two letters accordingly (e.g. $N^X N^Y$).

A.1.1 Angular bins

The Landy-Szalay estimator for the multi-tracer correlation function is

$$\left(\hat{\xi}^{XY}\right)_a^c = \frac{(N^X N^Y)_a^c}{(R^X R^Y)_a^c}. \quad (\text{A.1})$$

We make the following shorthand for the multi-tracer covariance:

$$(C^{XY,ZW})_{ab}^{cd} \equiv \text{cov} \left[\left(\hat{\xi}^{XY} \right)_a^c, \left(\hat{\xi}^{ZW} \right)_b^d \right]. \quad (\text{A.2})$$

It is important to note that the shot-noise approximation works only for the same-tracer overdensities:

$$(\delta_i^X)^2 \approx \frac{\alpha_{\text{SN}}^X}{n_i^X} (1 + \delta_i^X). \quad (\text{A.3})$$

The full multi-tracer expression for the model covariance is from Philcox et al. (2020) but with some changes in notation:

$$\begin{aligned} \left(\tilde{C}^{XY,ZW} \right)_{ab}^{cd}(\alpha_{\text{SN}}) &= ({}^4C^{XY,ZW})_{ab}^{cd} + \frac{\alpha_{\text{SN}}^X}{4} \left[\delta^{XW} ({}^3C^{X,YZ})_{ab}^{cd} + \delta^{XZ} ({}^3C^{X,YW})_{ab}^{cd} \right] \\ &\quad + \frac{\alpha_{\text{SN}}^Y}{4} \left[\delta^{YW} ({}^3C^{Y,XZ})_{ab}^{cd} + \delta^{YZ} ({}^3C^{Y,XW})_{ab}^{cd} \right] \\ &\quad + \frac{\alpha_{\text{SN}}^X \alpha_{\text{SN}}^Y}{2} \left[\delta^{XW} \delta^{YZ} + \delta^{XZ} \delta^{YW} \right] ({}^2C^{XY})_{ab}^{cd} \end{aligned} \quad (\text{A.4})$$

with

$$\begin{aligned} ({}^4C^{XY,ZW})_{ab}^{cd} &= \frac{1}{(R^X R^Y)_a^c (R^Z R^W)_b^d} \sum_{i \neq j \neq k \neq l} n_i^X n_j^Y n_k^Z n_l^W w_i^X w_j^Y w_k^Z w_l^W \Theta^a(r_{ij}) \quad (\text{A.5}) \\ &\quad \times \Theta^c(\mu_{ij}) \Theta^b(r_{kl}) \Theta^d(\mu_{kl}) \left[\cancel{\eta_{ijkl}^{(c),XYWZ}} + \xi_{ik}^{XZ} \xi_{jl}^{YW} + \xi_{il}^{XW} \xi_{jk}^{YZ} \right] \\ ({}^3C^{Y,XZ})_{ab}^{cd} &= \frac{4}{(R^X R^Y)_a^c (R^Z R^W)_b^d} \sum_{i \neq j \neq k} n_i^X n_j^Y n_k^Z w_i^X (w_j^Y)^2 w_k^Z \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) \\ &\quad \times \Theta^b(r_{jk}) \Theta^d(\mu_{jk}) \left[\zeta_{ijk}^{XYZ} + \xi_{ik}^{XZ} \right] \\ ({}^2C^{XY})_{ab}^{cd} &= \frac{2\delta^{ab}\delta^{cd}}{(R^X R^Y)_a^c (R^X R^Y)_b^d} \sum_{i \neq j} n_i^X n_j^Y (w_i^X w_j^Y)^2 \Theta^a(r_{ij}) \Theta^c(\mu_{ij}) [1 + \xi_{ij}^{XY}], \end{aligned}$$

where δ^{XY} , δ^{ab} and δ^{cd} are Kronecker deltas; $\xi_{ij}^{XY} = \xi^{XY}(r_{ij}, \mu_{ij})$ is the 2PCF of tracers X and Y evaluated at the separation between points number i and j (in practice the value is obtained by bicubic interpolation from the input grid of correlation function values).

Analogously, ζ_{ijk}^{XYZ} and $\eta_{ijkl}^{(c),XYZW}$ are the 3-point and connected 4-point correlation functions of the tracers listed in the superscript evaluated at the separations between i, j, k and i, j, k, l points, respectively. These non-Gaussian higher-point functions are included for

completeness but dropped in practice, we reflected this by crossing them out in the expressions.

The expressions for multi-tracer jackknife covariance can be constructed similarly. However, to the best of our knowledge, only the single-tracer jackknife auto-covariances have been used in practice. The shot-noise rescaling values have been tuned on each tracer separately. Therefore, the computation of the jackknife model for the other blocks of the full matrix has been unnecessary; the single-tracer expressions are sufficient.

A.1.2 Projected Legendre

Note that the Legendre projection factors (Equation (2.18)) are the same for all tracers. Accordingly, the multi-tracer covariance is still projected linearly from the angularly binned one:

$$\left(\tilde{C}^{XY,ZW}\right)_{ab}^{\ell\ell'} \equiv \text{cov} \left[\left(\hat{\xi}^{XY}\right)_a^\ell, \left(\hat{\xi}^{ZW}\right)_b^{\ell'} \right] = \sum_{c,d} \left(\tilde{C}^{XY,ZW}\right)_{ab}^{cd} F_c^\ell F_d^{\ell'}. \quad (\text{A.6})$$

Therefore the full covariance model is constructed analogously to Equation (A.4):

$$\begin{aligned} (C^{XY,ZW})_{ab}^{\ell\ell'} = & ({}^4C^{XY,ZW})_{ab}^{\ell\ell'} + \frac{\alpha_{\text{SN}}^X}{4} \left[\delta^{XW} ({}^3C^{X,YZ})_{ab}^{\ell\ell'} + \delta^{XZ} ({}^3C^{X,YW})_{ab}^{\ell\ell'} \right] \\ & + \frac{\alpha_{\text{SN}}^Y}{4} \left[\delta^{YW} ({}^3C^{Y,XZ})_{ab}^{\ell\ell'} + \delta^{YZ} ({}^3C^{Y,XW})_{ab}^{\ell\ell'} \right] \\ & + \frac{\alpha_{\text{SN}}^X \alpha_{\text{SN}}^Y}{2} (\delta^{XW} \delta^{YZ} + \delta^{XZ} \delta^{YW}) ({}^2C^{XY})_{ab}^{\ell\ell'}. \end{aligned} \quad (\text{A.7})$$

with the following terms (similar to Equation (2.20)):

$$\begin{aligned} ({}^4C^{XY,ZW})_{ab}^{\ell\ell'} = & \sum_{i \neq j \neq k \neq l} n_i^X n_j^Y n_k^Z n_l^W w_i^X w_j^Y w_k^Z w_l^W \Theta^a(r_{ij}) \Theta^b(r_{kl}) \\ & \times \left[\cancel{\eta_{ijkl}^{(c),XYWZ}} + 2\xi_{ik}^{XZ} \xi_{jl}^{YW} \right] \sum_c \frac{\Theta^c(\mu_{ij}) F_c^\ell}{(R^X R^Y)_a^c} \sum_d \frac{\Theta^d(\mu_{kl}) F_d^{\ell'}}{(R^Z R^W)_b^d}, \\ ({}^3C^{Y,XZ})_{ab}^{\ell\ell'} = & 4 \sum_{i \neq j \neq k} n_i^X n_j^Y n_k^Z w_i^X (w_j^Y)^2 w_k^Z \Theta^a(r_{ij}) \Theta^b(r_{jk}) \left[\zeta_{ijk}^{XYZ} + \xi_{ik}^{XZ} \right] \\ & \times \sum_c \frac{\Theta^c(\mu_{ij}) F_c^\ell}{(R^X R^Y)_a^c} \sum_d \frac{\Theta^d(\mu_{jk}) F_d^{\ell'}}{(R^Y R^Z)_b^d}, \\ ({}^2C^{XY})_{ab}^{\ell\ell'} = & 2\delta^{ab} \sum_{i \neq j} n_i^X n_j^Y (w_i^X w_j^Y)^2 \Theta^a(r_{ij}) [1 + \xi_{ij}^{XY}] \sum_c \frac{\Theta^c(\mu_{ij}) F_c^\ell F_c^{\ell'}}{[(R^X R^Y)_a^c]^2}. \end{aligned} \quad (\text{A.8})$$

The multi-tracer covariance model for Legendre multipoles of the 2PCF is implemented in the code and has been employed for overlapping tracers’ cross-correlation in Valcin et al. (2025). We omit the corresponding jackknife expressions as they have not been used.

A.2 Statistics of comparison metrics for noisy sample covariance matrix

Here we provide derivations of the expectation values for comparison metrics between a noisy sample covariance and the true covariance/precision matrix. This is useful for testing how close RASCALC results are to the latter.

