THE UNIVERSITY OF CHICAGO

MODELING THE NON-LINEAR SCALES IN COSMOLOGY VIA THE GALAXY-HALO CONNECTION WITH APPLICATIONS IN DES Y3

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ASTRONOMY AND ASTROPHYSICS

BY
GEORGIOS ZACHAREGKAS

CHICAGO, ILLINOIS
AUGUST 2023

To my mother and to the memory of my grandfather,

without the support of whom I would not have completed this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I left this part to the end of my thesis writing. That is because this is supposed to be the conclusion of my PhD journey, and I assumed that it would be the most fun and easy part to write. It is very fun, but it is also harder than I expected it to be. That is, however, for a good reason. I realized that I have many people to thank for being part of my about six years in Chicago as a student, as well as before that.

I honestly cannot begin to describe how much I have changed over these six year. Remarkably, a lot of this change has taken place in the last one to two years. Part of these changes reflect how much I have grown as a scientist, which has been the slow process of learning from my collaboration with various people and groups at the University of Chicago (UChicago). The other part has to do with how much I feel that I have changed as a person in "real life" and this is the most rapid change that has taken place over the last couple of years. I am equally proud of both.

To begin with, I want to thank my advisors. When I first came to Chicago, I started working with Prof. Damiano Caprioli who was also my first contact with UChicago before I left my home country. I feel that I owe a lot to Prof. Caprioli for helping me through the tough first two years in Chicago and for giving me exciting projects to work on. I especially appreciated his dedication to discussing science with me for hours on end even after already long and exhausting days of back-to-back meetings and teaching.

After a while I decided that I wanted to explore more of the science that takes place at the University and thus I started attending Prof. Wayne Hu's group meetings. Soon after, I started collaborating on a project with Prof. Hu and Dr. Marco Raveri, who was a postdoc at the time. I learned a lot about statistics as a result which I needed the most at that time in order to delve deeper into Cosmology later. I want thus to express my gratitude to them for being willing to work with me.

Later on, I started collaborating with Prof. Chihway Chang, who became my new advisor

and with whom I have done the bulk of my work as a graduate student at UChicago. The projects I worked on with Chihway and the "survey science group" were, to be honest, very different from what I was expecting to work on when I arrived at Chicago. However, I could not be more thankful for doing that, as I soon realized that I enjoy using data to extract information from it just as as much as I enjoy doing theoretical work. Related to that, I want to thank here my undergraduate advisor, Prof. Loukas Vlahos, who taught me to keep an open mind before I left for Chicago.

Through my work with Chihway I was introduced to the Dark Energy Survey (DES) collaboration. I have met and collaborated with so many scientists in DES at this point that any list of names I might attempt to create will definitely be incomplete if I want to keep this part of the thesis reasonably brief. Instead, I will say a big Thank You to the whole collaboration, and especially to Dr. Judit Prat and Prof. Bhuvnesh Jain with whom I have had the most collaboration.

I want now to especially thank my family. Especially, I want to thank my mother whose support is the driving force behind all that I have achieved. If it was not for that support, I would not have been able to join UChicago. My grandfather also played an equally important role in my journey and so I want to say an especially big thank you to him. I dedicate this thesis to my mother and to the memory of my grandfather.

While at UChcago, my life was made so much more fun due to my friends. I was lucky enough to have met a lot of people I can call friends which is a great achievement for me, given how much of an introvert I am. I want to express my deep gratitude to all of them.

Last but not least, I want to thank the administration staff at UChicago who took care of countless small tasks for me, making sure I was on track to graduate.

ABSTRACT

Our concordance cosmological model, known as ΛCDM, is remarkably successful in explaining a wide range of observations. Using data to fit that model to we are able to constrain its parameters. In the era of precision cosmology, we have reached a point where we can constrain those parameters remarkably well and as we move into the next generation of galaxy surveys, we expect to be able to do even better. However, this also means that we need to build models that keep up with the increasing precision of the data. Moreover, we need to be able to utilize the whole data set, which we are not capable of doing well enough presently. Specifically, cosmological studies from galaxy surveys by collaborations such as the Dark Energy Survey (DES) only use the large, linear scales due to our inability to accurately model the non-linear regime. The removed scales, however, contain a wealth of information that could lead to increased constraining power on cosmology, as well as on the physics of how galaxies relate to their dark matter halos, which is a key component of the ΛCDM model. Furthermore, the ΛCDM model itself can be stress-tested on the non-linear scales, which can reveal more about the nature of the dark matter and even put General Relativity to test. It becomes clear, therefore, that we need to model small scales if we want to construct a complete and comprehensive picture of the cosmological model of the Universe. In this thesis, we focus on developing galaxy-halo connection models in order to model the non-linear scales in 2-point statistics in Cosmology. We use our framework to analyze data from the Year 3 (Y3) of the DES, self-consistently describing the large scales used in cosmological studies and the small scales that probe the connection between galaxies and the dark matter distribution in the Universe. We successfully obtain constraints on the relation between galaxies and their dark matter host halos in DES Y3 at fixed cosmology. Our model can be utilized in the future to perform studies where, in contrast, we fix this "galaxy-halo connection" to model all scales in the 2-point statistics of interest to constrain the cosmological parameters, or where both sets of parameters vary at the same time.

CHAPTER 1

INTRODUCTION

1.1 Cosmological preliminaries

Throughout this thesis we assume the standard cosmological model of the Universe. Based on this model, the cosmological background is isotropic and homogeneous, and our expanding Universe consists of cold dark matter (CDM) with a cosmological constant Λ; this is the so-called ΛCDM model (Peebles, 1984; Carroll et al., 1992; Peebles & Ratra, 2003). Thus, we express all formulas using the solutions of General Relativity (GR) under these assumptions and using the Friedmann-Robertson-Walker (FRW) metric, namely $g_{\mu\nu}$ with $\mu, \nu = 0, 1, 2, 3$. In this cosmology, the homogeneous and isotropic hypersurfaces of constant time evolve with time according to the scale factor of the Universe, $a(t)$, which only depends on time t.

According to General Relativity, two events in spacetime with coordinates that differ by dx^μ are separated by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu = c^2 dt^2 - a^2(t) d\ell^2 \ , \tag{1.1}$$

where we have used the FRW metric, and where c is the speed of light. In more detail, isotropy requires that clocks can be synchronized such that $g_{0i} = 0$, while the spatial part of the metric, g_{ij} $(i, j = 1, 2, 3)$, can only contract isotropically with the scale factor such that $g_{ij} dx^i dx^j = a^2(t) d\ell^2$, where $d\ell$ is the line element of the isotropic and homogeneous three-space. In addition, the fundamental postulates of the concordance cosmological model state that: *when averaged over sufficiently large scales, there exists an average motion of matter and radiation in the Universe with respect to which all averaged observables are isotropic, and that these averaged observables are the same for all fundamental observers (homogeneity).* These postulates then lead to the conclusion that the mean motion is described by $dx^i = 0$, in which case $ds^2 = g_{00} dt^2 = c^2 dt^2$; for the last equality, we have equated the eigentime of the fundamental observers to the cosmic time, i.e. $g_{00} = c^2$.

The above spatial hypersurfaces with geometry described by the line element $d\ell^2$ can be either flat or curved, with the only constraint being isotropy, i.e. they need to be spherically symmetric. Also, due to homogeneity we can arbitrarily select a point of origin for our coordinate system and then choose two angles, θ and ϕ, with which to uniquely identify positions on the unit sphere and a radial coordinate R. In this case, the most general form of the spatial line element one can write down is the following:

$$d\ell^2 = dR^2 + f_K^2(R)(d\phi^2 + \sin^2\theta d\theta^2) \equiv dR^2 + f_K^2(R)d\omega^2 \ . \tag{1.2}$$

As for the radial function $f_K(R)$, homogeneity requires that it is either a trigonometric, linear, or hyperbolic function of R depending upon the curvature K; namely:

$$f_K(R) = \begin{cases} K^{-1/2}\sin(\sqrt{K}R), & K > 0 \\ R, & K = 0 \\ (-K)^{-1/2}\sinh(\sqrt{-K}R), & K < 0 \end{cases} \tag{1.3}$$

for a positive, zero, or negative curvature respectively. Note here that since $f_K(R)$ has units of length, $|K|^{-1/2}$ has also units of length.

It is worth here to define some basic concepts in cosmology that are relevant to what follows. Note that this is a very concise list of definitions, and that more can be found in pretty much any textbook or review on cosmology.

- **Redshift**: Due to the expansion of the Universe, photons that travel through space experience the phenomenon of *redshift*, i.e. their wavelength increases according to the scale factor of the Universe. If z in the redshift factor, then we have:

$$1 + z = \frac{a(t_o)}{a(t_e)} = \frac{1}{a(t_e)} \tag{1.4}$$

where t_o is the time of observation and t_e is the time of emission of the radiation. Since we usually set the scale factor $a(t_o) = 1$ at the present time (i.e. when the observation takes place at $R = 0$), we reach the last part of the equation.

- **Expansion rate**: Generally, one can derive the Friedmann equations and thus the expansion rate of the Universe starting from the Einstein Field equations,

$$G_{\mu\nu} = \frac{8\pi G}{c^2} T_{\mu\nu} + \Lambda g_{\mu\nu} \ , \tag{1.5}$$

where G is the gravitational constant, which relate the Einstein tensor $G_{\mu\nu}$ to the stress-energy tensor $T_{\mu\nu}$. In the cosmology described above, and for a perfect fluid, characterized by density ρ and pressure p, the field equations greatly simplify to the two equations below:

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda}{3} \ , \tag{1.6}$$

which is Friedman's equation, and

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda}{3} \ , \tag{1.7}$$

where Λ is the cosmological constant. From the above equations, once the value of a is fixed at one time, $a(t)$ can be determined.

- **The Hubble parameter**: We define the Hubble parameter as the relative expansion rate $H \equiv \dot{a}/a$. It's value today, $H_0 \equiv H(t_0) = (\dot{a}/a)|_{t_0=0}$, is called the *Hubble constant* and is commonly expressed in terms of the dimensionless quantity $h \equiv H_0/(100 \text{ km/s/Mpc})$.

- **Critical density**: We define the following combination of parameters as the critical

density of the Universe:

$$\rho_{\rm cr} \equiv \frac{3H_0^2}{8\pi G} \,. \tag{1.8}$$

Whether this quantity is less than or grater than or equal to unity determines whether the Universe has a closed, flat, or open geometry, as we also mention below.

- **Density parameters**: We define the density parameter $\Omega_i \equiv \rho_i/\rho_{\rm cr}$ as the mean energy density of "species" i with respect to the critical density of the Universe. In cosmological studies we usually want to constrain these parameters by fitting models to the data. For example, the matter density parameter is $\Omega_m \equiv \rho_m/\rho_{\rm cr}$, the radiation parameter is $\Omega_R \equiv \rho_R/\rho_{\rm cr}$, the curvature parameter is Ω_K and the parameter that corresponds to the dark energy (or the cosmological constant in ΛCDM) is written as $\Omega_\Lambda \equiv \rho_\Lambda/\rho_{\rm cr} = \Lambda/(3H_0^2)$. Note here that when we define the matter parameter, we assume that it includes all types of non-relativistic matter we consider to exist in the Universe, e.g. baryonic matter, cold dark matter, etc. The sum over all density parameters has to then be equal to unity:

$$\sum_i \Omega_i = 1 \tag{1.9}$$

in a flat Universe, which is what we will be considering in what follows. This is equivalent to stating that if the mean density in the Universe is equal to the critical one, the spatial geometry is that of a flat space. Figure 1.1 demonstrates the Cosmic makeup of the Universe according to what we know today based on various experiments, from which we can see that dark energy by far dominates over all other components, while Dark Matter accounts for most of the matter.

- **Spatial expansion**: Using the definitions from above, we can conveniently rewrite

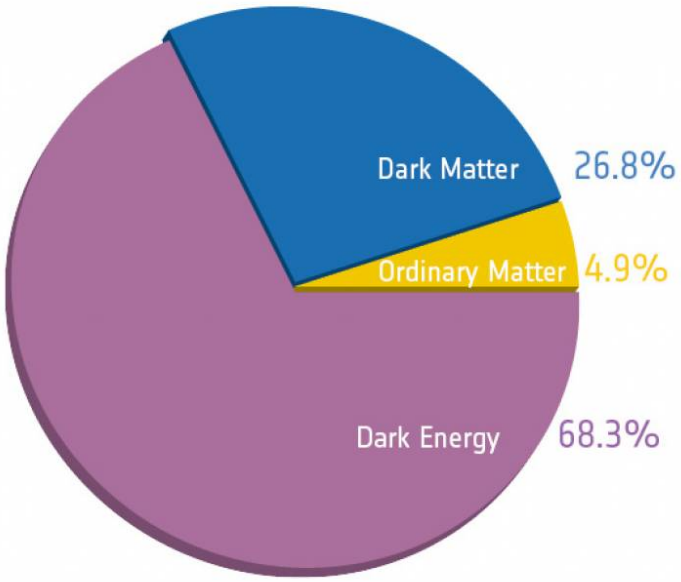


Figure 1.1: Makeup of the Universe according to the concordance cosmology. *Image credit: ESA/Planck.*

Friedmann's equation (1.6) as follows:

$$E^2(z) \equiv \frac{H^2(z)}{H_0^2} = \Omega_{R,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{K,0}a^{-2} + \Omega_\Lambda a^{-3(w+1)} \,, \tag{1.10}$$

where the subscript "0" denotes values today, at $t_0 = 0$, and we have introduced the equation-of-state parameter w for dark energy, which is defined as the ratio of the pressure over the density of dark energy. In ΛCDM $w = -1$ and thus the "cosmological constant", in which case Ω_Λ does not change with time.