A.2.1 KL divergence mean and variance

A more generic setup – two sample covariance matrices based on draws from a multivariate normal distribution – has been considered in Appendix D of Philcox et al. (2020). However, the derivation was limited to the expectation value of the KL divergence between them, and we have not been able to find a reference about the metric’s scatter around the mean (variance or standard deviation). In addition, we believe the final result there is slightly incorrect, namely 1 should be subtracted from the number of samples. This is because for the estimate of sample covariance commonly used with mocks

$$X_{ab} = \frac{1}{n_S - 1} \sum_{i=1}^{n_S} (x_{a,i} - \bar{x}_a)(x_{b,i} - \bar{x}_b) \quad (\text{A.9})$$

the mean is not known beforehand but estimated from the sample as well: $\bar{x}_a \equiv \frac{1}{n_S} \sum_{i=1}^{n_S} x_{a,i}$. This reduces the number of degrees of freedom by one. Then the covariance of the sample covariance matrix elements is

$$\text{cov}(X_{ab}, X_{cd}) = \frac{C_{0,ac}C_{0,bd} + C_{0,ad}C_{0,bc}}{n_S - 1} \quad (\text{A.10})$$

instead of

$$\text{cov}(X_{ab}, X_{cd}) = \frac{C_{0,ac}C_{0,bd} + C_{0,ad}C_{0,bc}}{n_S}$$

as in Philcox et al. (2020). $\mathbf{C}_0$ is the true underlying covariance matrix of the Gaussian distribution the samples are drawn from. The sample covariance estimate is unbiased, meaning that the expectation value is the true covariance: $\langle \mathbf{X} \rangle = \mathbf{C}_0$. Then, considering two sample covariance matrices $\mathbf{X}_i$ obtained from $n_S^{(i)}$ samples each, decomposing them as $\mathbf{X}_i = \mathbf{C}_0 + \delta \mathbf{X}_i$, Taylor expanding and only leaving the leading nontrivial (quadratic) order in $\delta \mathbf{X}_i$, we obtain

$$\langle D_{\text{KL}}(\mathbf{X}_1^{-1}, \mathbf{X}_2) \rangle \approx \frac{N_{\text{bins}}(N_{\text{bins}} + 1)}{4} \left(\frac{1}{n_S^{(1)} - 1} + \frac{1}{n_S^{(2)} - 1} \right) \quad (\text{A.11})$$

instead of

$$\langle D_{\text{KL}}(\mathbf{X}_1^{-1}, \mathbf{X}_2) \rangle \approx \frac{N_{\text{bins}}(N_{\text{bins}} + 1)}{4} \left(\frac{1}{n_S^{(1)}} + \frac{1}{n_S^{(2)}} \right)$$

as in Philcox et al. (2020).

Since in this work we only consider one sample covariance matrix, while the RASCALC results are not expected to follow the Wishart distribution, the more relevant result is for the true precision matrix Ψ_0 :

$$\langle D_{\text{KL}}(\Psi_0, \mathbf{X}) \rangle \approx \frac{N_{\text{bins}}(N_{\text{bins}} + 1)}{4(n_S - 1)}, \quad (\text{A.12})$$

which can be obtained from Equation (A.11) by setting the first number of samples to infinity, reducing the noise in $\mathbf{X}_1^{-1}$ to zero.

For the further derivations, it is convenient to “normalize” the covariance. Let us take

$$\mathbf{y}_i = \Psi_0^{1/2} \mathbf{x}_i, \quad (\text{A.13})$$

where $\Psi_0^{1/2}$ means the matrix square root of Ψ_0 – a matrix with the same eigenvectors and eigenvalues equal to the square roots of corresponding eigenvalues of the original matrix. Then

$$\text{cov}(y_{a,i}, y_{b,j}) = \delta^{ij} \delta^{ab}. \quad (\text{A.14})$$

Let us also introduce $\bar{y}_a \equiv \frac{1}{n_S} \sum_{i=1}^{n_S} y_{a,i}$, define $\delta y_{a,i} \equiv y_{a,i} - \bar{y}_a$ (so that $\langle \delta y_{a,i} \rangle = 0$) and finally compute the “normalized” covariance matrix:

$$Y_{ab} \equiv \frac{1}{n_S - 1} \sum_{i=1}^{n_S} \delta y_{a,i} \delta y_{b,i}. \quad (\text{A.15})$$

Then also

$$\mathbf{Y} = \mathbf{\Psi}_0^{1/2} \mathbf{X} \mathbf{\Psi}_0^{1/2}. \quad (\text{A.16})$$

Let us compute

$$\text{cov}(\delta y_{a,i}, \delta y_{b,j}) = \langle \delta y_{a,i} \delta y_{b,j} \rangle = \delta^{ij} \delta^{ab} - 2 \times \frac{1}{n_S} \delta^{ab} + n_S \times \frac{1}{n_S^2} \delta^{ab} = \left(\delta^{ij} - \frac{1}{n_S} \right) \delta^{ab}. \quad (\text{A.17})$$

As a consequence, $\langle \mathbf{Y} \rangle = \mathbb{I}$, and we can expand

$$\mathbf{Y} = \mathbb{I} + \delta \mathbf{Y} \quad (\text{A.18})$$

while $\langle \delta \mathbf{Y} \rangle = 0$. Also,

$$\text{cov}(Y_{ab}, Y_{cd}) = \langle \delta Y_{ab} \delta Y_{cd} \rangle = \frac{\delta^{ac} \delta^{bd} + \delta^{ad} \delta^{bc}}{n_S - 1}. \quad (\text{A.19})$$

Now let us expand the KL divergence using the “normalized” covariance matrix, starting from

$$2D_{\text{KL}}(\mathbf{\Psi}_0, \mathbf{X}) = \text{tr}(\mathbf{\Psi}_0 \mathbf{X}) - N_{\text{bins}} - \ln \det(\mathbf{\Psi}_0 \mathbf{X}), \quad (\text{A.20})$$

we can write $\mathbf{\Psi}_0 = \mathbf{\Psi}_0^{1/2} \mathbf{\Psi}_0^{1/2}$, use the cyclic property of trace and determinant to arrive to

$$2D_{\text{KL}}(\mathbf{\Psi}_0, \mathbf{X}) = \text{tr}(\mathbf{Y}) - N_{\text{bins}} - \ln \det(\mathbf{Y}), \quad (\text{A.21})$$

remembering Equation (A.16). Then we expand in $\delta \mathbf{Y}$ (Equation (A.18)):

$$2D_{\text{KL}}(\mathbf{\Psi}_0, \mathbf{X}) = \text{tr}(\delta \mathbf{Y}) - \ln \det(\mathbb{I} + \delta \mathbf{Y}). \quad (\text{A.22})$$

Now using $\ln \det\{\mathbf{A}\} = \text{tr} \ln \mathbf{A}$ and expanding the second term in Taylor series up to quadratic order in $\delta \mathbf{Y}$ we obtain

$$2D_{\text{KL}}(\mathbf{\Psi}_0, \mathbf{X}) \approx \frac{1}{2} \text{tr}[(\delta \mathbf{Y})^2] = \frac{1}{2} \sum_{a,b=1}^{N_{\text{bins}}} \delta Y_{ab} \delta Y_{ab}. \quad (\text{A.23})$$

Taking the expectation value of Equation (A.23) and using Equation (A.19), one can re-derive

Equation (A.12). We will proceed to compute the variance:

$$\begin{aligned}\text{Var} \left\{ \text{tr} \left[(\delta \mathbf{Y})^2 \right] \right\} &= \left\langle \left(\text{tr} \left[(\delta \mathbf{Y})^2 \right] \right)^2 \right\rangle - \left\langle \text{tr} \left[(\delta \mathbf{Y})^2 \right] \right\rangle^2 \\ &= \sum_{a,b,c,d=1}^{N_{\text{bins}}} [\langle \delta Y_{ab} \delta Y_{ab} \delta Y_{cd} \delta Y_{cd} \rangle - \langle \delta Y_{ab} \delta Y_{ab} \rangle \langle \delta Y_{cd} \delta Y_{cd} \rangle].\end{aligned}\quad (\text{A.24})$$

Full expansion gives

$$Y_{ab} Y_{ab} Y_{cd} Y_{cd} = \frac{1}{(n_S - 1)^4} \sum_{i,j,k,l=1}^{n_S} \delta y_{a,i} \delta y_{b,i} \delta y_{a,j} \delta y_{b,j} \delta y_{c,k} \delta y_{d,k} \delta y_{c,l} \delta y_{d,l}. \quad (\text{A.25})$$

δy are normally distributed and have zero means, so for them, we can use Wick's theorem to split this into all possible pairs. ($\delta \mathbf{Y}$ also has zero mean, but not a Gaussian distribution; this is why we need to go to a deeper level.) The total number of pairs is $8!/(4! \times 2^4) = 105$, so it is easy to go over them in a computer program. Additionally, it is useful to check which ones are similar. It is apparent that the following five index permutations leave the expression unchanged: $a \leftrightarrow b$, $c \leftrightarrow d$, $i \leftrightarrow j$, $k \leftrightarrow l$ and $(a, b, i, j) \leftrightarrow (c, d, k, l)$. Finally, some of the pairs will not contribute to variance and can be excluded: contraction of a pair inside the same Y contribute to $\langle Y \rangle = \mathbb{I}$ and must be subtracted; and if all the contracted pairs correspond to Y 's with the same indices, that contributes to the mean of D_{KL} and has to be subtracted too.