- **Distance measures**: In curved spacetimes, there is no one unique measure of distance. Here we define some useful distance measures commonly utilized in cosmology. The first of them is the *proper distance*, $D_{\text{prop}}(z_1, z_2)$, which measures the distance traveled by a light ray from its emission at redshift z_2 to the observer at redshift $z_1 < z_2$, or equivalently from $a_2(t_2)$ to $a_1(t_1) > a_2(t_2)$. By assuming that the observer is at the

origin of the coordinate system, we can write:

$$D_{\rm prop}(z_1, z_2) \equiv -c \int_{t_1}^{t_2} dt = -\frac{c}{H_0} \int_{a_1}^{a_2} \frac{da}{aE(a)} = \frac{1}{H_0} \int_{z_1}^{z_2} \frac{a(z)}{E(z)} dz \ . \tag{1.11}$$

The second distance measure is the *comoving distance* which, as the name implies, is the distance on the hypersurfaces $t = t_0$ between the wordlines of the source and the observer that comoves with the cosmic flow; thus:

$$D_{\rm com}(z_1, z_2) \equiv -c \int_{t_1}^{t_2} \frac{dt}{a} = -\frac{c}{H_0} \int_{z_1}^{z_2} \frac{dz}{E(z)} \ . \tag{1.12}$$

We note here that, sometimes the *angular diameter distance*, $D_{\rm ang}$, is utilized in applications relevant to our work. We therefore define this distance measure here in analogy to the Euclidean relation between the physical cross section δA of an object at $z = z_2$ and the solid angle $\delta\Omega$ that it subtends for an observer at $z = z_1$: $\delta\Omega D_{\rm ang}^2 = \delta A$. Then, in a flat spacetime we can write it as:

$$\begin{aligned} D_{\rm ang}(z_1, z_2) \equiv \left(\frac{\delta A}{\delta\Omega}\right)^{1/2} &= a(z_2) f_K(D_{\rm com}(z_1, z_2)) \\ &= a(z_2) D_{\rm com}(z_1, z_2) = \frac{D_{\rm com}(z_1, z_2)}{1 + z_2} \ . \end{aligned} \tag{1.13}$$

1.2 The Large-Scale Structure of the Universe

1.2.1 Linear growth of perturbations

According to the concordance cosmology briefly discussed above, the structure we observe in the Universe today is formed from small initial perturbations in the matter field. In fact, according to the inflationary paradigm, these perturbations are due to quantum fluctuations which rapidly grew during the inflationary expansion of the Universe at the Big Bang (see figure 1.2 for an illustration of the Cosmic timeline in the ΛCDM model). Later on, due

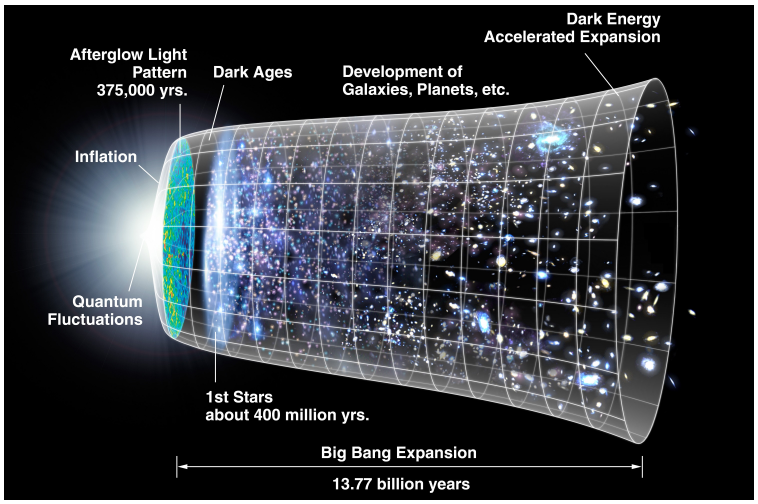


Figure 1.2: Timeline of the Universe according to ΛCDM. *Image credit: NASA/WMAP.*

to the gravitational instability, the overdensities in the field collapsed (while voids grew in size) forming the structures in the observed Universe. In this picture, the fluctuations are uncorrelated and the distribution of their amplitudes is Gaussian.

The growth of these perturbations can be sufficiently described by the linear perturbation theory as long as they are less than unity, or, in other words, as long as their amplitude is sufficiently small compared to the homogeneous and isotropic background. For a qualitative, as well as quantitative, review of the linear theory see Padmanabhan (1993). Here we will only go over some basics that are useful for this work.

As we discussed in Chapter 1.1, we usually make the assumption that our Universe is dominated by a weakly interacting dark matter component at the late times (which is mostly responsible for the weak gravitational lensing we observe). If we define the mean density for the background to be $\bar{\rho}$ and if $\rho(\mathbf{x})$ is the density at point $\mathbf{x}$, then we can define the fluctuations in density as:

$$\delta(\mathbf{x}) \equiv \frac{\rho(\mathbf{x})}{\bar{\rho}} - 1 \, . \tag{1.14}$$

Then, the applicability of the linear theory is restricted to $|\delta| \ll 1$.

We can statistically describe the density field by calculating the average (denoted by $\langle \cdot \rangle$)

in real coordinate space (denoted by the subscript "c") of the product of $\delta(\mathbf{x}_i)$ at n points:

$$\langle \delta(\mathbf{x}_1)\delta(\mathbf{x}_2)...\delta(\mathbf{x}_n)\rangle_c \equiv \xi_n(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) \ , \tag{1.15}$$

which is the so-called n-point correlation function in real space. Usually, due to the isotropic and homogeneous nature of the density field, every $\mathbf{x}$ is the same, and thus the above breaks down to:

$$\langle \delta \rangle_c = \langle \delta \rangle \equiv \mu \tag{1.16}$$

$$\langle \delta^2 \rangle_c = \langle \delta^2 \rangle - \mu \equiv \sigma^2 \tag{1.17}$$

$$\langle \delta^3 \rangle_c = \langle \delta^3 \rangle - 3\mu\sigma^2 - \mu^3 \tag{1.18}$$

and so on. In the above, we have introduced the mean μ and variance σ of the field. Moreover, we almost always consider the case of $\mu = 0$, by definition. It is interesting to note here that, in the case of a Gaussian random field, the 2-point correlation function fully describes the statistics of the field, while according to Wick's theorem any odd-point function is zero. This means that any deviation from zero of odd-point correlation functions is an indication of non-Gaussianity.

In linear theory, we usually perform calculations in the Fourier space because operations simplify significantly. Therefore, we need to introduce the Fourier pair $\delta(\mathbf{k})$, where $\mathbf{k}$ is the wavenumber, of the density field; the two relate through the following transformation:

$$\delta(\mathbf{x}) = \int \frac{d^3\mathbf{k}}{(2\pi)^3} \delta(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}} \ . \tag{1.19}$$

We can thus now express the statistics of the fluctuations in Fourier space as:

$$\langle \delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\rangle = (2\pi)^3\delta_D(\mathbf{k}_{12})P(k_1) \tag{1.20}$$

$$\langle \delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)\rangle = (2\pi)^3\delta_D(\mathbf{k}_{123})B(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3) \tag{1.21}$$

$$\langle \delta(\mathbf{k}_1)\delta(\mathbf{k}_2)\delta(\mathbf{k}_3)\delta(\mathbf{k}_4)\rangle = (2\pi)^3\delta_D(\mathbf{k}_{1234})T(\mathbf{k}_1,\mathbf{k}_2,\mathbf{k}_3,\mathbf{k}_4) \tag{1.22}$$

where P, B and T are known, respectively, as power spectrum, bispectrum and trispectrum, and where we have used the convention

$$\delta_D(\mathbf{k}_{12\ldots n}) \equiv \int \frac{d^3\mathbf{x}}{(2\pi)^3}\exp\left[-i\mathbf{x}\cdot(\mathbf{k}_1+\mathbf{k}_2+\ldots+\mathbf{k}_n)\right] \tag{1.23}$$

for simplicity, with $\mathbf{k}_{12\ldots n} \equiv \mathbf{k}_1+\mathbf{k}_2+\ldots+\mathbf{k}_n$.

In what follows, we will be extensively using the power spectrum in our calculations and its Fourier transform counterpart 2-point function in real space

$$\xi_2(r) = \int \frac{d^3\mathbf{k}}{(2\pi)^3}P(k)e^{i\mathbf{k}\cdot\mathbf{r}} \tag{1.24}$$

(later on we will be simply writing $\xi(r)$ for the 2-point functions as this is the only n-point function we are interested in for this work). Note also that, it is common to instead of working with $P(k)$ utilize the dimensionless power spectrum which is defined as:

$$\Delta(k) \equiv \frac{k^3P(k)}{2\pi^2} \tag{1.25}$$

and expresses the power per logarithmic interval in wavenumber k. The variance of the power spectrum is the quantity:

$$\sigma^2(R) = \int d\ln k\Delta(k)|W(kR)|^2 \tag{1.26}$$

smoothed over a scale R. For the window function $W(kR)$ a top-hat function is usually assumed, of the form:

$$W_{\rm tophat}(kR) = \frac{3[\sin(kR) - (kR)\cos(kR)]}{(kR)^3} \tag{1.27}$$

or sometimes the Gaussian window function:

$$W_{\rm gauss}(kR) = \exp(-k^2R^2/2) \tag{1.28}$$

when the real-space Gaussian is $\exp[-(1/2)(r/R)^2]/\sqrt{2\pi R^2}$.

1.2.2 The halo model

In this section we go over the basics of the *halo model*, which allows us to go beyond the linear theory of structure formation. A framework similar to the modern "halo model" was developed by Neyman & Scott (1952); Neyman et al. (1953). Its modern theoretical framework was introduced in Seljak (2000); Ma & Fry (2000); Peacock & Smith (2000a). Cooray et al. (2000) was one of the earliest papers to show how this model can be used in weak lensing applications. In what follows below, we only briefly introduce the basics of the halo model as is needed to understand our modeling in Chapters 3 and 4.

According to the halo model, the dark matter in the Universe is partitioned into distinct units that we call *dark matter halos*. A dark matter halo is an overdensity of dark matter that is bound by gravity (see figure 1.3 for an example in a simulation). In ΛCDM, structure formation is hierarchical, as in small halos form first to then combine under the influence of gravity to form larger ones. Then, the small-scale structure, that is, what happens within a dark matter halo, depends on how the dark matter is spatially distributed inside a halo. On the other hand, this small-scale information is not important for the large-scale distribution of halos. In contrast to the large scales, beyond a single dark matter halo, the small-scale

regime cannot be described by the linear theory and thus needs a more advanced treatment, as we discuss in Section 1.3. In practice, that regime is usually ignored in cosmological studies which instead focus on the linear regime to constrain cosmological models.

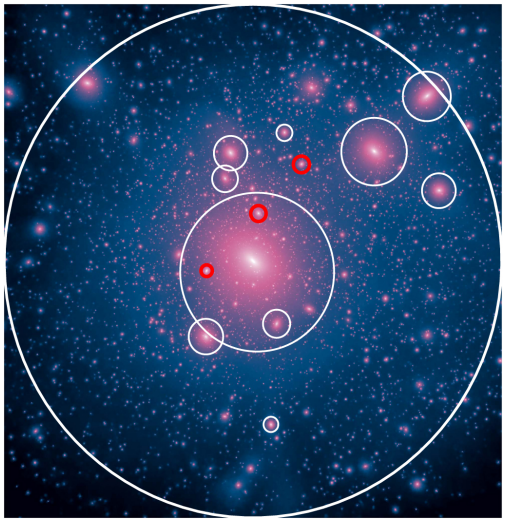


Figure 1.3: Illustration of a dark matter halo and subhalos (in circles) in simulations, from the *Caterpillar Project* (Griffen et al., 2016).

To construct a halo model, we first need to find a definition of a halo, and then model their abundance and distribution in the linear regime. If we then model how dark matter is distributed within halos, we can combine all these ingredients to form a complete model for the large-scale structure in the Universe. A commonly used approach to this is the so-called *spherical collapse model*, which is a simple, but very useful, approximation. In that model, halos are spherical object of mass $M_0 = (4\pi R_0^3/3)\bar{\rho}(1+\delta_i)$ within a comoving Lagrangian radius R_0, at a point where the initial overdensity was δ_i. As the Universe evolves over time, the radius will change according to $(R_0/R)^3 = 1+\delta$. For such an object to collapse, the initial overdensity $\delta_i \ll 1$ must increase and, in linear theory, exceed the threshold $\delta_{\rm sc}(z)$ at redshift z. Usually, we consider this threshold to be $\delta_{\rm sc}(z=0) \approx 1.686$ and is the result of the linear extrapolation of $\delta_{\rm sc}(z)$ to the present, $z = 0$.

We can now use the above model to describe the abundance of dark matter halos. Using $\delta_{\rm sc}$ we can introduce a new quantity named *peak height* that combines the critical overdensity and variance of the density field as follows:

$$\nu \equiv \frac{\delta_{\rm sc}(z)}{\sigma(M)} \tag{1.29}$$

with which to quantify the abundance of regions that have had time to collapse by z. Then, in terms of peak height, we assume that the abundance of halos of mass M that have collapsed by redshift z is described by the *Halo Mass Function* (HMF), denoted by $dn(M,z)/dM$. The formalism for the HMF that we use in this thesis is provided via the relation:

$$\frac{dn(M,z)}{d\ln M} = \frac{\bar{\rho}_m}{M}\left|\frac{dF}{dM}\right| , \tag{1.30}$$

where we have defined

$$F(M,z) = \int_{-1}^{\infty} C_{\rm coll}(\delta,z)p(\delta)d\delta$$

to be the probability that a region in an overdensity in the original overdensity field smoothed on a mass scale M, $\delta_M(\mathbf{x})$, will collapse into a halo of mass $\geq M$, by the redshift z. Also, in the above, $p(\delta)$ is the probability density function (PDF) of $\delta_M(\mathbf{x})$, while $C_{\rm coll}(\delta,z)$ is the probability that a given overdensity $\delta_M(\mathbf{x})$ at $\mathbf{x}$ will collapse by the epoch z. A popular choice of $C_{\rm coll}$ is provided by Press & Schechter (1974). In most applications, we use this *Press-Schechter ansatz* which adopts the spherical top-hat collapse model, according to which any point in space with linearly extrapolated $\delta_M(\mathbf{x})a(z) \geq \delta_{\rm sc}$ will collapse into a halo of mass $M_h \geq M$ by redshift z; in that case, $C_{\rm coll} = \Theta(\delta - \delta_{\rm sc})$, where Θ is the Heavyside step-function.

In what follows, we will be using the above ansatz and we will be adopting HMF's calibrated based on simulations. Figure 1.4 shows what the shape of the halo mass function looks like as various redshifts. As one would expect under the spherical collapse approximation, lower-mass objects are more abundant than massive ones, while there is a sharp cut-off at the very large masses. Furthermore, as time increases after the Big Bang we see more and more massive halos forming hierarchically.

Furthermore, as a consequence of the inability of the linear model to correctly capture the dark matter's distribution on the small scales, corrections to the power spectrum, $P(k)$,

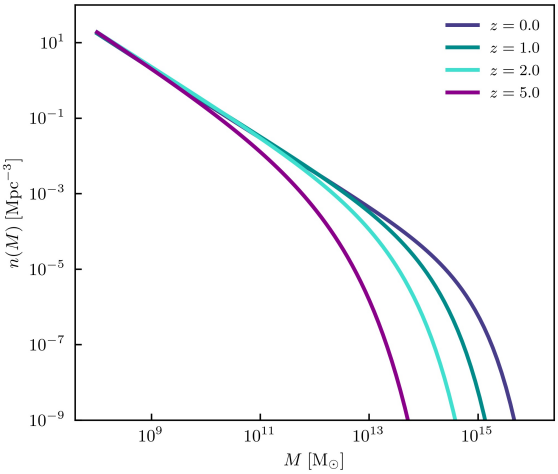


Figure 1.4: Halo mass function at various redshifts.

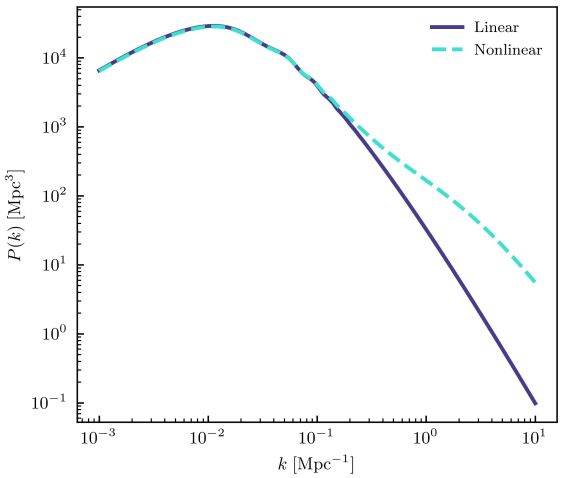


Figure 1.5: Comparison between the linear and nonlinear matter power spectrum at $z = 1$.

must be introduced on large k values. This modifications to $P(k)$ attempt to correct for the "missing power" in the small scales as predicted by the linear theory and usually take the form of fitting function based on simulations. Alternatively, one can simply utilize $P(k)$ as measured from simulations directly. Efficient emulators have been developed for this purpose (e.g. Heitmann et al., 2006; Habib et al., 2007; Heitmann et al., 2010; Heitmann et al., 2009; Heitmann et al., 2013; Kwan et al., 2013; Kwan et al., 2015; Heitmann et al., 2016; Lawrence et al., 2017); however, we do not consider such an approach in this thesis. A commonly used approach is what we call the *halofit* (see e.g. Takahashi et al., 2012) correction to the linear power spectrum, $P^{\rm L}(k)$, to obtain the non-linear power spectrum, namely $P^{\rm NL}(k)$. For an example of a comparison between the linear and nonlinear power spectrum see figure 1.5. Another example of tools that have been developed to calculate the nonlinear power spectrum is `HMCODE` (Mead et al., 2021), which is a code that fits the Halo Model to simulations. Moreover, building fast *emulators*, which is an approach in the category of *forward modeling* is becoming more popular, especially as *Machine Learning* develops rapidly and is extensively utilized in cosmological studies (see also Chapter 7 for a

relevant discussion).

1.3 The galaxy-halo connection

In this section we introduce the galaxy-halo connection model we utilize in this thesis. For influential papers on the topic we refer the reader to Benson et al. (2000); Berlind & Weinberg (2002). In what follows we briefly go over the basics that are needed to understand Chapters 3 and 4.

In order to make sense of the features of a galaxy-halo connection model, we first need to discuss how dark matter is structured within a halo. As mentioned before, a halo is the distinct unit, almost like a spheroid, of dark matter consisting of dark matter particles that are bound together by gravity. However, as it is probably clear (see also figure 1.3), there is not a unique definition of a halo. What we call a halo is a region where there is an overdensity of dark matter particles: starting from the "center" where the density is the largest we usually draw imaginary spheres around it that enclose some amount of the surrounding dark matter. If $\bar{\rho}$ is some mean density (which can be ρ_m or $\rho_{\rm cr}$), we usually define a halo to be that sphere that encloses density ρ_h that is a multiple of $\bar{\rho}$: $\rho_h = \Delta\rho$. Usual values of that multiple are: $\Delta = \{200, 300, \Delta_{\rm vir}\}$, where the Virial overdensity is $\Delta_{\rm vir}$.

The mass of a halo has a big impact on the shape of the dark matter distribution inside it, especially on how concentrated the distribution is. Higher peaks, that correspond to the more massive virialized halos, would be less concentrated than small halos, which is what is found in simulations. This dependence of the concentration of the dark matter distribution on halo mass is called the *concentration-mass relation*, and it is demonstrated in figure 1.6.

Another aspect of the halo models worth introducing here is the notion of halo biasing. The *halo bias* quantifies the excess clustering of halos over the clustering of dark matter and will be denoted by $b_h(M_h, z)$ in what follows, and is in general a function of halo mass and redshift. As shown in figure 1.7 the more massive the halo is the more biased it will be.

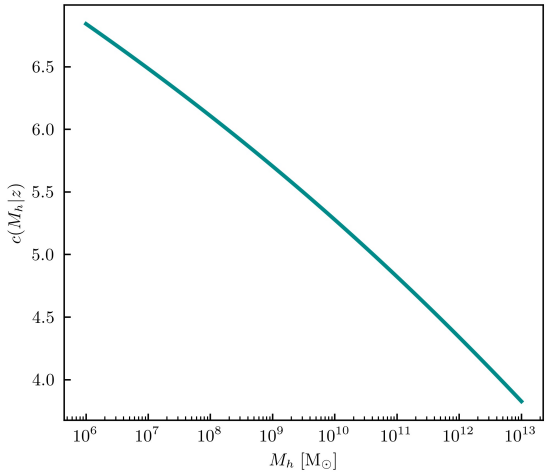


Figure 1.6: The NFW profile of a halo of mass $10^9 M_\odot$, normalized by the characteristic density ρ_s, as a function of radius in units of the scale radius r_s, at redshift $z = 1$.

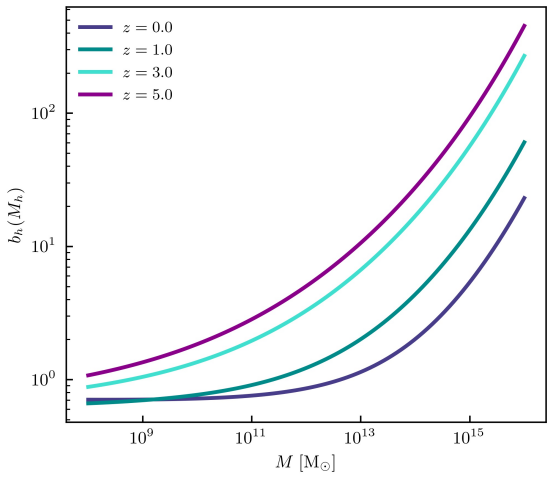


Figure 1.7: The halo bias as a function of halo mass at various redshifts.

When studying the dark matter distribution within a halo, expressions of the following form have been heavily studied:

$$\rho(r|M) = \frac{\rho_s}{(r/r_s)^\alpha (1 + r/r_s)^\beta} \ , \tag{1.31}$$

where r_s is called the *scale radius* and ρ_s is the *characteristic density* at the scale radius, while α and β are parameters that control the slope of the distribution at different radii. The most commonly used profile is the NFW (Navarro et al., 1996), for which $(\alpha, \beta) = (1, 2)$, and is shown in figure 1.8. From now on we will be assuming the NFW profile in our applications. For the concentration-mass relation we have used the fitting functions from Bhattacharya et al. (2013).

As we have previously mentioned, however, in the applications to follow, we will be working the Fourier space, and we thus need to construct the Fourier-transformed NFW profile for when we are modeling the inside of a halo. Thankfully, this is straightforward to

do using the transformation:

$$\tilde{u}(k|M_h) = \frac{\int d^3\mathbf{x}\rho(\mathbf{x}|M_h)e^{-i\mathbf{k}\cdot\mathbf{x}}}{\int d^3\mathbf{x}\rho(\mathbf{x}|M_h)} = \int 4\pi r^2 dr \frac{\sin(kr)}{kr}\frac{\rho(r|M_h)}{M_h} \tag{1.32}$$

where we have assumed spherically symmetric halos to obtain the last equality. More specifically, for the NFW profile the above simplifies even more, since we can derive the following analytic expression:

$$\begin{aligned}\tilde{u}_{\rm NFW}(k|M_h) = \frac{4\pi\rho_s r_s^3}{M_h}\Bigg\{ &\sin(kr)\left[{\rm Si}([1+c]kr_s) - {\rm Si}(kr_s)\right] - \frac{\sin(ckr_s)}{(1+c)kr_s} \\ &+ \cos(kr_s)\left[{\rm Ci}([1+c]kr_s) - {\rm Ci}(kr_s)\right]\Bigg\},\end{aligned} \tag{1.33}$$

where the sine and cosine intrgral functions are defined as:

$${\rm Si}(x) = \int_0^x \frac{\sin t}{t}dt\ ,\ \ {\rm Ci}(x) = -\int_x^\infty \frac{\cos t}{t}dt\ .$$

Figure 1.9 shows what the Fourier transform of the NFW profile looks like for halos of various mass, as a function of wavenumber.

We need to explicitly note here something that was implied before and can be seen in figure 1.3 regarding the structure in the dark matter on small scales that plays a key role in this work. Specifically, as halo is not a smooth distribution of dark matter, but rather a collection of small particles and *subhalos*, i.e. smaller halos within the central large ones. This is what we call *substructure* of the halos and we are discussing it more in what follows below. In all of the above, we focused on how the distribution of dark matter that belongs to the large central halo looks, e.g. it can follow the NFW profile, but a similar discussion needs to be had about the spatial distribution of the subhalos within a large halo.

Given the hierarchical nature of structure formation in the ΛCDM model, it is reasonable to assume that large dark matter halos will not exist completely independently from one

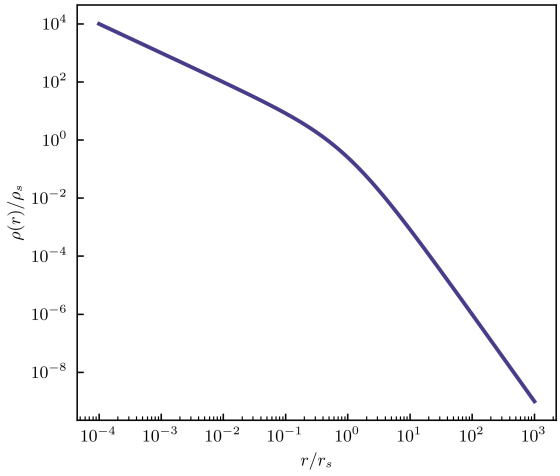


Figure 1.8: The NFW profile of a halo of mass $10^9 M_\odot$, normalized by the characteristic density ρ_s, as a function of radius in units of the scale radius r_s, at redshift $z = 1$.

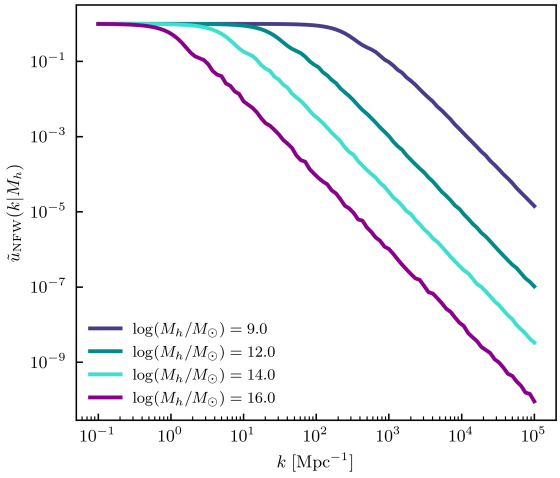
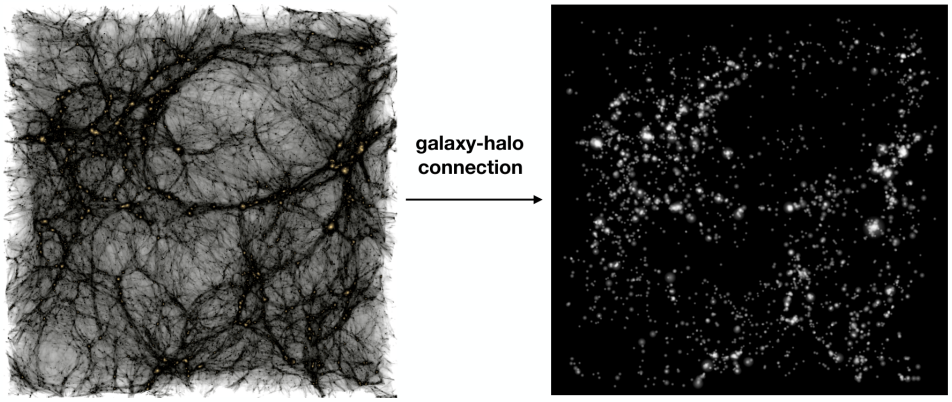


Figure 1.9: Fourier transform of the NFW profile of halos of various mass, at redshift $z = 1$.

another, but rather that larger halos will be orbited by smaller ones. The latter are commonly called *satellite halos* and the large ones in the center are the *central halos*. Moreover, in the concordance cosmological model, galaxies form within dark matter halos, the latter providing the deep gravitational potential necessary for galaxy formation. Therefore, galaxies and their properties are tightly connected to the properties of the dark matter halos they reside in, and this relation is called the *galaxy-halo connection*. In this work, we use this relation in order to study the non-linear scales, as we will discuss below. Additionally, the scales within a single dark matter halo is called the *1-halo regime*, while the large scales beyond the boundaries of a halo form the so-called *2-halo regime*. The intermediate scales, as we transition from the 1-halo to the 2-halo regime, is what we will be referring to as the *transition regime*. Modeling these regimes is all we need in order to form a complete picture to describe scales from the linear regime, which is usually the only one considered in cosmological studies, to the non-linear one which adds valuable information and constraining power (see also Chapter 2).



Approaches to modeling the galaxy-halo connection

← physical models — empirical models →

Hydrodynamical Simulations	**Semi-analytic Models**	**Empirical Forward Modeling**	**Subhalo Abundance Modeling**	**Halo Occupation Models**
Simulate halos & gas; Star formation & feedback recipes	Evolution of density peaks plus recipes for gas cooling, star formation, feedback	Evolution of density peaks plus parameterized star formation rates	Density peaks (halos & subhalos) plus assumptions about galaxy—(sub)halo connection	Collapsed objects (halos) plus model for distribution of galaxy number given host halo properties

Figure 1.10: Summary of various approaches to modeling the galaxy-halo connection. *Image credit:* Wechsler & Tinker (2018).