We find there are 56 pair assignments contributing to $\approx 16(n_S - 1)^4 \text{Var}[D_{\text{KL}}(\Psi_0, X)]$, but no more than 8 of them are distinct after using symmetries:

$$\begin{aligned}8 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{a,j} \rangle \langle \delta y_{b,i} \delta y_{c,k} \rangle \langle \delta y_{b,j} \delta y_{c,l} \rangle \langle \delta y_{d,k} \delta y_{d,l} \rangle = \\ 8 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ij} - \frac{1}{n_S} \right) \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{bc} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{bc} \left(\delta^{kl} - \frac{1}{n_S} \right) = \\ 8N_{\text{bins}}^3 (n_S - 1)\end{aligned}\quad (\text{A.26})$$

$$\begin{aligned}
& 16 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{a,j} \rangle \langle \delta y_{b,i} \delta y_{c,k} \rangle \langle \delta y_{b,j} \delta y_{d,l} \rangle \langle \delta y_{d,k} \delta y_{c,l} \rangle = \\
& 16 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ij} - \frac{1}{n_S} \right) \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{bc} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{bd} \left(\delta^{kl} - \frac{1}{n_S} \right) \delta^{dc} = \\
& 16 N_{\text{bins}}^2 (n_S - 1)
\end{aligned} \tag{A.27}$$

$$\begin{aligned}
& 8 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{b,j} \rangle \langle \delta y_{b,i} \delta y_{c,k} \rangle \langle \delta y_{a,j} \delta y_{d,l} \rangle \langle \delta y_{d,k} \delta y_{c,l} \rangle = \\
& 8 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ij} - \frac{1}{n_S} \right) \delta^{ab} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{bc} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{ad} \left(\delta^{kl} - \frac{1}{n_S} \right) \delta^{cd} = \\
& 8 N_{\text{bins}} (n_S - 1)
\end{aligned} \tag{A.28}$$

$$\begin{aligned}
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{c,k} \rangle \langle \delta y_{b,i} \delta y_{d,k} \rangle \langle \delta y_{a,j} \delta y_{c,l} \rangle \langle \delta y_{b,j} \delta y_{d,l} \rangle = \\
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{bd} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{bd} = \\
& 4 N_{\text{bins}}^2 (n_S - 1)^2
\end{aligned} \tag{A.29}$$

$$\begin{aligned}
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{c,k} \rangle \langle \delta y_{b,i} \delta y_{d,k} \rangle \langle \delta y_{a,j} \delta y_{d,l} \rangle \langle \delta y_{b,j} \delta y_{c,l} \rangle = \\
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{bd} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{ad} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{bc} = \\
& 4 N_{\text{bins}} (n_S - 1)^2
\end{aligned} \tag{A.30}$$

$$\begin{aligned}
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{c,k} \rangle \langle \delta y_{b,i} \delta y_{c,l} \rangle \langle \delta y_{a,j} \delta y_{d,k} \rangle \langle \delta y_{b,j} \delta y_{d,l} \rangle = \\
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{il} - \frac{1}{n_S} \right) \delta^{bc} \left(\delta^{jk} - \frac{1}{n_S} \right) \delta^{ad} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{bd} = \\
& 4 N_{\text{bins}} (n_S - 1)
\end{aligned} \tag{A.31}$$

$$\begin{aligned}
& 8 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{c,k} \rangle \langle \delta y_{b,i} \delta y_{c,l} \rangle \langle \delta y_{a,j} \delta y_{d,l} \rangle \langle \delta y_{b,j} \delta y_{d,k} \rangle = \\
& 8 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{il} - \frac{1}{n_S} \right) \delta^{bc} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{ad} \left(\delta^{jk} - \frac{1}{n_S} \right) \delta^{bd} = \\
& 8N_{\text{bins}}(n_S - 1)
\end{aligned} \tag{A.32}$$

$$\begin{aligned}
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \langle \delta y_{a,i} \delta y_{c,k} \rangle \langle \delta y_{b,i} \delta y_{d,l} \rangle \langle \delta y_{a,j} \delta y_{c,l} \rangle \langle \delta y_{b,j} \delta y_{d,k} \rangle = \\
& 4 \sum_{a,b,c,d=1}^{N_{\text{bins}}} \sum_{i,j,k,l=1}^{n_S} \left(\delta^{ik} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{il} - \frac{1}{n_S} \right) \delta^{bd} \left(\delta^{jl} - \frac{1}{n_S} \right) \delta^{ac} \left(\delta^{jk} - \frac{1}{n_S} \right) \delta^{bd} = \\
& 4N_{\text{bins}}^2(n_S - 1)
\end{aligned} \tag{A.33}$$

Gathering all together gives

$$\begin{aligned}
\text{Var} \left\{ \text{tr} \left[(\delta \mathbf{Y})^2 \right] \right\} & \approx \frac{4N_{\text{bins}}}{(n_S - 1)^3} [2N_{\text{bins}}^2 + 4N_{\text{bins}} + 4 + (N_{\text{bins}} + 1)n_S] \\
& = \frac{4N_{\text{bins}}}{(n_S - 1)^3} [(N_{\text{bins}} + 1)(n_S + 2N_{\text{bins}} + 2) + 2].
\end{aligned} \tag{A.34}$$

Then, according to Equation (A.23), the variance of the KL divergence is approximately 16 times smaller:

$$\text{Var}[D_{\text{KL}}(\Psi_0, \mathbf{X})] \approx \frac{N_{\text{bins}}[(N_{\text{bins}} + 1)(n_S + 2N_{\text{bins}} + 2) + 2]}{4(n_S - 1)^3}. \tag{A.35}$$

This is an approximation because in Equation (A.23) the Taylor expansion was truncated at the quadratic order in $\delta \mathbf{Y}$. Other derivation steps are exact.

A.2.2 Inverse test

We considered how different χ^2 would result from one matrix compared to the other. If $\mathbf{u}$ is an unit vector in N_{bins} -dimensional space ($\mathbf{u}^T \mathbf{u} = 1$), then $\mathbf{w} = \mathbf{C}_2^{1/2} \mathbf{u}$ gives an unit χ^2 according to C_2 : $\mathbf{w}^T \mathbf{C}_2^{-1} \mathbf{w} = 1$. Here $\mathbf{C}_2^{1/2}$ means the matrix square root of $\mathbf{C}_2$ – a matrix with the same eigenvectors and eigenvalues equal to the square roots of corresponding eigenvalues of the original matrix. Then we consider the χ^2 with respect to $\mathbf{C}_1$ (Ψ_1): $\mathbf{w}^T \Psi_1 \mathbf{w}$ and subtract

the expected value of 1:

$$\mathbf{w}^T \boldsymbol{\Psi}_1 \mathbf{w} - 1 = \mathbf{u}^T \left[\mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right] \mathbf{u}. \quad (\text{A.36})$$

Taking the RMS over all directions of $\mathbf{u}$, one arrives at the RMS eigenvalue of this matrix, which can be expressed through the Frobenius norm:

$$R_{\text{inv}}(\boldsymbol{\Psi}_1, \mathbf{C}_2) = \frac{1}{\sqrt{N_{\text{bins}}}} \left\| \mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right\|_F. \quad (\text{A.37})$$

The Frobenius norm can be recast as a trace and simplified further using its cyclic property:

$$\begin{aligned} \left\| \mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right\|_F &= \text{tr} \left[\left(\mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right)^T \left(\mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right) \right] \\ &= \text{tr} \left[\left(\mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - \mathbb{I} \right)^2 \right] \\ &= \text{tr} \left(\mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2 \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} - 2 \mathbf{C}_2^{1/2} \boldsymbol{\Psi}_1 \mathbf{C}_2^{1/2} + \mathbb{I} \right) \\ &= \text{tr} (\boldsymbol{\Psi}_1 \mathbf{C}_2 \boldsymbol{\Psi}_1 \mathbf{C}_2 - 2 \boldsymbol{\Psi}_1 \mathbf{C}_2 + \mathbb{I}) = \text{tr} \text{tr} \left[(\boldsymbol{\Psi}_1 \mathbf{C}_2 - \mathbb{I})^2 \right]. \end{aligned} \quad (\text{A.38})$$

This allows us to compute the quantity as

$$R_{\text{inv}}(\boldsymbol{\Psi}_1, \mathbf{C}_2) = \sqrt{\frac{\text{tr} \left[(\boldsymbol{\Psi}_1 \mathbf{C}_2 - \mathbb{I})^2 \right]}{N_{\text{bins}}}}, \quad (\text{A.39})$$

which is more computationally robust – it is better to avoid factorizing (and inverting) covariance matrices (especially the sample one) when possible.

It is notable that this metric is related to the discriminant matrix

$$\mathbf{P} = \sqrt{\boldsymbol{\Psi}_R}^T \mathbf{C}_S \sqrt{\boldsymbol{\Psi}_R} - \mathbb{I}, \quad (\text{A.40})$$

where $\sqrt{\boldsymbol{\Psi}_R}$ is the lower Cholesky decomposition, used in Philcox et al. (2020); Philcox & Eisenstein (2019). Through a similar procedure, one can find that its Frobenius norm is the same as above:

$$\|\mathbf{P}\|_F^2 = \text{tr} \left[(\boldsymbol{\Psi}_1 \mathbf{C}_2 - \mathbb{I})^2 \right]. \quad (\text{A.41})$$

However, the interpretation of elements of the discriminant matrix is less clear.