The spatial distribution of satellite galaxies (or equivalently of the dark matter substructure) is modeled in a very simple way in what follows. In particular, we assume that it also follows the NFW profile, but with a different concentration parameter. More specifically, due to astrophysical processes that take place on the small scales, such as baryonic feedback, the spatial distribution of satellites is less concentrated that that of the central halo's dark matter. Therefore, if $c_{\rm dm}$ and $c_{\rm sat}$ are the concentration parameters of dark matter and satellites, respectively, we make the assumption that $c_{\rm sat} = a_c c_{\rm dm}$ where generally we expect that $a_c \lesssim 1$. This assumption is based on the fact that feedback processes lead to the satellite galaxies being less concentrated compared to dark matter and this is what is found in, e.g., Carlberg et al. (1997); Nagai & Kravtsov (2005); Hansen et al. (2005); Lin et al. (2004).

To summarise, we have so far introduced the basics of structure formation, from the linear theory to the halo model that allows us to go beyond that, and we know that galaxies are connected to dark matter. To complete the picture, we still need to come up with models of the galaxy-connection itself, and then we will be able to fully model the spatial distribution of galaxies in the Universe. There are various ways of doing that, from theoretically motivated models, to semi-analytical approaches, to purely empirical ways, as summarized in figure 1.10. For a great review on this see also Wechsler & Tinker (2018). As is also demonstrated in figure 1.10, knowing how to model the galaxy-halo connection allows us to populate dark matter halos with galaxies, given the properties of the halos. In particular, the mass of the halo is one of the most important properties that determines what galaxies it is expected to reside within them. To put this in a quantitative framework, we denote the expectation number of galaxies that reside in a halo of mass M_h at redshift z as $\langle N(M_h|z)\rangle$. Then, we usually model the occupation of halos by central and satellite galaxies separately, as we discuss in more detail in Chapters 3 and 4. Modeling the two populations of galaxies in this way was introduced in Kravtsov et al. (2004). Generally speaking, more massive halos

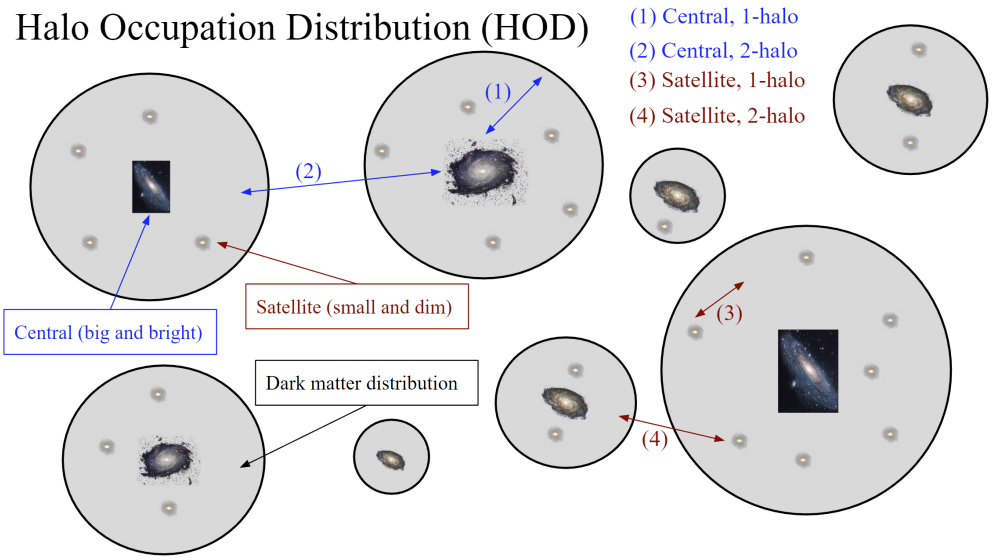


Figure 1.11: Schematic of the Halo Occupation Distribution (HOD); gray spheres represent the dark matter halos of the central galaxies, and within each halo we show the central galaxy at the center and the satellite galaxies orbiting around it.

are home to more massive central galaxies and to a larger number of satellites.

For now, we will only briefly introduce the idea of the *Halo Occupation Distribution* (HOD) which is the approach to building a framework to model the galaxy-halo connection of our choice. In short, the HOD is an empirical framework that, based on the properties of the halos and the galaxies of interest, provides us with an expected number of satellite and central galaxies. This is demonstrated in figure 1.11. Being an empirical approach, the HOD consists of a set of equations that are constructed so that we can fit the data at hand, which are only minimally motivated by theory, and is parameterized by a "small" number of parameters. However, note here that "small" is a relative term and one can make this kind of models as complex as they desire by introducing more parameters or more properties of the galaxies or halos on which the model depends. However, as we will see later, we must take caution when we select the number of free parameters to tune in our model.

The HOD framework is elegant in its simplicity and in many cases surprisingly successful in describing the data. And combined with the halo model we can build a framework to

model the clustering statistics of galaxies to compare with data to constrain models. This is exactly the approach we will be following for most of what is to follow. We will use the HOD to model galaxy-galaxy lensing and galaxy clustering (see Section 1.4) by modeling the correlation functions between central/satellite galaxies with dark matter in the 1-halo/2-halo regime. These correlations give rise to the 1-halo central, 1-halo satellite, 2-halo central and 2-halo satellite terms, as illustrated in figure 1.11. The features of each of these correlation functions is case-specific, but they generally look similar. We go into more detail on this in Sections 1.4, 3 and 4.

As a concluding remark, as we have implicitly stated before, the galaxy-halo framework gives us a way of modeling the *galaxy bias*, denoted by $b_{\rm gal}$, of a sample of galaxies (or at least this is the main way in which we are interested in using it for this work). This, of course, is simply a different way of saying that galaxies are biased tracers of the dark matter. And to be even more explicit, we can use the HOD to calculate $b_{\rm gal}$ and thus connect the spatial clustering of galaxies to that of the underlying dark matter halos. On large scales this relation simply takes the form $P_{\rm gg} = b_{\rm gal}^2 P_{\rm mm}$, where $P_{\rm gg}$ and $P_{\rm mm}$ are the power spectra of the galaxy distribution and dark matter, respectively. In the 1-halo regime this relation becomes a lot more complex, as we have to account for the specifics of the dark matter's distribution. We will omit the details here, as we go into the detail in what follows.

1.4 Correlation functions and 2-point statistics

1.4.1 Weak gravitational lensing

Gravitational lensing of light is a well-established result of the theory of General Relativity. The main idea behind this phenomenon is that as light travels through space, from its source to the observer, it experiences the gravitational potential of massive objects in the line-of-sight. Therefore, its trajectory gets deflected by some amount, $\hat{\boldsymbol{\alpha}}$, which is called

the deflection angle. This distorts the image of the observed objects, and as a result we observe their lensed form. One of the most powerful applications of this phenomenon is weak lensing, which allows us to probe the (otherwise unobservable) dark matter distribution which underlies the distribution of the visible galaxies in the Universe. This leads to a better understanding of how the properties of the galaxies are related to the properties of their host dark matter halos.

The large-scale structure in the universe generates gravitational potentials, namely Φ, that perturb the flat space-time. Usually it is the case that these potentials are weak, and more specifically the dimensionless potential $|\Phi|/c^2 \ll 1$. This is the main feature of the theory of weak gravitational lensing. Under this assumption it can be shown that the perturbed metric is diagonal, coordinate on gauge choice and non-relativistic matter, and the line element takes the form

$$ds^2 = -c^2\left(1 - \frac{2\Phi}{c^2}\right)dt^2 + \left(1 + \frac{2\Phi}{c^2}\right)d\ell^2 \,. \tag{1.34}$$

The mass distribution then has the same effect as a regular lens would have in classical optics, as we discuss briefly below. The massless photons follow null geodesics, $ds = 0$, which leads to the definition of an "index of refraction", corresponding to the mass distribution causing the lensing, as

$$n \equiv \frac{c}{v} = \left[\left(1 + \frac{2\Phi}{c^2}\right)\left(1 - \frac{2\Phi}{c^2}\right)^{-1}\right]^{1/2} \approx 1 + \frac{2\Phi}{c^2} \,, \tag{1.35}$$

where v is the speed of light in the lens. Then we can calculate the deflection angle by applying Fermat's theorem of minimizing the optical path τ, $\delta\tau = \delta\int_{\lambda_A}^{\lambda_B}(c/n)d\lambda = 0$, of the photons to get

$$\hat{\boldsymbol{\alpha}} = -\frac{2}{c^2}\int \nabla_\perp \Phi d\lambda \,, \tag{1.36}$$

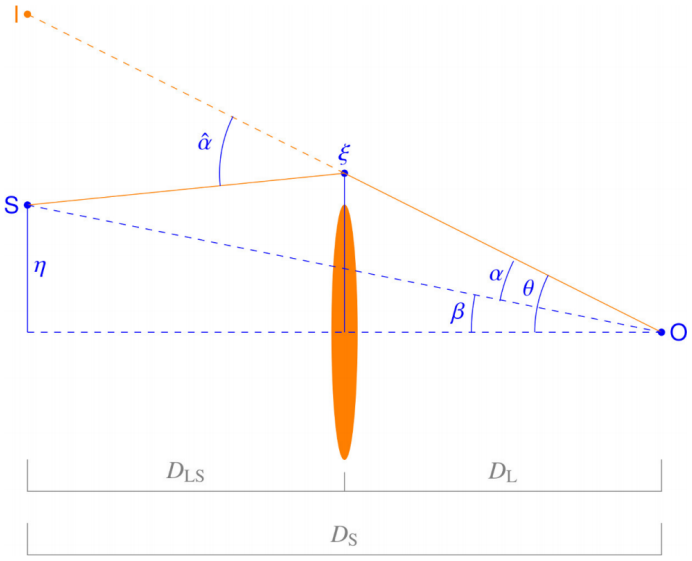


Figure 1.12: Sketch that demonstrates the deflection of light when it passes near a mass overdensity (gravitational lensing) in the weak regime.

where λ parametrizes the photon's path and $\nabla_\perp$ is taken to be perpendicular to the photon's trajectory.

As shown in figure 1.12 the deflection angle $\hat{\boldsymbol{\alpha}}$ is measured at the lens (for more see, e.g. Bartelmann & Maturi (2017)). However, it is useful to define the angle $\boldsymbol{\alpha}$ which is, instead, measured at the observer. Simple geometry and the small-angle approximation, since we are in the regime of weakly lensed systems, can relate these two angles by $\boldsymbol{\alpha} = (D_{LS}/D_S)\hat{\boldsymbol{\alpha}}$, where D_{LS} and D_S are the angular diameter distances from the lens to the source and from the observer to the source, respectively. Note here that both of these angles should be thought of as vectors since they have a direction as well as a magnitude. Conventionally, $\boldsymbol{\alpha}$ is called the "reduced" deflection angle. To understand where this definition comes from, we can use the lens equation which, according to figure 1.12, yields:

$$\boldsymbol{\theta} D_S = \boldsymbol{\beta} D_S + \hat{\boldsymbol{\alpha}} D_{LS} \quad \Rightarrow \quad \boldsymbol{\theta} = \boldsymbol{\beta} + \hat{\boldsymbol{\alpha}} \frac{D_{LS}}{D_S} \equiv \boldsymbol{\beta} + \boldsymbol{\alpha} \, .$$

Then, we can go further and write things down on a dimensionless form by defining a length

scale ξ_0 on the lens plane that has a corresponding length scale on the source plane which is η_0. The two are obviously related by $\xi_0/D_L = \eta_0/D_S$, yielding that

$$\eta_0 = \frac{D_S}{D_L}\xi_0 \ .$$

We can then use this to rewrite the deflection angle as

$$\boldsymbol{\alpha} = \frac{D_{LS}D_S}{D_L}\hat{\boldsymbol{\alpha}} \ .$$

Since we are interested in modeling the lensing signal as a function of the angular distance, $\boldsymbol{\theta}$, measured from the axis that connects the observer to the center of the lens halo, below we introduce a transformation which allow us to do so. Some manipulation of equation (1.36) combined with the transformation $\nabla_\perp = D_L^{-1}\nabla_{\boldsymbol{\theta}}$, which includes derivatives with respect to $\boldsymbol{\theta}$, yield the equation $\boldsymbol{\alpha} = \nabla_{\boldsymbol{\theta}}\hat{\Psi}$, where we define the gravitational potential as

$$\hat{\Psi}(\boldsymbol{\theta}) \equiv \frac{2}{c^2}\frac{D_{LS}}{D_L D_S}\int \Phi(\boldsymbol{\xi}, z)dz \ . \tag{1.37}$$

In the above expression we have used the z spatial coordinate as λ, D_L is the angular diameter distance from the observer to the lens and $\boldsymbol{\xi}$ is the two-dimensional vector on the lens plane. Then, $\Psi(\boldsymbol{\theta})$ can be thought of as the projected three-dimensional Newtonian potential $\Phi(\mathbf{x}) = \Phi(D_L\boldsymbol{\theta}, z)$ onto the two-dimensional plane of the lens, re-scaled by the geometric factor $D_{LS}/D_L D_S$ to account for the relative positions of the lens, the source and the observer. The dimensionless counterpart of the gravitational potential defined above is given by

$$\Psi = \frac{D_L^2}{\xi_0}\hat{\Psi} \ , \tag{1.38}$$

which would change the geometric prefactor from $D_{LS}/(D_L D_S)$ to $D_{LS}D_L/D_S$.

The lensing potential needs to satisfy two very important properties: (a) Its gradient gives

us the deflection angle, $\nabla_{\boldsymbol{\xi}}\Psi(\boldsymbol{\xi}) = \boldsymbol{\alpha}(\boldsymbol{\xi})$; (b) Its Laplacian is equal to twice the convergence, $\kappa(\boldsymbol{\xi}) \equiv \Sigma(\boldsymbol{\xi})/\Sigma_c$, and thus $\nabla^2_{\boldsymbol{\xi}}\Psi(\boldsymbol{\xi}) = 2\kappa(\boldsymbol{\xi})$. $\Sigma(\boldsymbol{\xi})$ in the definition of convergence is the surface density on the lens plane and

$$\Sigma_c(z_L, z_S) \equiv \frac{c^2}{4\pi G}\frac{D_S}{D_L D_{LS}} \tag{1.39}$$

is called the critical surface density and characterizes the lensing system.

In comoving coordinates the critical surface density will have factors of $(1+z)$ in its expression. A simple dimensional analysis should give us the correct formula. Denoting the dimensions of a quantity with square brackets, $[\Sigma_c] = M_\odot/kpc^2$ is physical units. Since the comoving length ℓ is related to the physical length x through $x = a\ell = \ell/(1+z)$, in comoving units we get $\Sigma_c^{\rm phys} \to \Sigma_c^{\rm com}/(1+z)^2$. Also, the angular diameter distance is related to the comoving distance via $D = a\chi = \chi/(1+z)$. Therefore, from equation (1.39) we get for the critical surface density in comoving coordinates

$$\Sigma_c(z_L, z_S) = \frac{c^2}{4\pi G}\frac{D_S}{(1+z_L)^2 D_L D_{LS}} = \frac{c^2}{4\pi G}\frac{\chi_S}{(1+z_L)\chi_L\chi_{LS}} \tag{1.40}$$

for a lens at redshift z_L.

In comoving coordinates equation (1.38) then gives

$$\hat{\Psi}(\boldsymbol{\theta}) \equiv \frac{2}{c^2}\int \frac{D_{LS}D_L}{D_S}\Phi(\boldsymbol{\xi}, z)dz = \frac{2}{c^2}\int \frac{\chi_{LS}\chi_L}{(1+z)^2\chi_S}\Phi(\boldsymbol{\xi}, z)d\chi\ . \tag{1.41}$$

Now we can easily also derive the convergence in comoving units, which will be

$$\begin{aligned}\kappa &= \frac{1}{2}\nabla^2\Phi = \frac{1}{c^2}\int d\chi \nabla^2\Phi\frac{\chi_{LS}\chi_L}{(1+z)^2\chi_S} = \frac{4\pi G}{c^2}\int d\chi\rho_m(z)\frac{\chi_{LS}\chi_L}{(1+z)^2\chi_S} \\ &= \frac{3}{2}\frac{H_0^2}{c^2}\Omega_{m,0}\int d\chi(1+z)\frac{\chi_{LS}\chi_L}{\chi_S}\delta_m\end{aligned} \tag{1.42}$$

using Poisson's law, $\nabla^2\Phi = 4\pi G\delta_m\rho_m(z) = 4\pi G\delta_m\Omega_{m,0}\rho_{c,0}(1+z)^3$, with the critical density being $\rho_{c,0} = 3H_0^2/(8\pi G)$, and where we have defined the mass fluctuation $\delta_m = \rho_m(z)/\bar{\rho}_m(z) - 1$ (note that $\bar{\rho}(z) = \Omega_{m,0}\rho_{c,0}(1+z)^3$).

The lensing potential is what introduces the distortion and magnification in the shape of the source galaxies as they are being lensed by the large-scale structure in the line-of-sight between them and the observer. The shear distortion on the shape of the source can be defined by introducing the angle $\boldsymbol{\beta} \equiv \boldsymbol{\theta} - \boldsymbol{\alpha} = \boldsymbol{\theta} - \nabla_{\boldsymbol{\xi}}\Psi(\boldsymbol{\xi})$ and quantifies the angular distance of the source from the axis connecting the observer to the lens center. The shear distortion on the source is then $\delta\boldsymbol{\beta} \approx A\delta\boldsymbol{\theta}$, where $A \equiv \partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}$ is the Jacobian matrix of the lens mapping. The elements of the Jacobian are $A_{ij} = \delta_{ij} - \Psi_{,ij}$, where δ_{ij} is the Kronecker delta symbol and $\Phi_{,ij} \equiv \partial^2\Psi(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$. The Jacobain matrix builds the foundations of weak lensing applications in astrophysics. By definition it maps small distortions in the image, $\delta\boldsymbol{\theta}$, to small distortions in the source, $\delta\boldsymbol{\beta}$.

It can be shown that we can decompose the Jacobian matrix into a symmetric part proportional to $(1-\kappa)$ and an anti-symmetric, trace-free part which in the appropriate coordinate system is proportional the tangential shear shear, γ_t. Therefore,

$$A = (1-\kappa)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \gamma_t \begin{pmatrix} \cos 2\phi & \sin 2\phi \\ \sin 2\phi & -\cos 2\phi \end{pmatrix} . \tag{1.43}$$

This equation makes the meaning of shear and convergence very clear: (a) Shear stretches the image of the source tangentially along a privileged direction and there exists a rotation coordinate system by an angle ϕ in which the shear matrix is proportional to γ_t; (b) Convergence, on the other hand, induces a distortion which is isotropic and just re-scales the image by a constant factor.

1.4.2 Modeling the 2-point statistics

We now discuss the convergence power spectrum, how it relates to the shear and how to apply the Limber approximation to simplify the equations (under the assumption that this is a valid approximation in our case). First, lets see what Limber's approximation is about. Lets consider for this purpose a quantity $x(\boldsymbol{\theta})$ which is a function of a two-dimensional variable $\boldsymbol{\theta}$ and write down its two-point correlation function

$$\xi_x(\boldsymbol{\phi}) = \langle x(\boldsymbol{\theta}) x(\boldsymbol{\theta} + \boldsymbol{\phi}) \rangle \ .$$

The corresponding power spectrum results from Fourier transforming that above, and thus

$$C_x(\ell) = \int d^2\boldsymbol{\phi} \ e^{-i\boldsymbol{\ell}\cdot\boldsymbol{\phi}} \xi_x(\boldsymbol{\phi})$$

In essence, Limber's approximation states that if $x(\boldsymbol{\theta})$ is the two-dimensional projection of a three-dimensional quantity, say $y(\boldsymbol{\theta})$, i.e.

$$x(\boldsymbol{\theta}) = \int_0^\chi d\chi' w(\chi') y(\boldsymbol{\theta}\chi', \chi') \ , \tag{1.44}$$

defined using the weight function $w(\chi)$, then the power spectrum of $x(\boldsymbol{\theta})$ can be calculated by

$$C_x(\ell) = \int_0^\chi d\chi' \ \frac{w^2(\chi')}{\chi'^2} P_y(\ell/\chi') \ . \tag{1.45}$$

For the above to hold, however, y must vary on length scales much smaller than the typical length scale of the weight function w. We will discuss below what this means specifically in our case.

By direct comparison of equations (1.42) and (1.44) we can see what the weight function of convergence in Limber's equation is; by setting y to be the density fluctuation, δ_m, we get

that:

$$W_\kappa(\chi|z') = \frac{3}{2}\frac{H_0^2}{c^2}\Omega_{m,0}(1+z)\frac{\chi(\chi'-\chi)}{\chi'} = \rho_{m,0}\Sigma_c^{-1}(z,z') \tag{1.46}$$

where, $1+z = 1/a$, and this is what we call the *weight function* or *lensing efficiency*. In addition, we know that the power spectrum of $y = \delta_m$ is the matter power spectrum; thus:

$$P_y = P_{\delta_m} \equiv P_m(k,z) \propto \sigma_8^2 D^2(z)\mathcal{R}(k)$$

with D being the growth factor, $\mathcal{R}$ being the primordial power spectrum and

$$\sigma_8^2 \equiv \sigma^2(R = 8\ \mathrm{Mpc}/h) \tag{1.47}$$

if the variance of the power spectrum (see equation (1.26)) smoothed over a scale of 8 Mpc/h which we use to set the normalization of the dark matter's power spectrum. Then, Limber's approximation, i.e. equation (1.45), gives us the convergence power spectrum:

$$\begin{aligned} C_\kappa(\ell) &= \int_0^{\chi_S} d\chi\ \frac{W_\kappa^2(\chi|z_S)}{\chi^2} P_m(k = (\ell+1/2)/\chi, z) \\ &= \frac{9}{4}\frac{H_0^4}{c^4}\Omega_{m,0}^2 \int_0^{\chi_S} d\chi\ \frac{(1+z)^2}{\chi^2}\left[\frac{\chi(\chi_S-\chi)}{\chi_S}\right]^2 P_m(k = (\ell+1/2)/\chi, z)\ . \end{aligned} \tag{1.48}$$

Following the same procedure as above we can now use the *galaxy-matter cross power spectrum*, namely P_{gm}, in order to model the galaxy-galaxy lensing signal. For this purpose, we calculate projected the cross correlation between the dark matter density field and galaxy

positions and use it, under the Limber approximation, to obtain:

$$\begin{aligned} C_{gm}(\ell|z,z') &= \frac{W_\kappa(\chi|z')}{\chi^2} P_{gm}(k=(\ell+1/2)/\chi,z) \\ &= \frac{3H_0^2\Omega_{m,0}}{2c^2}\frac{1+z}{\chi^2}\left[\frac{\chi(\chi'-\chi)}{\chi'}\right] P_{gm}(k=(\ell+1/2)/\chi,z) \\ &= \frac{4\pi G\rho_{m,0}}{c^2}\frac{1+z}{\chi^2}\left[\frac{\chi(\chi'-\chi)}{\chi'}\right] P_{gm}(k=(\ell+1/2)/\chi,z) \\ &= \rho_{m,0}\Sigma_c^{-1}(z,z') P_{gm}(k=(\ell+1/2)/\chi,z)\ . \end{aligned} \tag{1.49}$$

However, in applications like the ones we are interested in, we deal with a large number of galaxies that are distributed in space in a particular way, and the features of their distribution is important in precision cosmology. Therefore, if we define the normalized distribution of lens and source galaxies in redshift as $n_L(z)$ and $n_S(z)$, respectively, we can average over these PDF's to obtain the mean shear within a redshift range. It is convenient in this case to calculate the average of equation (1.46) over the PDF of source galaxies,

$$\begin{aligned} W_\kappa(\chi) &= \frac{4\pi G\rho_{m,0}}{c^2}(1+z)\int_\chi^\infty d\chi'\ n_S(\chi(z'))\chi(z)\frac{\chi(z')-\chi(z)}{\chi(z')} \\ &= \rho_{m,0}\int_z^\infty dz'\ n(z')\Sigma_c^{-1}(z,z'), \end{aligned} \tag{1.50}$$

and to define the weighting function of galaxies as

$$W_g(\chi) = n_L(\chi(z))\frac{dz}{d\chi}\ , \tag{1.51}$$

which is also called the *radial galaxy selection function*. With the help of the above, we then simply write down the galaxy-galaxy lensing Limber power spectrum as:

$$C_{gm}(\ell) = \int d\chi\ \frac{W_\kappa(\chi)W_g(\chi)}{\chi^2} P_{gm}(k=(\ell+1/2)/\chi,z)\ . \tag{1.52}$$

Therefore, the average galaxy-galaxy lensing signal in real space, is computed by performing a Hankel transform to the above, defined as

$$\gamma_t(\theta) = \int \frac{\ell d\ell}{2\pi} J_2(\ell\theta) C_{gm}(\ell) \ , \tag{1.53}$$

where J_2 is the Bessel function of the first kind of the second order. The above equations, can then be combined with an HOD prescription, as discussed in Section 1.3, to calculate individual components of the galaxy-galaxy lensing signal; e.g. from 1-halo centrals or satellites, etc. We discuss in detail these expressions in Chapters 3 and 4.

Finally, discuss how we model the statistics of galaxy clustering, following a similar process as for galaxy-galaxy lensing above. When we generally talk about *galaxy clustering* there is a number of different ways we can think of it: clustering of galaxies in real three-space, the projected clustering on the two-sphere, or even redshift-space clustering. Then, there are various ways of modeling this statistic: one might be interested in the angular clustering or the three-dimensional clustering as a function of radius. In what follows, however, we will be interested in how galaxies cluster in real space and the corresponding Fourier-space functions which we will be projecting on the two-dimensional sky as a function of angular scale, θ. This we will be denoting by $w(\theta)$.

We start by constructing the *galaxy position auto-correlation*, P_{gg}, which is the equivalent of $P_{gm}(k, z)$ for galaxy clustering. This represents the three-dimensional power spectrum, i.e. the Fourier transform of the real-space ξ_{gg} correlation function of the clustering statistics. From that, the next step is to project on the two-dimensional sky and assume use the Limber approximation to obtain:

$$C_{gg}(\ell) = \int d\chi \frac{W_g^2(\chi)}{\chi^2} P_{gg}(k = (\ell + 1/2)/\chi, z) \tag{1.54}$$

where, since P_{gg} is the galaxy position auto-correlation power spectrum we have two factors

of W_g, the selection function of the galaxies of interest in a redshift range. Then, the observable of interest can be computed by a Hankel transform as:

$$w(\theta) = \int \frac{\ell d\ell}{2\pi} J_0(\ell\theta) C_{gg}(\ell) \ , \tag{1.55}$$

similarly to γ_t but using the Bessel function of different order. For a detailed discussion on how we choose to model $w(\theta)$ under the HOD framework see Chapter 4

In Chapter 2 we are going to discuss more what the purpose of using the above statistics is in Cosmology. Later on, in Chapters 3 and 4, we will apply the model to real-world data and dive deep into the specific details of the modeling.

CHAPTER 2

CONSTRAINING COSMOLOGY WITH GALAXY SURVEYS

2.1 An overview of galaxy surveys

A *galaxy survey* here refers to astronomical observations designed to map the distribution of galaxies in the sky in order to probe the large-scale structure of the Universe. Since galaxies are biased tracers of the dark matter density field, we can use data from galaxy surveys to make inferences about the dark matter's distribution and other fundamental properties. The way this is usually done is by extracting the observables of interest (such as weak lensing shear and galaxy counts) from the raw data, measuring statistics of the observables and then compare with cosmological models in order to constrain the parameters in the model and to prove or disprove the validity of the theory. In this section we will discuss in general terms the types of galaxy surveys and their purpose, go over a brief historical overview of the efforts of mapping the sky and explain why galaxy surveys are so important in our quest to learn more about the Cosmos.

When we carry out galaxy surveys we are aiming to map as large of an area of the sky, and to go as deep (i.e. as high in redshift) as we possibly can within some range of the electromagnetic spectrum. For this reason we need big telescopes with high sensitivity and good ways of estimating distances to the observed objects. The latter is an especially important and tricky point: while the angular positions of the objects we see is easy to measure on the 2-dimensional (2D) sky, to construct the 3-dimensional (3D) signal we need to estimate redshifts to the objects. The challenge is amplified by the sheer number of objects that we observe in a galaxy survey, which can be of the order of hundreds of millions of galaxies or even more. Why this is an issue will become more clear in what follows below.