Now let us consider the square of this metric (to remove the root):

$$R_{\text{inv}}^2(\Psi_0, \mathbf{X}) = \frac{1}{N_{\text{bins}}} \text{tr}[\Psi_0 X \Psi_0 X - 2\Psi_0 X + \mathbb{I}]. \quad (\text{A.42})$$

Remembering Equation (A.16), we arrive to

$$R_{\text{inv}}^2(\Psi_0, \mathbf{X}) = \frac{1}{N_{\text{bins}}} \text{tr}[\mathbf{Y}^2 - 2\mathbf{Y} + \mathbb{I}]. \quad (\text{A.43})$$

Furthermore, expanding in $\delta\mathbf{Y}$ (according to Equation (A.18)), we get

$$R_{\text{inv}}^2(\Psi_0, \mathbf{X}) = \frac{1}{N_{\text{bins}}} \text{tr}[(\delta\mathbf{Y})^2], \quad (\text{A.44})$$

which is similar to Equation (A.23) up to a constant factor of $1/N_{\text{bins}}$ and lack of approximations.

We can obtain the expectation value by plugging in Equation (A.19):

$$\langle R_{\text{inv}}^2(\Psi_0, \mathbf{X}) \rangle = \frac{N_{\text{bins}} + 1}{n_S - 1}. \quad (\text{A.45})$$

For variance we can use Equation (A.34):

$$\text{Var}[R_{\text{inv}}^2(\Psi_0, \mathbf{X})] = \frac{4[(N_{\text{bins}} + 1)(n_S + 2N_{\text{bins}} + 2) + 2]}{N_{\text{bins}}(n_S - 1)^3}. \quad (\text{A.46})$$

Assuming $\text{Var}[R_{\text{inv}}^2(\Psi_0, \mathbf{X})] \ll \langle R_{\text{inv}}^2(\Psi_0, \mathbf{X}) \rangle^2$, we can take the square root to estimate the mean and variance of not-squared metric as

$$\langle R_{\text{inv}}(\Psi_0, \mathbf{X}) \rangle \approx \sqrt{\langle R_{\text{inv}}^2(\Psi_0, \mathbf{X}) \rangle} = \sqrt{\frac{N_{\text{bins}} + 1}{n_S - 1}}, \quad (\text{A.47})$$

$$\text{Var}[R_{\text{inv}}(\Psi_0, \mathbf{X})] \approx \frac{\text{Var}[R_{\text{inv}}^2(\Psi_0, \mathbf{X})]}{4\langle R_{\text{inv}}^2(\Psi_0, \mathbf{X}) \rangle} = \frac{(N_{\text{bins}} + 1)(n_S + 2N_{\text{bins}} + 2) + 2}{N_{\text{bins}}(N_{\text{bins}} + 1)(n_S - 1)^2}. \quad (\text{A.48})$$

Also, it is useful to note that R_{inv} is related to the χ^2 approximation of minus log-likelihood, considering the covariance of all covariance matrix elements. This is rather easy to see with $\delta\mathbf{Y}$ expression (Equation (A.44)). Since $\delta\mathbf{Y}$ is real and symmetric, we obtain

$$N_{\text{bins}} \times R_{\text{inv}}^2(\Psi_0, \mathbf{X}) = \|\delta\mathbf{Y}\|_F^2 = \sum_{a,b=1}^{N_{\text{bins}}} (\delta Y_{ab})^2. \quad (\text{A.49})$$

From Equation (A.19) we conclude that distinct elements of $\mathbf{Y}$ matrix have zero covariance (excluding the pairs symmetric with respect to the diagonal), its diagonal elements have a

variance of $2/(n_S - 1)$ and off-diagonal elements have a variance of $1/(n_S - 1)$. Then this covariance is trivial to invert, and we have to sum the squares of the deviation of independent elements divided by their variance to get the χ^2 :

$$\begin{aligned}\chi^2 &= (n_S - 1) \left[\frac{1}{2} \sum_{a=1}^{N_{\text{bins}}} (\delta Y_{aa})^2 + \sum_{a=1}^{N_{\text{bins}}} \sum_{b=a+1}^{N_{\text{bins}}} (\delta Y_{ab})^2 \right] \\ &= \frac{n_S - 1}{2} \sum_{a,b=1}^{N_{\text{bins}}} (\delta Y_{ab})^2 = \frac{(n_S - 1) N_{\text{bins}}}{2} \times R_{\text{inv}}^2(\Psi_0, \mathbf{X}).\end{aligned}\quad (\text{A.50})$$

Since elements of $\mathbf{Y}$ are independent linear combinations of elements of sample covariance matrix $\mathbf{X}$ (via Equation (A.16), since Ψ_0 is not degenerate), the same holds for $\mathbf{X}$, but a direct computation without the “rotation” into $\mathbf{Y}$ would be significantly longer as the covariance of X_{ab} elements (Equation (A.10)) has a more generic and complex structure.

A.2.3 Mean chi-squared

We consider the sum of χ^2 associated with the deviation of data vectors in individual samples from the estimate of the average:

$$\sum_{i=1}^{n_S} \sum_{a,b=1}^{N_{\text{bins}}} (x_{a,i} - \bar{x}_a) \Psi_{0,ab} (x_{b,i} - \bar{x}_b) = (n_S - 1) \sum_{a,b=1}^{N_{\text{bins}}} \Psi_{0,ab} X_{ab} = (n_S - 1) \text{tr}(\Psi_0 X). \quad (\text{A.51})$$

Remembering Equation (A.13), we can rewrite the LHS as

$$\sum_{i=1}^{n_S} \sum_{a,b=1}^{N_{\text{bins}}} (y_{a,i} - \bar{y}_a) \delta_{ab} (y_{b,i} - \bar{y}_b) = \sum_{a=1}^{N_{\text{bins}}} \sum_{i=1}^{n_S} (y_{a,i} - \bar{y}_a)^2 \sim \sum_{a=1}^{N_{\text{bins}}} \chi^2(n_S - 1) \sim \chi^2[N_{\text{bins}} \times (n_S - 1)]. \quad (\text{A.52})$$

Therefore, the corresponding reduced χ^2 is

$$\chi_{\text{red}}^2 = \frac{1}{N_{\text{bins}}} \text{tr}(\Psi_0 \mathbf{X}). \quad (\text{A.53})$$

A.2.4 Validation of means and standard deviations

To check the theoretical results from the previous sections, we have performed a quick Monte-Carlo validation. 10,000 batches of 999 samples having 45 bins each have been generated. For simplicity, we have taken the true covariance and precision to be unity matrices $C_0 = \Psi_0 = \mathbb{I}$.

		$D_{\text{KL}}(\Psi_0, X)$	$R_{\text{inv}}(\Psi_0, X)$	$\chi_{\text{red}}^2(\Psi_0, X)$
Measurement space (45 bins)	Theoretical	0.519 ± 0.024	0.2147 ± 0.0049	1.0000 ± 0.0067
	Sampled	0.526 ± 0.023	0.2146 ± 0.0050	1.0000 ± 0.0067
Parameter space (5 proj. from 25 bins)	Theoretical	0.0075 ± 0.0028	0.078 ± 0.014	1.000 ± 0.020
	Sampled	0.0077 ± 0.0028	0.077 ± 0.014	1.000 ± 0.020

Table A.1: Theoretical versus sampled mean $\pm$ std of covariance matrix comparison metrics with the true precision matrix. A close agreement can be seen.

This should not affect the results, save for numerical instabilities in matrix operations. Full 45-bin sample covariance and precision matrices have been estimated in each batch. Then, 25 bins were selected and projected into 5 quantities, for simplicity, using 5 random orthonormal vectors as parameter derivatives. Comparison between theoretical and sampled means and standard deviations is presented in Table A.1. Differences are most pronounced in D_{KL} , but the disagreement is only in the second digit of standard deviation and fractions of standard deviation on the mean. Therefore, we report a close agreement, more than enough for the main part of the paper, where the scatter of RASCALC results is significantly larger than these standard deviations. We note that the results for D_{KL} and R_{inv} (Equations (A.12), (A.35), (A.47) and (A.48)) are approximate and we expect meaningful deviations from them, especially as N_{bins} increases, while the derivations for R_{inv}^2 and χ_{red}^2 (Equations (A.45) and (A.52)) are exact.

We have repeated this test with a realistic covariance matrix (RASCALC Average NG for pre-reconstruction) and derivatives of the observables with respect to the parameters (accordingly, for the BAO model before reconstruction) to confirm whether the comparison measures are indeed not affected. We have obtained the same numbers as in simpler test (Table A.1), and close agreement of $\sigma[\sigma(\alpha_{\text{BAO}})]/\sigma(\alpha_{\text{BAO}})$ with $[2(n_S - 1)]^{-1/2}$ according to Equation (2.50).

A.3 Covariance for the combination of two regions

In this section, we document the procedure to compute the covariance matrix for the combination of two (or more) disjoint regions (volumes) using the covariances in each region. It is

fully consistent with the correlation function treatment in the DESI data processing pipeline.

A.3.1 Angular bins

During the combination of regions labeled by G , DESI scripts simply add the total counts of every relevant kind QQ (where each Q can be D or R , data or randoms¹):

$$(Q^X Q^Y)_a^c = \sum_G (Q^X Q_G^Y)_a^c. \quad (\text{A.54})$$

The following technical details are important for the strictly correct implementation with PYCORR. The counts entering the correlation function estimator (Equation (A.1)) need to be normalized properly. Originally the main cause was the possibly different number of galaxies and random points (Landy & Szalay, 1993; Pons-Bordería et al., 1999). Now we are often weighting both types of points, and the pair counts weighted by the product of the two individual point weights can be normalized by the product of the sums of the weights². PYCORR stores both the weighted counts (**wcounts**) and their normalization (**wnorm**). It is their ratio, the normalized counts, that enter the correlation function estimator (Equation (A.1)). Before adding the pair counts for the two regions, DESI scripts bring counts of different types to the same normalization within each region (via the **normalize** method³). During the combination of regions **wnorm** are added as well as **wcounts**. Therefore, the change of normalization within each region alters the relative contributions of the regions to different types of counts, even though it does not affect the region’s correlation function. Therefore, this step is important to reproduce. With all count types brought to the same normalization (in each region and therefore in their combination), the **wnorm** cancel out in the Landy-Szalay estimator (Equation (A.1)) and so it is sufficient to substitute only **wcounts**.

Then the total correlation function (Equation (A.1)) is simply the average of the two

¹I.e. $Q^X Q^Y$ encompasses $D^X D^Y$, $D^X R^Y$, $R^X D^Y$ and $R^X R^Y$ count arrays if standard BAO reconstruction is not used. After reconstruction we also have S – shifted randoms, and the counts relevant to the Landy-Szalay estimator are $D^X D^Y$, $D^X S^Y$, $S^X D^Y$, $S^X S^Y$ and $R^X R^Y$.

²Exactly $\sum_i w_i^X \times \sum_j w_j^Y$ for cross-correlations ($X \neq Y$) and $(\sum_i w_i^X)^2 - \sum_i (w_i^X)^2$ for auto-correlations.

³By default, this method preserves the **wnorm** of $D^X D^Y$ counts.

regions' correlation functions weighted by RR counts:

$$(\xi^{XY})_a^c = \frac{1}{(R^X R^Y)_a^c} \sum_G (R^X R_G^Y)_a^c (\xi_G^{XY})_a^c \equiv \sum_G (W_G^{XY})_a^c (\xi_G^{XY})_a^c, \quad (\text{A.55})$$

$$(W_G^{XY})_a^c \equiv \frac{(R^X R_G^Y)_a^c}{(R^X R^Y)_a^c}. \quad (\text{A.56})$$

Then the covariance matrix for the combined region is simply

$$(C^{XY,ZW})_{ab}^{cd} \approx \sum_G (C_G^{XY,ZW})_{ab}^{cd} (W_G^{XY})_a^c (W_G^{ZW})_b^d, \quad (\text{A.57})$$

where we neglect the covariance between the correlation functions in different regions, the estimation of which poses extra challenges. We expect this to be a safe approximation for the North and South Galactic caps in the DESI footprint because the separation between galaxies in different caps is significantly larger than the maximum separation for correlation function measurements ($200 h^{-1} \text{Mpc}$).

A.3.2 Legendre

We can build upon the results for angular bins, with conversions both ways between them and Legendre moments. The Legendre multipoles are estimated from the angular bins via Equations (2.17) and (2.18):

$$(\xi^{XY})_a^\ell = \sum_c (\xi^{XY})_a^c F_c^\ell. \quad (\text{A.58})$$

One can do the reverse approximately with bin-averaged values of the Legendre polynomials⁴:

$$L_\ell^c \equiv \frac{1}{\Delta\mu_c} \int_{\Delta\mu_c} d\mu L_\ell(\mu) = \frac{F_c^\ell}{(2\ell+1)\Delta\mu_c}, \quad (\text{A.59})$$

$$(\xi^{XY})_a^c \approx \sum_\ell (\hat{\xi}^{XY})_a^\ell L_\ell^c. \quad (\text{A.60})$$

These do not depend on the region. Additional approximation comes from the fact that we limit the multipole index — in this work, we only consider $\ell = 0, 2, 4$.

With this, we can work out the partial derivative of the Legendre moment of the combined

⁴Or possibly with just the middles of the bins, but this way they end up very much related to already computed projection factors F_c^ℓ .

correlation function with respect to one in each of the regions in the following steps:

$$\frac{\partial(\xi^{XY})_a^\ell}{\partial(\xi_G^{XY})_a^{\ell_1}} = \sum_c \frac{\partial(\xi^{XY})_a^\ell}{\partial(\xi^{XY})_a^c} \frac{\partial(\xi^{XY})_a^c}{\partial(\xi_G^{XY})_a^c} \frac{\partial(\xi_G^{XY})_a^c}{\partial(\xi_G^{XY})_a^{\ell_1}} \approx \sum_c F_c^\ell (W_G^{XY})_a^c L_{\ell_1}^c \equiv (W_G^{XY})_a^{\ell, \ell_1}; \quad (\text{A.61})$$

the different separation bins and correlation functions stay independent.

Then the covariance matrix for the combined region's 2PCF is

$$(C^{XY, ZW})_{ab}^{\ell \ell'} \approx \sum_G \sum_{\ell_1, \ell'_1} \left(C_G^{XY, ZW} \right)_{ab}^{\ell_1 \ell'_1} (W_G^{XY})_a^{\ell, \ell_1} (W_G^{ZW})_b^{\ell', \ell'_1}, \quad (\text{A.62})$$

here we also neglect the covariance between the correlation functions in different regions.

Appendix B

Appendix to Hubble tension with clumpy recombination (Chapter 4)

B.1 Justifying simplifications

B.1.1 RECFAST vs HyREC

In the main text we used the recombination code RECFAST, as it is faster than the more precise HyREC; here we justify that this choice does not bias our results. RECFAST is sufficiently accurate for the analysis of *Planck* data, though this code will not be satisfactory for future CMB missions (Lee & Ali-Haïmoud, 2020). Moreover, highly nonstandard hydrogen densities, which can appear in the $\pm$ zones in M3, might limit the RECFAST applicability even further.

For our purposes, the change of zero-point is not the most important, as we focus on the shift introduced by clumping. Therefore, we compare the relative changes in $C_\ell^{TT/EE}$ between Λ CDM with standard recombination and M3 with $b = 1$ clumping obtained with RECFAST and HyREC (with full hydrogen model) in Figure B.1. We overlay the CMB-S4 error bars to assess the difference. There is no notable difference for $\ell \leq 2000$ -2500, so for current *Planck* data both codes can be considered equivalent. In particular, the differences between each M3 model and Λ CDM are $\Delta\chi_{Planck,RECFAST}^2 = 93$ and $\Delta\chi_{Planck,HyREC}^2 = 90$,