There are various ways of extracting the redshift information for an astronomical object, but here we will only discuss a few that are relevant to what follows. To begin with, however,

we need to make the distinction between a *spectroscopic* and a *photometric* survey. In spectroscopic surveys, we measure the spectra of objects, say of a galaxy, that is, we measure the energy emitted by an object as a function of wavelength. This can precisely characterize the nature of the object (e.g. separate between galaxies and stars, etc) and it provides one of the more accurate ways of estimating redshifts via methods such as template fitting where we fit templates generated based on theoretical models to observed galaxy spectra in order to constrain the galaxy's redshift. However, this technique is time consuming and cannot easily be performed for a very large number of objects in a relatively short amount of time. This is where photometric surveys shine, i.e. in scanning the sky for objects in a more time-efficient manner.

As its name suggests, in this case we perform photometry on an object, which is the process of measuring the integrated flux of light emitted by it within a relatively wide band of the electromagnetic spectrum. The shortcoming of this technique is that it can only give us a rough estimate of an object's redshift via various ways that we will briefly mention below. But before we move on to that, we should emphasize here that in this work we will be working with data from a photometric survey, introduced in Section 2.2 and analyses of its data descussed in Sections 3, 4 and 5.

We will omit discussing measuring redshifts from spectroscopy here, as it will not be directly relevant to what follows, but we will rather provide a brief introduction to how some techniques of photometric redshift estimation, also referred to as *photo-z*, work (for a review on the topic see also Newman & Gruen, 2022). In addition, we are not interested in estimating the redshift of individual objects, known as *performance*, but rather in obtaining an stimate of the distribution of redshifts, $n(z)$, of an ensemble of galaxies, a process referred to as *characterization*, for as many high-precision measurements as possible. From the pool of methodologies to estimate photo-z's, some that are employed by modern cosmological surveys and are of interest to us include the following.

- **Template-based methods:** This family of methods rely on a *priori* knowledge of models based on the range of galaxy *Spectral Energy Distributions* (SED's) that exist. These algorithms usually construct the PDF of the redshift of a galaxy by applying the *Bayes theorem*. Without going into too much detail, the Bayesian Photo-z (BPZ)[1] method (see also Benítez, 2000, and figure 2.1), an application of which we will see in Chapter 4, commonly performs χ^2-based likelihood analyses to fit already existing templates of galaxy SED's to the photometry of the galaxies whose photo-z we want to estimate. We note here that these methods belong to the even broader family of Bayesian hierarchical methods we briefly discuss below.

- **Clustering-based redshift estimation:** Whenever an object in a photometric survey is near another object for which we have spectroscopic measurements, and thus a good redshift estimate, it is possible that we match the objects and in this way compute the photo-z of the object of interest. The redshift estimation from this cross-correlation between a photometric and a spectrocsopic galaxy sample is known as *clustering redshift*, it is introduced in Newman (2008), and it is a method that is employed by modern galaxy surveys.

- **Machine Learning methods:** This method relies again on our pre-existing knowledge of the redshift and other features (such as magnitudes) of a sample of galaxies, which is usually smaller than the full sample of galaxies for which we want to estimate photo-z's. Using the former as the *training data* we teach a Machine Learning (ML) model how to find correlations between redshift and other galaxy properties. We then apply the model to our full set of galaxies whose photo-z's we are trying to compute. We will see an application of that in Chapter 4.

- **Bayesian Hierarchical methods:** It might be already clear from the above, but

1. https://www.stsci.edu/ dcoe/BPZ/intro.html

we should note here that the photo-z methods are necessarily hierarchical, speaking in a statistical way, given that there are at least two sets of parameters relevant to the redshift inference: one that is a property of the single object whose photo-z we are trying to estimate and one that is describing the properties of the ensemble of galaxies. The Bayesian Hierarchical techniques exploit this fact and take it to its limits by constructing posterior distribution estimates for both the redshift of the individual object and the properties of the underlying ensemble of galaxies via Bayesian parameter estimation. However, the Bayesian Hierarchical methods are still in their early stages of development, but they look promising in addressing many challenges in photo-z estimation, as discussed in Newman & Gruen (2022) for example. One example of these models is the *Forward-modeling methods* (Herbel et al., 2017; Tortorelli et al., 2021) where the model parameters are constrained via *Markov Chain Monte-Carlo* (MCMC) methods to resemble the observed distributions of galaxy properties, including the measured redshifts, of spectroscopic objects.

Having the redshift information of a galaxy can tell us a lot about its intrinsic properties and allows us to perform studies where we divide galaxies into tomographic bins in redshift. This reveals trends in how the properties of the Large-scale structure of the Universe evolve over time. We note here that, all the above photo-z methods, albeit useful in conducting cosmological analyses, have their shortcomings. Especially in the era of precision cosmology, the demands for accurate redshift estimations are high. Therefore, new and better methods of estimating redshifts to objects, or improvements in the aforementioned ones, are always under development. This also implies that, given the photo-z uncertainties and how they can impact our results, we need to account for it in our work, as part of our model. We will see examples of how this is done in Chapters 3 and 4 where we marginalize over those uncertainties in our MCMC chains.

We will now close this section on a historical note (see also Coil, 2013; Okamura, 2019;

Wang & Zhao, 2020) by talking about the evolution of galaxy surveys in brief. The study of whether galaxies are distributed uniformly in space can be traced back to Edwin Hubble (Hubble, 1926), who used his catalog of 400 "extragalactic nebulae" to test the homogeneity of the Universe and found it to be uniform on large scales. In 1932, the larger Shapley-Ames catalog was published (Shapley & Ames, 1932), in which it was found that the galaxies projected onto the plane of the sky are unevenly distributed with roughly a factor of 2 difference in the numbers of galaxies in the northern and southern galactic hemispheres. Using this larger sample, Hubble (Hubble, 1934) found that on angular scales less than $\sim 10^{\circ}$ there is an excess in the number counts of galaxies above what would be expected for a random Poisson distribution, though the sample follows a Gaussian distribution on larger scales. Therefore, while the Universe appears to be homogeneous on the largest scales, on smaller scales it is clumpy.

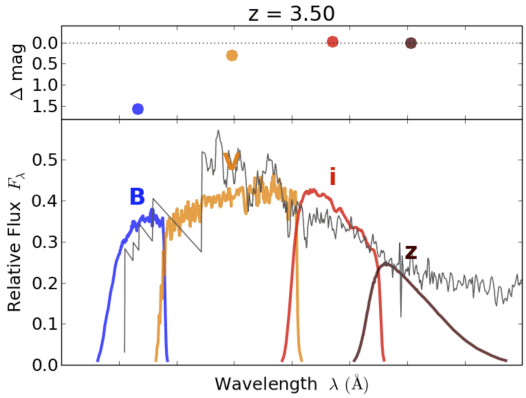


Figure 2.1: Illustration of fitting a template to photometry in the BPZ method of redshift estimation. The algorithm tries to minimize the χ^2 (i.e. find the smallest set of Δmag points in the upper panel for all bands) of the fit of the model (black line in the bottom panel), i.e. the spectrum redshifted by z (shown in the title), to the photometry in each band (colorful lines). *Image credit: Illustration from the BPZ's website.*

The study of the large-scale structure of the Universe significantly moved forward with the Lick galaxy catalog produced by Shane & Wirtanen (1967), which contained roughly one million galaxies. Seldner et al. (1977) published maps of the counts of galaxies in angular cells on the sky, which showed in much greater detail that the projected distribution of galaxies on the sky is not uniform, but it rather contains walls and filaments, clusters, and large voids. The statistical spatial distribution of galaxies from this catalog and that of Zwicky et al. (1961) was analyzed by Jim

Peebles and collaborators in a series of papers (e.g., Peebles & Groth, 1975) who showed that the angular 2-point correlation function roughly follows a power law distribution over angular scales of $\sim 0.1^\circ - 5^\circ$.

These results played a key role in motivating the first large-scale redshift surveys which obtained optical spectra of individual galaxies in order to measure the redshifts and the spatial distribution of large galaxy samples. Two galaxy surveys were the KOS survey (Kirshner et al., 1978) and the original CfA survey (Center for Astrophysics; Davis et al. (1982)). The KOS survey measured redshifts for 164 galaxies brighter than magnitude 15 in eight separate fields on the sky, covering a total of 15 deg^2. Part of the motivation for the survey was to study the three-dimensional spatial distribution of galaxies, and it was found that the galaxies are strongly clustered in velocity space. The original CfA survey, completed in 1982, contained redshifts for $2,400$ galaxies brighter than magnitude 14.5 across the north and south galactic poles, covering a total of 2.7 steradians. The goals of the survey were cosmology and quantifying the clustering of galaxies in three dimensions. From this survey, one could identify galaxy clusters, voids, and a "filamentary connected structure" between groups of galaxies. This paper also performed a comparison of the so-called "complex topology" of the large scale structure seen in the galaxy distribution with that seen in N-body dark matter simulations, paving the way for future studies of theoretical models of structure formation. The second CfA redshift survey, which ran from 1985 to 1995, contained spectra for $\sim 5,800$ galaxies and revealed the existence of the so-called "Great Wall", a supercluster of galaxies that extends over 170 Mpc/h. Large voids were also found, with a density 20% of the mean density.

Redshift surveys have rapidly progressed with the development of multi-object spectrographs, which allow simultaneous observations of hundreds of galaxies, and larger telescopes, which allow deeper surveys of both lower-luminosity nearby galaxies and more distant luminous galaxies. Some of the largest redshift surveys of galaxies are: the Las Campanas

redshift survey (Shectman et al., 1996), the Sloan Digital Sky Survey (SDSS, York et al. (2000)), the Two Degree Field Galaxy Redshift Survey (2dFGRS, Colless et al. (2001)), the Galaxy and Mass Assemply (GAMA, Baldry et al. (2010)) survey, etc.

Figure 2.2: The Dark Energy Survey (DES) logo.

Today we live in a very exciting era of data-driven cosmology – the amount of data is almost overwhelming, which necessitates the development of new techniques to analyze them, is many cases based on ML (and thus the high demand of such algorithms for cosmology), over a wide sky coverage, large redshift range and large density of observed galaxies. The largest galaxy surveys today that probe the Large-scale structure and dark energy are: the Dark Energy Survey (DES, Flaugher et al. (2015)), the Kilo-Degree Survey (KiDS, de Jong et al. (2015)) and the Hyper Suprime-CamSubaru Strategic Program (HSC-SSP, Aihara et al. (2017)) – what we refer to as *Stage-III galaxy surveys*. All of these surveys have provided exquisite measurements with which we have been able to obtain cosmological constraints of high significance. And the precision is only going to increase as we step into the *Stage-IV galaxy surveys*, such as the Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration (2016)) and the *Vera C. Rubin observatory* which will conduct the *Legacy Survey of Space and Time* (LSST) survey, as the number of observed objects is expected to skyrocket, which will require improved theoretical models as well as statistical techniques and great computational power to perform the analyses. For the remaining of this section we will introduce the Dark Energy Survey, before we delve deeper into the specifics of how one can use data from such a survey to constraint cosmology and astrophysics.

2.2 The Dark Energy Survey

In this section we give a brief overview of the Dark Energy Survey (DES, Flaugher et al., 2015), since we will be exclusively using data from this experiment in our main analyses in Chapters 3 and 4, as well as in Chapter 5.

The Dark Energy Survey (its logo shown in figure 2.2) is an international collaborative effort to survey the sky and map millions of galaxies, thousands of Supernovae, and to probe the large-scale structure of the Universe and reveal the nature of the dark energy. It is a photometric survey that covers about a quarter of the southern sky to a depth of $r > 24$ and images $\sim$ 300 million galaxies in 5 broadband filters ($grizY$), up to redshift $z \sim 1.4$ (Flaugher et al., 2015; DES Collaboration, 2016).

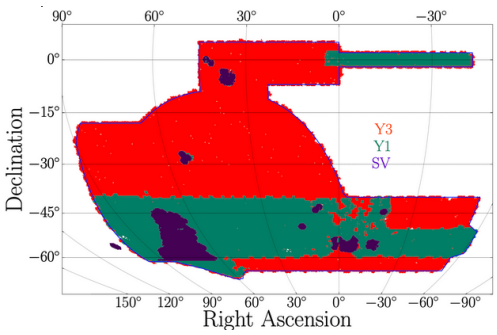


Figure 2.3: The DES footprint, showing the sky coverage split into the Science Verification (SV), Year 1 (Y1) and Year 3 (Y3) data sets.

The DES data was assembled from the imaging survey using the Blanco 4m telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile and observed in a range from $\sim$ 400 nm to $\sim$ 1060 nm in wavelength, with the Dark Energy Camera (DECam Flaugher et al., 2015). DES completed observations in January 2019, after 6 years of operations, with 10 overlapping dithered exposures at predefined positions in the sky in each filter. The primary goal of the DES is to study the origin of cosmic acceleration and the nature of dark energy, using a variety of cosmological probes enabled by this rich data set.

The Science Verification (SV) exposures covered an area of $\sim$ 250 deg^2 and it showed that the DES can provide a powerful data set with wich to constrain cosmology[2]. For

2. https://des.ncsa.illinois.edu/releases/sva1/docs

many of the Year 1 (Y1) analyses in DES (Abbott et al., 2018) the `Y1 GOLD` catalog was utilized, which is described in Drlica-Wagner et al. (2018). The DES Y1 data was publicly released in October 2018[3] and the release includes: the `Y1 GOLD` catalog that covers an area of ~ 1800 deg^2, the ancillary maps of the survey properties (Leistedt et al., 2016), the shear catalogs (Zuntz et al., 2018a), the photometric redshift catalogs (Hoyle et al., 2018), the Luminous Red Galaxy (LRG) `redMaGiC` catalogs (Rozo et al., 2016) used for the DES Y1 results, and the value-added catalogs (Sevilla-Noarbe et al., 2018; Tarsitano et al., 2018).