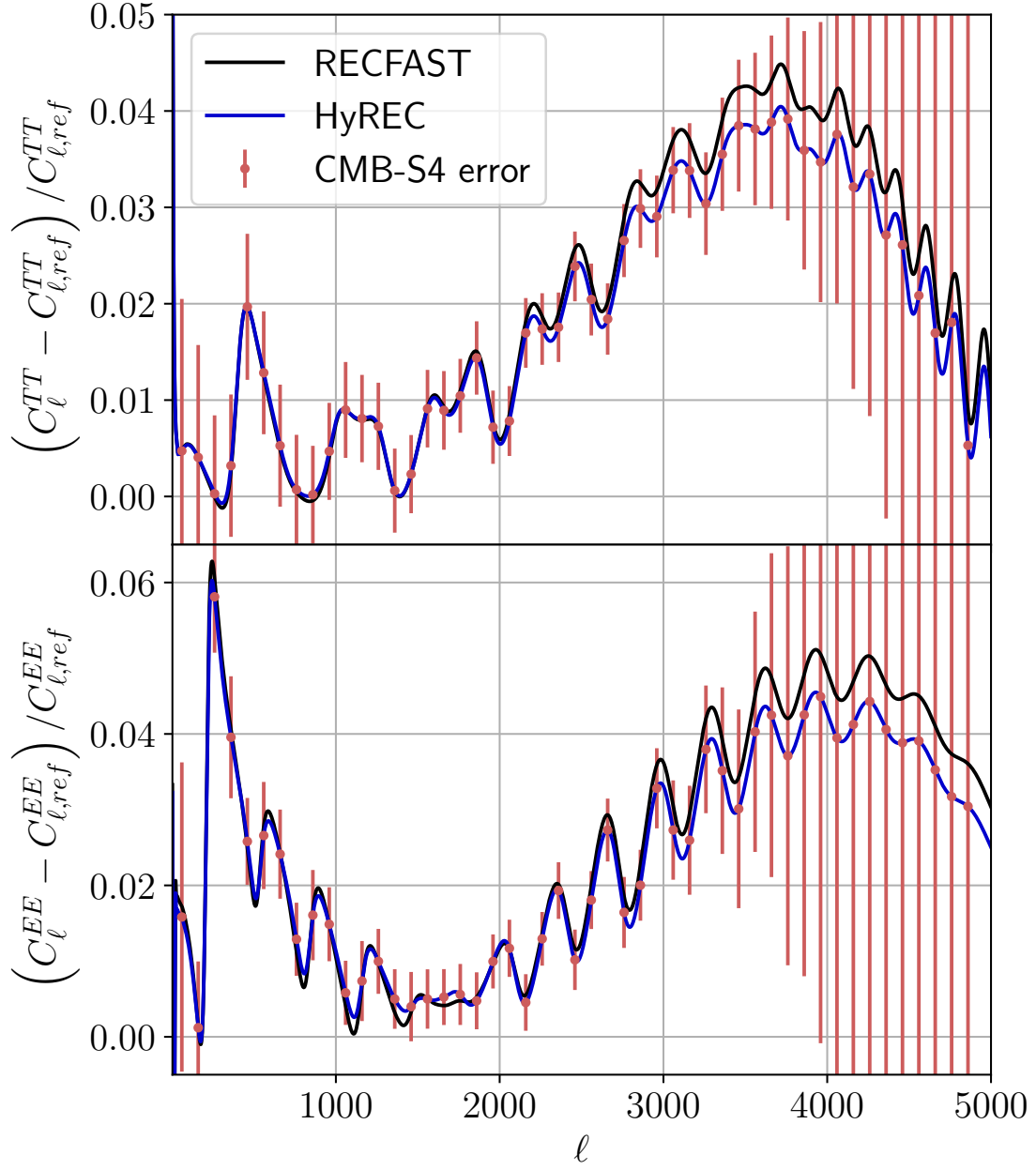


Figure B.1: Relative shifts between Λ CDM with standard recombination (taken as reference $C_{\ell,\text{ref}}$) and a particular M3 configuration ($\delta_- = -0.9$, $\delta_+ = 5/3$, $f_0 = 1/3$ giving $b = 1$, like in Figure 4.5), using RECFAST (in black) and HyREC (in blue). The red bands correspond to CMB-S4 errors binned with $\Delta\ell = 100$. The cosmological parameters (θ_s , ω_b , ω_{cdm} , A_s , n_s , τ_{reio}) are fixed to the Planck best fit.

which are very close (as we note we have not shifted any parameters here). For SO or CMB-S4, however, the difference between the recombination codes can be larger. For the lines shown, $\Delta\chi^2_{\text{CMB-S4,RECFAST}} = 1350$ and $\Delta\chi^2_{\text{CMB-S4,HyREC}} = 1150$, if one assumes the *Planck* best-fit cosmology (fixed), which shows a relative difference between the recombination codes of $\lesssim 20\%$. Near the fiducial, where most of posterior is, the absolute difference will naturally be lower. Also, for CMB-S4 precision the shifts in parameters are expected to be less than 1σ (Lee & Ali-Haïmoud, 2020). We conclude that an analysis of real SO and CMBS4 data should use HyREC, though RECFAST is sufficient for our forecasting purposes.

B.1.2 Neutrino masses

Throughout this paper we assumed massless neutrinos for efficiency, as it reduces the computational overhead by an order of magnitude. This increases our best fit H_0 compared to the *Planck* one. However, again, we are most interested in changes introduced by clumping with respect to standard recombination, so we compare them for massive and massless neutrinos in Figure B.2. The difference between both predictions in this plot is minuscule, and always smaller than even the CMB-S4 error bars. More quantitatively, for the lines shown, $\Delta\chi^2_{\text{Planck}, m_\nu=0.06\text{ eV}} = 92.6$ and $\Delta\chi^2_{\text{Planck}, m_\nu=0\text{ eV}} = 93.3$; $\Delta\chi^2_{\text{CMB-S4}, m_\nu=0.06\text{ eV}} = 1346$ and $\Delta\chi^2_{\text{CMB-S4}, m_\nu=0\text{ eV}} = 1334$ (if one assumes *Planck* best fit cosmology for fiducial). The relative difference in $\Delta\chi^2$ is $\lesssim 1\%$ in both cases. Therefore computing with massless neutrinos suffices our purposes.

B.2 Full contours from *Planck* runs

In Figure B.3 we show posteriors for all parameters in runs of M3 with *Planck* 2018 data (without and with SH0ES). The posteriors on clumping parameters $\log_{10}(-\delta_-)$, $\log_{10}|\delta_+/\delta_-|$ and f_0 are largely flat, like the priors, except the decrease for higher $|\delta_-|$ for *Planck* and increase in the same place for *Planck*+SH0ES. The clumping parameters also show almost no correlations with standard cosmological parameters, except a weak H_0 increase for the highest $|\delta_-|$. Addition of SH0ES causes some increase in n_s , $\Omega_b h^2$, H_0 ; a weak increase in A_s and τ_{reio} ;

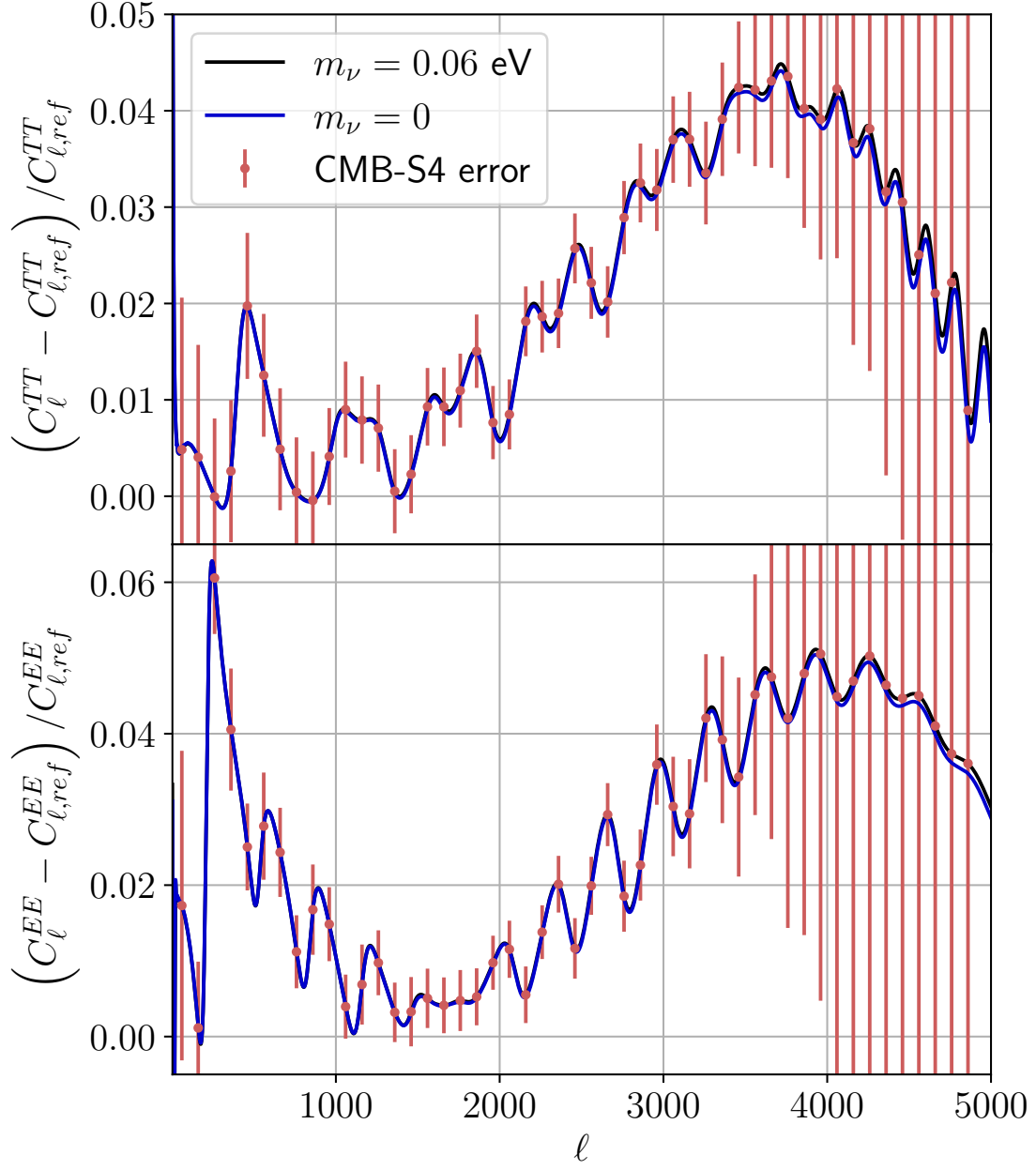


Figure B.2: Same as Figure B.1 but comparing the cases of massive (black) and massless (blue) neutrinos.

some decrease in $\Omega_{\text{cdm}} h^2$.