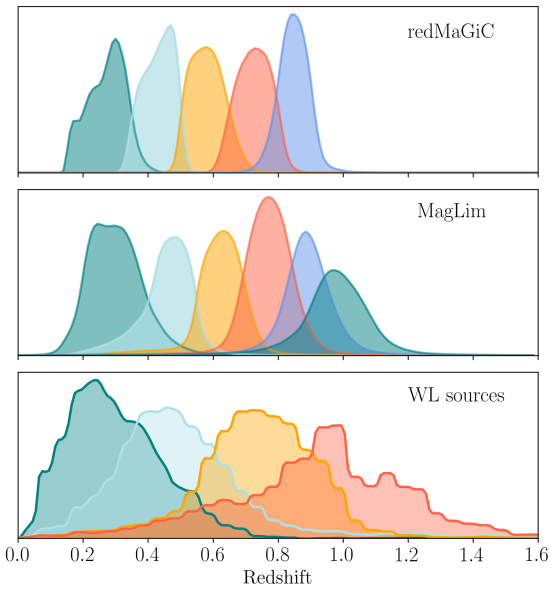


Figure 2.4: The DES Y3 redshift distributions of the lens galaxy samples (`redMaGiC`, in the top panel, and `MagLim`, in the middle panel) and source galaxies, in the bottom panel. The different colors represent different tomographic bins (see also, Prat et al., 2022).

The most resent public data release of the DES, at the time of the writing of this thesis in 2023, consists of observations from the first three years (Y3) of observations. The data used for the Y3 analyses (Abbott et al., 2022) is the `Y3 GOLD` catalog and is described in Sevilla-Noarbe et al. (2021). The area covered in Y3 increased to ~ 4132 deg^2 and two survey modes were chosen: the *Wide Survey* which is optimized for the weak lensing, galaxy clustering, and galaxy cluster cosmological probes, and the *Supernova Survey* which enables the discovery of thousands of Supernovae. In Chapters 3, 4 and 5 we will focus on some applications that use data from the Wide Survey of Y3 of the DES. More specifically, in Chapters 3 and 5 we will be presenting results that utilize the two lens galaxies samples named `MagLim` (Porredon et al., 2021), which is optimized for cosmology, and the Y3 `redMaGiC` sam-

3. https://des.ncsa.illinois.edu/releases/y1a1

ple, which uses its own point estimate from the red-sequence template fitting, as described in Rozo et al. (2016). As our source galaxy sample we use the shape catalog described in Gatti et al. (2020) and Jarvis et al. (2020), which is based on the METACALIBRATION technique (Huff & Mandelbaum, 2017; Sheldon & Huff, 2017), which is able to accurately measure weak lensing shear using the available imaging data. The distributions in redshift of all these three galaxy samples are shown in figure 2.4. In Chapter 4 we are going to introduce a new lens galaxy sample, created from the DES `Y3 GOLD` catalog, which also provides the stellar-mass and color information of the galaxies.

2.3 Cosmological constraints from galaxy surveys

In this section we will go over some methodologies that are commonly used to extract information from galaxy-survey data with the goal of constraining the cosmological parameters. The general idea is the following:

1. Starting from the calibrated data, we produce redshift estimates of the galaxies, e.g. using methods that are mentioned in Section 2.1. Depending on what we want to achieve in the study, we can produce other property estimates, such as galaxy color and stellar mass, for the galaxies. An example where we also estimate the stellar mass and color for our lens sample in presented in Chapter 4.

2. We then choose the lens and source samples and split the galaxies into tomographic bins in redshift, and possibly in bins in other properties. We also construct the distributions of the galaxies per bin in redshift and the other properties of our choice.

3. After that, we measure the statistics – e.g., the 2-point statistics such as galaxy-galaxy lensing and clustering – over a range of scales in every bin (auto-correlations) or bin combination (cross-correlations).

4. We usually want to validate the measurements; hence, we perform various tests on the data and measurements in order to test for systematic effects.

5. On the theory part, after we construct the theoretical framework we want to work in, we produce the code to numerically calculate the theory predictions of the observables of interest (e.g. the two-point functions), given an input set of parameters, redshift distributions, angular scales, etc.

6. We then build the numerical pipeline to produce the theoretical expectations, read in the measurements, and compare the two in a Bayesian likelihood analysis, to find the set of parameters that best describe the data. To do so, we usually need a good *prior* knowledge of the distribution of the parameters we vary.

7. As a result of running the MCMC chains to fit the model to the data, we obtain the parameter *posterior distributions*, which we usually report as a value, which can be the *Maximum Likelihood (ML)* point or the *Maximum a Posteriori (MAP)* point or something else, and the 1σ uncertainty on that value. It is also common to show the full posterior distributions in a triangle plot (with the 1D distribution of each parameter on the diagonal and the joint 2D distributions between two of the parameters below the diagonal), which can reveal important correlations – also called degeneracies – between parameters.

8. Depending on our goals, we then translate the above posterior constraints into constraints on derived quantities or parameters, i.e. on things that depend on the baseline free parameters of the model.

The above procedure is commonly utilized by collaborations such as the DES to analyze the data and constrain the cosmological parameters. Note here, however, that there are many more analysis choices that feed into the above general procedure. One of them, which is of great significance to this thesis, are the *scale cuts* on the data. By "scale cuts" we refer

to the decision we make on what scales to consider in a given analysis, or in other words, what scales to discard. The reasons we might choose to simply ignore scales can be many; for example, we might find by performing tests on the data that some measurements are of very poor quality, or maybe our model cannot be trusted to provide an accurate prediction in some regime. The ways that scale cuts are determined can also vary depending upon the collaboration and the specifics of the analysis. For example, the fiducial scale cuts on lensing and clustering in the DES cosmological analyses are more or less fully determined based on the applicability of the linear theory, i.e. scales below which the 1-halo terms are needed, or even scales in the quasi-linear regime, are removed (see e.g., Abbott et al., 2022; Prat et al., 2022; Sánchez et al., 2022a, and references therein). This has simply to do with the complications that come with trying to model the non-linear scales. More on this can also be found in Section 5.1 where we talk about how we can use "contaminated data sets" as a test to determine scale cuts, as presented in Prat et al. (2022).

Before we conclude this section, it is worth introducing some useful terminology. In DES the main cosmological results are produced from fitting jointly three 2-point statistics: galaxy-galaxy lensing (γ_t), galaxy clustering (w) and cosmic shear ($\xi_\pm$), which is the so-called 3×2-point (or 3×2pt) analysis. Cosmic shear is the 2-point correlation between the shapes of the source galaxies, while γ_t and w have been defined in Chapter 1. In what follows, however, especially in Chapters 3 and 4, we will be working only with the first two, γ_t and w. The type of analyses that utilize one 2-point statistic are known as 1×2-point (or 1×2pt), while the 2×2-point (or 2×2pt) analyses only consider two 2-point functions. Also, from now on we will be referring to the scales considered by the DES in their 3×2pt analyses as *large scales*, of a few Mpc or above, while the scales that are excluded we will call *small scales*, which are of order or below a Mpc and can be as small as $\lesssim 100$ kpc. Our focus for the majority of this thesis will be on extending the modeling to also account for the small scales, using the halo model and the galaxy-halo connection.

2.4 Modeling the non-linear scales in Cosmology

There is no doubt that the linear scales contain a wealth of information that is valuable in constraining the cosmological parameters (see e.g., Abbott et al., 2018, 2022, as an example). It is also true, however, that the vast majority of the cosmological analyses get rid of most of the part of the data vector via scale cuts. The small scales that are excluded are powerful in terms of constraining power as they contain a lot of additional information, but can also be a source of systematic errors if modeled incorrectly. Beyond the "standard cosmology", modeling small scales provides a direct way of combining cosmology on the largest scales with the astrophysical processes that take place on smaller scales, thus allowing us to form a complete picture of how the Universe operates. In the era of precision cosmology, the measurements are in many cases so precise that the error bars they come with are tiny, especially on small scales. This, in addition to the complex nature of modeling the non-linear regime, makes it clear why the 1-halo scales are usually ignored. However, in order to make meaningful progress in cosmology and to take full advantage of the data sets we need to build a framework that can model the full range of scales.

In what follows, we will discuss how the small-scale modeling in DES can lead to constraints on the galaxy-halo connection. Specifically, in Chapter 3 we will present constraints on the average halo mass, galaxy bias and satellite fraction of the DES Y3 `redMaGiC` and `MagLim` samples from galaxy-galaxy lensing alone. In Chapter 4 we will expand our framework to include galaxy clustering on the small scales in order to perform a 2×2pt analysis on the DES Y3 data set with the goal to constrain the stellar-to-halo mass relation of the lenses. In that work, we will in addition describe a new lens galaxy sample that we constructed from the `Y3 GOLD` catalog which includes the stellar-mass and color information of the galaxies. Building the framework and pipeline to perform this 2×2pt analysis on small and large scales simultaneously is the necessary first step in the direction of running a cosmological analysis (e.g. the equivalent of the 3×2pt analysis in DES, Abbott et al., 2022) in which

we utilize the full range of scales, as we will briefly discuss in Chapter 7.

To conclude this chapter, we will summarize the reasons why the galaxy-halo connection is important in cosmology (for more see also the review by Wechsler & Tinker, 2018):

- **It helps us study the physics of galaxy formation and evolution:** As we discussed in Chapter 1, the formation and evolution of galaxies is inevitably connected to dark matter. Thus, through modeling the galaxy-halo connection and using real data to constrain the parameters in our models we can say a lot about which models are valid and we can learn more about the relation between galaxies and their host dark matter halos. For example, we are usually interested in understanding which halo properties are the most important in setting the properties of the galaxies they host.

- **It can help with cosmological parameter inference:** As we have mentioned in Section 2.1, future galaxy surveys will provide an unprecedented amount and quality of data which we will want to utilize. This will lead to very high-precision measurements, especially in galaxy clustering, that can be pushed to smaller scales, deep into the non-linear regime. On those scales, high-order perturbation theory breaks down and the galaxy bias becomes very complex, thus necessitating the use of a galaxy-halo connection model that possesses the required flexibility to fit the data in, e.g., a 3×2pt analysis that is sensitive to cosmological information. However, we should note here that, beyond the 3×2pt studies, one can also think of other probes, such as *redshift-space distortions* and *baryonic processes*, that are very sensitive to cosmology as well.

- **It probes the nature and properties of dark matter:** We have talked about the success of the ΛCDM model in Chapter 1. These successes mostly come from tests on the large scales, however, and the concordance cosmological model is to a lesser degree tested on the small scales. In the non-linear regime, alternative models of the dark matter can modify the power spectrum or the density profile of dark matter, depending

on the dark matter properties. Therefore, by modeling the galaxy-halo connection we can perform various tests on small scales to study the fundamental properties of the dark matter (see also Chapter 7).

- **It informs numerical simulations:** Due to computational limitations, it is today impossible to run large-scale simulations that evolve the dark matter field and at the same time account for baryonic processes. Therefore, we usually end up running N-body dark matter-only simulations. If, however, we have a good understanding of the galaxy-halo connection, after we identify dark matter halos in the simulation box via the algorithm of our choice, we can populate the halos with galaxies, based on their properties. This can significantly help with performing simulation-based tests in a timely manner.

In the next two chapters we will present applications of the galaxy-halo connection to DES Y3 data, as we mentioned above. These studies are a big step forward in understanding the galaxy-halo connection. In the next decades, we expect that our understanding of this relation between galaxies and halos will be crucial in fully describing the data from galaxy surveys in order to constrain cosmology and astrophysical processes. At the same time, it is clear that for this to be done, the precision of models for the galaxy–halo connection will need to keep up with the increasing quality of the data.

The rest of this thesis is structured as follows. Chapter 3 introduces the main HOD framework that we will be using throughout this thesis to model small scales in cosmology. That chapter is based on Zacharegkas et al. (2022) and models galaxy-galaxy lensing down to small scales in DES Y3. Next, Chapter 4 expands upon that HOD work to also incorporate galaxy clustering in the model, so that a 2×2pt analysis in DES using small and large scales can be achieved with the goal of constraining the galaxy-halo connection of the lens sample. In addition, the enhanced model in Chapter 4 accounts for more galaxy properties, specifically stellar mass. For this reason, we also introduce a new lens galaxy

sample from DES Y3 that we use in our analysis. This work will be published in the near future (Zacharegkas et al. in prep). In Chapter 5 we discuss additional work within DES Y3 that has utilized the HOD framework for Chapter 3 to perform various tests. This chapter emphasizes further how important modeling the small scales in cosmology can be, even as a means to test our models and inform our modeling choices. Chapter 6 is based on Prat et al. (2023) and presents an alternative way of dealing with small scales compared to the previous chapters. Instead of modeling the non-linear regime, as we do in Chapters 3 and 4, in that chapter we discuss ways of removing that regime and the contamination from it on the linear scales that survive the scale cuts. Finally, in Chapter 7 we conclude by summarizing, as well as mentioning various interesting future directions we can take in cosmology using small scales.

CHAPTER 3

DARK ENERGY SURVEY YEAR 3 RESULTS: GALAXY-HALO CONNECTION FROM GALAXY-GALAXY LENSING

In this chapter we describe in detail a galaxy-halo connection model that we developed to analyze data from DES Y3. Our framework allows us to model the non-linear scales that are ignored in the key cosmology analyses in DES. By fitting this model to the data we are able to constrain the average halo mass, satellite fraction and galaxy bias of the two lens galaxy samples utilized for cosmology by the DES collaboration. What follows is taken from Zacharegkas et al. (2022).

3.1 Introduction

Understanding the connection between galaxies and dark matter, i.e. how the galaxy properties relate to the properties of their dark matter halo hosts, is essential in forming a comprehensive interpretation of the observed Universe. Cosmological analyses of large-scale Structure (LSS) in modern galaxy surveys have reached a point where ignoring the details of this connection (McDonald & Roy, 2009; Baldauf et al., 2012) can lead to significant biases in the inferred cosmological constraints (Krause et al., 2017). To avoid this problem, typically we remove data points on the smallest scales until the remaining data is in the linear to quasilinear regime, and a simple prescription of the galaxy-halo connection (e.g. linear galaxy bias) is sufficient (such as DES Collaboration, 2021). Alternatively, one can invoke more complicated galaxy bias models on small scales (such as the model of Heymans et al., 2021) and marginalise over the model parameters. For either approach, a data-driven model of the galaxy-halo connection on scales below a few Mpc could allow us to significantly improve the cosmological constraints achievable by a given dataset. It should be stressed, however, that galaxy bias has inherently non-linear characteristics (as discussed, for exam-

ple, in Dvornik et al., 2018), and should therefore be treated accordingly. Thus, accurate galaxy-halo connection models provide a wealth of crucial information when modeling galaxy bias. On the other hand, understanding the connection between different galaxy samples and their host halos also has implications for galaxy evolution (see Wechsler & Tinker, 2018, for a review of studies for galaxy-halo connection).