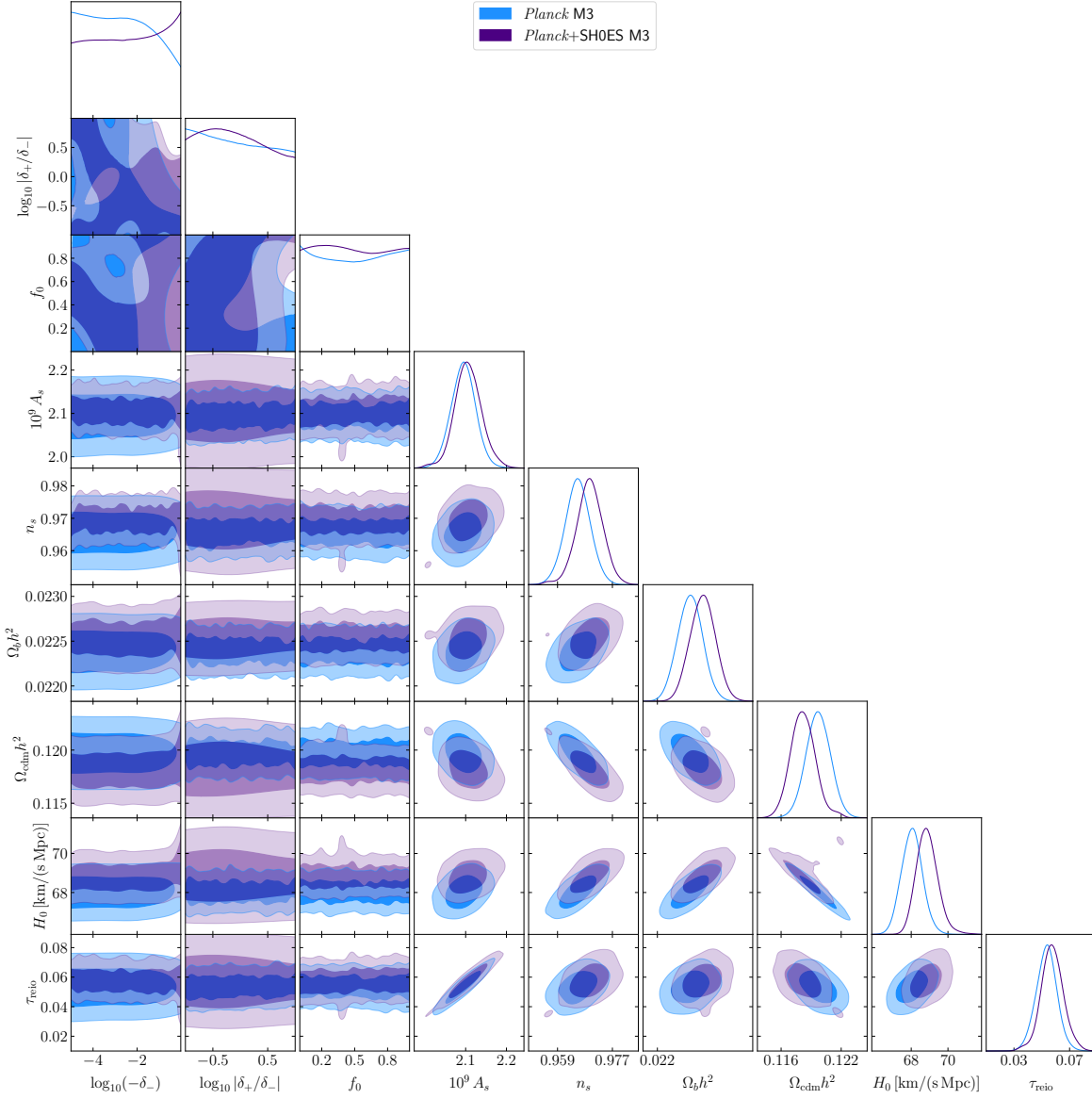


Figure B.3: Full contour plot for *Planck* runs of M3. Purple is with SH0ES, blue is without.

B.3 Best fit parameters

In Table B.1 we show the best fit parameters when fitting with full *Planck* data. The M3 best fit to *Planck*-only is not shown, as we have not found a better one than Λ CDM, and the Λ CDM is included in M3 when one sets either one of the δ 's to 0 or $f_0 = 1$. Adding SH0ES in Λ CDM naturally increases H_0 as well as most input parameters: A_s , n_s , θ_s , ω_b , τ_{reio} ; whereas ω_{cdm}

Model	Λ CDM	Λ CDM	M3
Fit to	<i>Planck</i>	<i>Planck</i> +SH0ES	
δ_-	n/a (0)	n/a (0)	-0.955
δ_+	n/a (0)	n/a (0)	1.320
f_0	n/a (1)	n/a (1)	0.652
b	n/a (0)	n/a (0)	0.439
$10^9 A_s$	2.1094	2.1440	2.1132
n_s	0.96604	0.97084	0.96552
$100\theta_s$	1.04192	1.04207	1.04177
$\Omega_b h^2$	0.022416	0.022569	0.022714
$\Omega_{\text{cdm}} h^2$	0.11945	0.11762	0.11999
τ_{reio}	0.0514	0.0555	0.0542
H_0 [km/(s Mpc)]	68.146	68.993	70.916
Ω_K	0		
m_ν [eV]	0		
$\chi^2_{\text{low } \ell TT}$	23.2	22.4	23.6
$\chi^2_{\text{low } \ell EE}$	395.7	396.1	395.9
$\chi^2_{\text{high } \ell TTTEE}$	582.2	584.0	585.4
χ^2_{lensing}	9.0	8.7	8.8
χ^2_{Planck}	1010.0	1011.1	1013.7
χ^2_{SH0ES}	(15.1)	10.5	3.1

Table B.1: Our best fit parameters to (full) *Planck* and *Planck*+SH0ES.

decreases slightly. M3 allows for larger H_0 , while the increase in A_s and τ_{reio} become smaller, n_s and θ_s decrease very slightly, $\Omega_b h^2$ increases further than in Λ CDM and $\Omega_{\text{cdm}} h^2$ increases, unlike in Λ CDM. The CMB χ^2 difference is dominated by high- ℓ TT, TE, EE (where “high” means $\ell > 30$).

In Table B.2 we show the best fit parameters when fitting with *Planck* $\ell < 1000$ data. The shifts in cosmological parameters are similar and small. Interestingly, the addition of SH0ES helped to find a better fit to *Planck* than Λ CDM, unlike with *Planck*-only data. This is likely because the optimal parameters region with high clumping is small. However, the fit improvement is not significant. The CMB χ^2 difference is also dominated by high- ℓ TT, TE, EE (where “high” now means $30 < \ell < 1000$).

Model	Λ CDM	Λ CDM	M3
Fit to	<i>Planck</i> $\ell < 1000$	<i>Planck</i> $\ell < 1000 + \text{SH0ES}$	
δ_-	n/a (0)	n/a (0)	-0.950
δ_+	n/a (0)	n/a (0)	1.196
f_0	n/a (1)	n/a (1)	0.301
b	n/a (0)	n/a (0)	0.794
$10^9 A_s$	2.0918	2.1315	2.0705
n_s	0.97026	0.97711	0.95944
$100\theta_s$	1.04150	1.04179	1.04034
$\Omega_b h^2$	0.022537	0.022773	0.022700
$\Omega_{\text{cdm}} h^2$	0.11831	0.11613	0.12198
τ_{reio}	0.0518	0.0553	0.0505
H_0 [km/(s Mpc)]	68.528	69.640	72.616
Ω_K	0		
m_ν [eV]	0		
$\chi^2_{\text{low } \ell TT}$	22.4	21.4	24.8
$\chi^2_{\text{low } \ell EE}$	395.7	395.9	395.6
$\chi^2_{\text{high } \ell TTTEEE}$	286.3	288.2	281.5
χ^2_{lensing}	8.9	9.0	8.8
χ^2_{Planck}	713.2	714.4	710.7
χ^2_{SH0ES}	(12.9)	7.5	0.2

Table B.2: Our best fit parameters to *Planck* $\ell < 1000$, without and with SH0ES.

Appendix C

Appendix to thermal Sunyaev-Zeldovich galaxy selection (Chapter 5)

C.1 Distribution of numbers of close neighbors for LRG in different tSZ SNR bins

We provide details on the distributions of the numbers of close neighbors in different tSZ SNR bins in Figure C.1. “Fraction relative to all” is

$$\frac{N(\text{galaxies in SNR bin with } k \text{ neighbors})}{N(\text{galaxies in SNR bin})} \times \frac{N(\text{all galaxies})}{N(\text{all galaxies with } k \text{ neighbors})} \quad (\text{C.1})$$

We obtained them by

- finding all the unique close galaxy pairs obeying the desired condition (Section C.2 details the numerical method);
- counting occurrences of each galaxy in the resulting array of pair member indices, which gives the number of close neighbors for that galaxy;
- counting how many galaxies have each number of neighbors (discrete histogram, e.g.

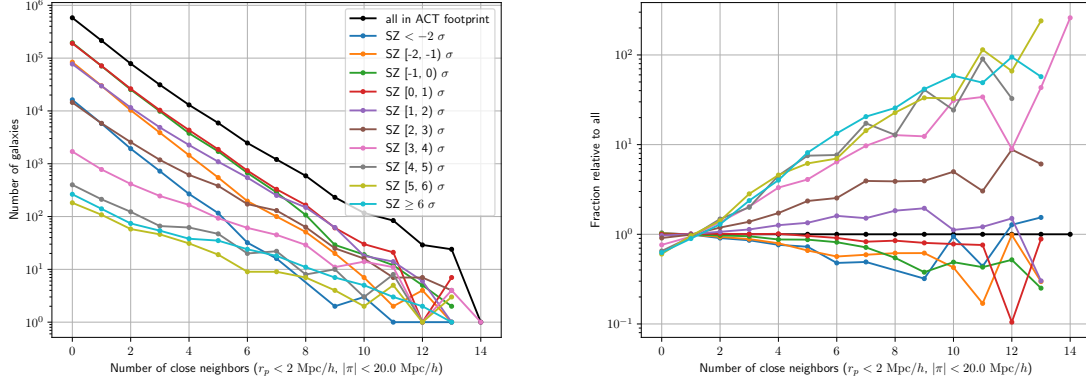


Figure C.1: Distribution of numbers of close neighbors for LRG in different tSZ SNR bins. Left: absolute number of galaxies in each bin with a given number of close neighbors. The neighbors counted are LRGs belonging to any bin. Right: an attempt to highlight the differences between different bins. “Fraction relative to all” is the fraction of galaxies with a given number of close neighbors in a given SNR bin, divided by the fraction of galaxies with the same number of close neighbors among the full LRG sample; see also Equation (C.1).

`numpy.bincount`) for the full LRG sample;

- dividing the galaxies into SNR bins and repeating the previous step in each bin.

C.2 Efficient method for near-neighbor search in cylinders

In this section, we describe our methodology for efficient search of close-neighbor galaxy pairs with different limits on parallel ($r_{\parallel, \text{max}}$) and perpendicular ($r_{\perp, \text{max}}$) separations within the pair. We define $r_{\parallel}$ and $r_{\perp}$ as components of the pair separation vector $\mathbf{r}_2 - \mathbf{r}_1$ with respect to the line-of-sight direction $\hat{n}$. In periodic boxes, this direction is typically fixed (flat-sky approximation). In the more complex case of a large realistic survey accounting for the curved sky, we choose $\hat{n}$ along the direction to the midpoint of the pair, $\mathbf{r}_1 + \mathbf{r}_2$.

We assume flat FLRW cosmology (as is DESI fiducial cosmology for comoving distance computations and coordinate conversion), allowing us to convert to 3D Cartesian space.

We build upon the k -d (k -dimensional) tree — an efficient data structure for searching for pairs closer than a given k -d Euclidean distance, or more generally with any other p -norm

with $p \geq 1$ for k -dimensional vectors:

$$\|\mathbf{v}\|_p \equiv \left(\sum_{i=1}^k |v_i|^p \right)^{1/p}. \quad (\text{C.2})$$

The $p \rightarrow +\infty$ limit (hereafter $p = \infty$ for brevity) is the maximum norm, which will be useful:

$$\|\mathbf{v}\|_\infty \equiv \max(|v_1|, \dots, |v_k|). \quad (\text{C.3})$$

The k -d tree pair-finding algorithm relies on the fact that $\|\mathbf{x}\|_p \geq |x_i|$ for any component index i . Many other distance metrics do not have such property.

Of course, any pair of galaxies with $|r_{\parallel}| \leq r_{\parallel, \max}$ and $r_{\perp} \leq r_{\perp, \max}$ also satisfy $|\mathbf{r}_2 - \mathbf{r}_1| \leq \sqrt{r_{\parallel, \max}^2 + r_{\perp, \max}^2}$. The latter criterion can be straightforwardly used for pair searching with a k -d tree, but this step produces extra pairs. We then need to store their information, compute $r_{\parallel}, r_{\perp}$ strictly for them, check the exact condition, and discard a large fraction of the pre-selected pairs (especially if the cylinder aspect ratio is very unequal, like $r_{\parallel, \max} = 10 \times r_{\perp, \max}$ in our use cases). Next, we explain several ideas to optimize the pair pre-selection with k -d tree, e.g., for the sake of memory usage.

C.2.1 Fixed line of sight

With a fixed line-of-sight direction, we can align one of the coordinate axes with it by rotation if they do not match already. Let us assume $\hat{n} = \hat{z}$ in the remainder of this subsection. Then we can scale the parallel and perpendicular separations independently via

$$x' = a_{\perp} x, \quad (\text{C.4})$$

$$y' = a_{\perp} y, \quad (\text{C.5})$$

$$z' = a_{\parallel} z. \quad (\text{C.6})$$

In principle, we can use $a_{\perp} = 1/r_{\perp, \max}$, $a_{\parallel} = 1/r_{\parallel, \max}$ and define a custom norm

$$\|\Delta \mathbf{r}'\|_{\text{custom}} \equiv \max \left[\sqrt{(\Delta x')^2 + (\Delta y')^2}, |\Delta z'| \right], \quad (\text{C.7})$$

which exactly defines the cylinder we want and maintains the property of being not less than the absolute difference in any coordinates, so k -d tree could work with it. However, the `scipy` (Virtanen et al., 2020; Gommers et al., 2023) implementation only supports the p -norm, Equation (C.2), including the $p = \infty$ case, Equation (C.3). This gives two options of imperfect pre-selection:

- (i) $\|\Delta \mathbf{r}'\|_\infty < 1$ criterion with $a_\perp = 1/r_{\perp, \max}$, $a_\parallel = 1/r_{\parallel, \max}$. This is, in essence, fitting a cube over a cylinder with a height (h) equal to its diameter ($2R$), and the ratio of volumes is $4/\pi$.
- (ii) using the Euclidean ($p = 2$) norm, essentially fitting a sphere over the cylinder. The optimal aspect ratio is less obvious than in the previous case, but the minimum sphere-to-cylinder volume ratio is $\sqrt{3}$ for $h/R = \sqrt{2}$. $\sqrt{3} > 4/\pi$ so this variant seems worse.

As a result, we used the first option and filtered the resulting pairs using the criterion $(\Delta x)^2 + (\Delta y)^2 \leq r_{\perp, \max}^2$, because $|\Delta z| \leq r_{\parallel, \max}$ is already guaranteed.

C.2.2 Varying (midpoint) line of sight

The problem is more challenging with varying line-of-sight direction $\hat{n}$ because it does not allow rescaling the component separately as straightforwardly.

However, it is possible to erase information along the line of sight by projecting each galaxy onto the unit sphere:

$$x' = x/r, \tag{C.8}$$

$$y' = y/r, \tag{C.9}$$

$$z' = z/r. \tag{C.10}$$

If we define $\Theta = \angle(\mathbf{r}_1, \mathbf{r}_2)$, the distance between the corresponding points $\mathbf{r}'_1, \mathbf{r}'_2$ on the unit sphere is $2 \sin(\Theta/2)$ whereas we prove that

$$r_\perp \geq \min(r_1, r_2) \sin \Theta. \tag{C.11}$$

Number of pairs	$r_{\perp} < 2, r_{\parallel} < 20$	$r_{\perp} < 3, r_{\parallel} < 30$
Simple distance	1.15×10^7	3.05×10^7
Sphere embedded in 3D	6.27×10^6	1.37×10^7
4D embedding, $p = 2$	1.16×10^6	2.70×10^6
4D embedding, $p = \infty$	1.11×10^6	2.51×10^6
Complete condition	4.40×10^5	9.25×10^5

Table C.1: Resulting number of pairs for different close-pair pre-selection methods for our two criteria for DESI DR1 LRG (curved sky) with redshifts $0.4 < z < 0.85$. The distances are in $h^{-1}\text{Mpc}$.

Thus, if we have a galaxy sample with a minimum distance to the origin $r_{\min} \neq 0$, we can use the unit sphere to apply a meaningful, albeit not a perfect pre-filter in $r_{\perp}$. To avoid excluding extra pairs, the (Euclidean) distance limit on the unit sphere should be slightly larger than $\sin \Theta_{\max} = r_{\perp, \max} / r_{\min}$. This is a 2-dimensional surface of the sphere embedded in 3-dimensional space.

Our addition is a fourth dimension for the radial coordinate:

$$w' = a \times r. \quad (\text{C.12})$$

It can allow to apply a filter in $r_{\parallel}$ because

$$|r_{\parallel}| \geq |r_2 - r_1|, \quad (\text{C.13})$$

although it is imperfect too.

Like with the fixed line of sight, for the combination of the $r_{\perp}$ and $r_{\parallel}$ constraints without re-implementing the k -d tree, we should choose between the maximum ($p = \infty$) and Euclidean ($p = 2$) norms. Similarly, the implementation with the maximum norm has more straightforward rescaling factors. Efficiency estimates are harder now, but the implementation with the maximum norm is more straightforward and it gave a marginally better pre-selection than with the Euclidean norm for our data, as we show in Table C.1. Both 4-dimensional embedding variants give a significantly stricter selection than the 3D distance filter and the $r_{\perp}$ -only selection with the sphere embedding in 3D.