A powerful probe of the galaxy-halo connection is *galaxy-galaxy lensing*. Galaxy-galaxy lensing refers to the measurement of the cross-correlation between the positions of foreground galaxies and shapes of background galaxies. Due to gravitational lensing, the images of background galaxies appear distorted due to the deflection of light as it passes by foreground galaxies and the dark matter halos they are in. As a result, this measurement effectively maps the average mass profile of the dark matter halos hosting the foreground galaxy sample. This is one of the most direct ways to connect the observable properties of a galaxy (brightness, color, size) to its surrounding invisible dark matter distribution (Tyson et al., 1984; Hoekstra et al., 2004; Mandelbaum et al., 2005; Seljak et al., 2005). A common approach to modeling this measurement is to invoke the *Halo Model* (Seljak, 2000; Cooray & Sheth, 2002) and the *Halo Occupation Distribution* (HOD) framework (Zheng et al., 2007; Zehavi et al., 2011). In this framework, we consider dark matter halos to be distinct entities with a large luminous *central galaxy* in their centers and smaller, less luminous *satellite galaxies* distributed within the halo, which are also surrounded by their own sub-halos (Kravtsov et al., 2004). The particular way that central and satellite galaxies occupy the dark matter halo is parametrised by a small number of HOD parameters, while all the dark matter halos contribute separately to the total galaxy-galaxy lensing signal according to the Halo Model. In this paper, we will invoke this HOD framework to model a new set of galaxy-galaxy lensing measurements using the Dark Energy Survey (DES) Year 3 (Y3) dataset.

Several previous studies have used galaxy-galaxy lensing to constrain the galaxy-halo connections for particular samples of galaxies. Mandelbaum et al. (2006a) performed an

analysis with the MAIN spectroscopic sample from the Sloan Digital Sky Survey (SDSS) DR4, characterising the HOD parameters for galaxies split in stellar mass, luminosity, morphology, colors and environment. The study was followed up by Zu & Mandelbaum (2015) using SDSS DR7 with a more sophisticated HOD model. The fact that all lens galaxies used in these studies have measured spectra allowed for good determination of the stellar mass and other galaxy properties. More recently, rapid development of large galaxy imaging surveys provide much more powerful weak lensing datasets to perform similar analyses. Gillis et al. (2013); Velander et al. (2013); Hudson et al. (2014) used measurements from the Canada-France-Hawaii Telescope Lensing Survey (CFHTLenS, Heymans et al., 2012; Erben et al., 2013), while Sifón et al. (2015); Viola et al. (2015); van Uitert et al. (2016) used data from the Kilo Degree Survey (KiDS, de Jong et al., 2013; Kuijken et al., 2015) to study the galaxy-halo connection for a range of different galaxy samples. Noticeably, these studies extend to higher redshifts as well as lower mass (including Ultra-Diffused Galaxies at low redshift). Furthermore, Bilicki et al. (2021) used photometry from KiDS, exploiting some overlap with Galaxy And Mass Assembly (GAMA, Driver et al., 2011) spectroscopy, to derive accurate galaxy-galaxy lensing measurements, split in red and blue bright galaxies, to constrain the stellar-to-halo mass relation by fitting the data with a halo model. All together these studies provide us with pieces of information to constrain models of galaxy formation. In parallel, Clampitt et al. (2017) derived constraints on the halo mass of a luminous red galaxies sample, the red-sequence Matched-filter Galaxy Catalog (redMaGiC) galaxies (Rykoff et al., 2014), using DES Science Verification data. The redMaGiC sample is particularly interesting as it is used heavily in many cosmological studies of LSS due to its excellent photometric redshift precision. For that reason, redMaGiC is one of the two samples we study in this work. From the studies above, it becomes evident that the basic HOD framework is capable of successfully describing the halo occupation statistics for a wide variety of galaxy samples, as long as it is modified accordingly to account for the specific

features of the dataset at hand.

The Clampitt et al. (2017) study was later combined with galaxy clustering to constrain cosmological models in Kwan et al. (2016), illustrating how understanding the small-scale galaxy-halo connection (and effectively marginalizing over them) could improve the cosmological constraints. Similar studies include Mandelbaum et al. (2013); Cacciato et al. (2013); Park et al. (2016); Krause & Eifler (2017); Singh et al. (2020). In particular, Park et al. (2016) demonstrated that to obtain robust constraints from combining large and small scale information, it is necessary to consistently model the full range of scales, and to have good priors on the HOD parameters due to degeneracies between HOD and cosmological parameters. When including the small-scale modeling from HOD in a cosmology analysis using galaxy clustering and weak lensing, Krause & Eifler (2017) showed that the statistical constraints on the dark energy equation of state w improves by up to a factor of three compared to standard analyses using only large-scale information. We leave for future work the exploration of gain in cosmological constraints including our HOD modeling in the DES Y3 cosmology analysis.

Many studies (e.g. Leauthaud et al., 2017; Lange et al., 2019; Singh et al., 2020; Wibking et al., 2019; Yuan et al., 2020; Lange et al., 2021) have shown that fitting galaxy clustering measurements with small-scale galaxy-halo connection models, at fixed cosmology, provides precise predictions of the lensing amplitude which is higher than the measured signal. This is the so-called "lensing is low" problem, which becomes especially evident when small scales are considered in the analysis. Figuring out whether this discrepancy can be explained by new physics, cosmology or by reconsidering our galaxy formation models is an open question. A better understanding of the galaxy-halo connection can play a crucial role in solving this mystery. For example, Zu (2020) found that the "lensing is low" tension can be resolved on small scales; however, the satellite fraction has to be very high, which is not in agreement with observations (e.g. Reid et al., 2014; Guo et al., 2014; Saito et al., 2016).

In this chapter we make use of data from Y3 of DES to study the galaxy-halo connection of two galaxy samples: REDMAGIC and an alternative magnitude-limited galaxy sample defined in Porredon et al. (2021). These two samples are used in the DES Y3 cosmological analysis combining galaxy clustering, galaxy-galaxy lensing and cosmic shear (commonly referred to as the 3×2pt analysis as it combines three two-point functions, DES Collaboration, 2021). We measure the galaxy-galaxy lensing signal to well within the 1-halo regime, demonstrating the extremely high signal-to-noise coming from the powerful, high-quality dataset. We model the measurements by combining the Halo Model and the HOD framework, fixing the background cosmology to be consistent with the DES Y3 cosmology analysis. This work presents one of the most powerful datasets for studying the galaxy-halo connection in a photometric survey and includes two main advances compared to previous work of similar nature: First, we include a number of model components that were previously mostly ignored in studies of the galaxy-halo connection via galaxy-galaxy lensing. Second, we borrow heavily from the tools used in cosmological analyses and carry out a set of rigorous tests for systematic effects in the data and modeling, making our results very robust. Both of these advances were driven by the supreme data quality – as the statistical uncertainties shrink, previously subdominant systematic effects in both the measurements and the modeling become important.

With our analysis, we place constraints on the HOD parameters, and derive the average halo mass, galaxy bias and satellite fraction of these samples. Our analysis provides complementary information from the small-scales to the large-scale cosmological analysis in Prat et al. (2022) and informs future cosmology analyses using these two galaxy samples. As shown in Berlind & Weinberg (2002); Zheng et al. (2002); Abazajian et al. (2005), combining HOD with cosmological parameter inference can greatly improve the cosmological constraints. Our results can also be incorporated into future simulations that include similar galaxy samples.

The structure of the chapter is as follows. In Section 3.2 we describe the baseline formalism for the HOD and Halo Model framework used in this paper. In Section 3.3 we detail the different components that contribute to the galaxy-galaxy lensing signal that we model. In Section 3.4 we describe the data products used in this paper. In Section 3.5 we describe the measurement pipeline, covariance estimation and the series of diagnostics tests performed on the data. In Section 3.6 we describe the model fitting procedure and the model parameters that we vary. We also describe how we determine the goodness-of-fit and quote our final constraints. In Section 3.7 we show the final results of our analysis. We conclude in Section 3.8 and discuss some of the implications of our results.

3.2 Two theoretical pillars

In this section we describe the two fundamental elements in our modeling framework: the halo occupation distribution model and the halo model. As we discuss later, the combination of the two allows us to predict the observed galaxy-galaxy lensing signal to very small scales given a certain galaxy-halo connection.

3.2.1 Halo Occupation Distribution

The halo occupation distribution (HOD) formalism describes the occupation of dark matter halos by galaxies. There are two types of galaxies that can occupy the halo (Kravtsov et al., 2004): central and satellite galaxies. A central galaxy is the large, luminous galaxy which resides at the center of the halo. The HOD model does not allow for more than one central galaxy to exist inside the halo. On the other hand, the HOD allows for many satellite galaxies to exist in a halo. The higher the mass of the halo the more satellites are expected to exist around the central. Satellite galaxies are smaller and less luminous than the central. They orbit around the center of the halo and give rise to the non-central part of the galaxy-galaxy lensing signal, as we discuss in more detail later. In what follows, we define the HOD of

a galaxy sample which has a minimum luminosity threshold, similarly to Clampitt et al. (2017).

The central galaxy is assumed to be exactly at the center of the halo, i.e. our model does not account for effects that might come from mis-centering of the central galaxy in its dark matter halo. The number of centrals in our HOD framework is given by a log-normal mass-luminosity distribution (Zehavi et al., 2004; Zheng et al., 2005; Zehavi et al., 2011) and its expectation value is denoted by $\langle N_c(M_h)\rangle$. The scatter in the halo mass-galaxy luminosity relation is parametrised by $\sigma_{\log M}$. The mass scale at which the median galaxy luminosity corresponds to the threshold luminosity will be denoted as $M_{\rm min}$. A third parameter is the fraction of occupied halos, $f_{\rm cen}$, which is introduced specifically for REDMAGIC and accounts for the number of central galaxies that did not make it into our sample due to how the galaxies are selected. In more detail, due to the selection process of the REDMAGIC algorithm, for halos of a fixed mass, not all the central galaxies associated with those halos will be selected into the lens sample. More specifically, the REDMAGIC selection depends on the photometric-redshift errors, which could result in excluding some galaxies even though they are above the mass limit for observation [1]. For most galaxy samples that are selected via properties intrinsic to the sample (luminosity, stellar mass, etc.), however, $f_{\rm cen} = 1$ is a natural choice.

The expectation value for the number of centrals is the smooth step function

$$\langle N_c(M_h)\rangle = \frac{f_{\rm cen}}{2}\left[1 + {\rm erf}\left(\frac{\log M_h - \log M_{\rm min}}{\sigma_{\log M}}\right)\right] , \tag{3.1}$$

where erf is the *error function*. Note that $M_{\rm min}$ in this expression essentially sets the mass of the lens halos, which makes it a crucial parameter to constrain.

1. Our model is slightly different from Clampitt et al. (2017) in that $f_{\rm cen}$ is multiplied to both the centrals and the satellites. This choice results in better matching to the MICE simulations (see Appendix A.2) and therefore facilitates our testing. Since $f_{\rm cen}$ and M_1 are fully degenerate, this difference does not alter the physical form of the model, although we have adjusted the prior ranges on M_1 to account for that.

The expectation number of satellites is modeled using a power-law of index α and normalization mass-scale M_1, and is written as

$$\langle N_s(M_h)\rangle = \langle N_c(M_h)\rangle \left(\frac{M_h}{M_1}\right)^{\alpha} . \tag{3.2}$$

This relation implies a power-law behaviour for the satellite galaxies at high halo masses only, as $\langle N_s(M_h)\rangle$ is coupled to $\langle N_c(M_h)\rangle$. The total number of galaxies in a dark matter halo is $\langle N(M_h)\rangle = \langle N_c(M_h)\rangle + \langle N_s(M_h)\rangle$. Figure 3.1 shows the number of galaxies as a function of halo mass as calculated by the HOD model described above. We note that significant modifications on top of our model have been developed for samples specifically defined by stellar mass or colors (Singh et al., 2020). Also, simple variants of the HOD we have adopted have been used in the literature, but given the nature of the two samples we study in this work we do not expect these modifications to be necessary as we discuss in Section 3.7.3.

3.2.2 Halo model

In the framework of the current cosmological model the large-scale structure in the universe follows a hierarchy based on which smaller structures interact and merge to give rise to structure of larger scale. The abundance of dark matter halos is described by the halo mass function (HMF) which is denoted by dn/dM and is a function of the halo mass M_h at redshift z. In this work we utilise analytic fitting functions to model the HMF following Tinker et al. (2008).

The root-mean-square (rms) fluctuations of density inside a sphere that contains on average mass M_h at the initial time, $\sigma(M_h)$, is defined as the square root of the variance in

the dark matter correlation function and is written as

$$\sigma^2(M_h) \equiv \int \frac{k^2 dk}{2\pi^2} |\tilde{W}(kR)|^2 P(k) \ , \tag{3.3}$$

where $P(k)$ is the dark matter power spectrum and k denotes the wave number. In Equation (3.3) the variance in the initial density field has been smoothed out with a top-hat filter $W(R)$ over scales of $R = (3M_h/4\pi\rho_m)^{1/3}$, where ρ_m is the mean matter density of the universe, and $\tilde{W}$ is the Fourier transform of the top-hat filter. We use this expression to calculate σ_8, the rms density fluctuations in a sphere of radius $R = 8$ Mpc$/h$, which we use as the normalization of the matter power spectrum.

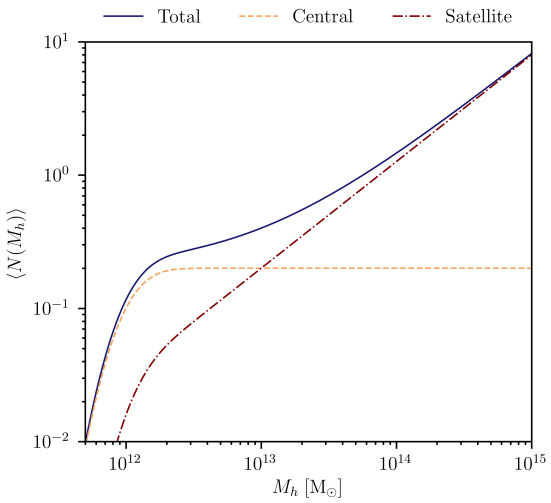


Figure 3.1: The HOD prediction for the expectation number of central (dashed), satellite (dash-dotted) and the total (solid) number of galaxies as a function of the mass of the dark matter halo inside of which they reside. The HOD parameters used to produce this plot are: $M_{\rm min} = 10^{12}\ M_\odot$, $M_1 = 10^{13}\ M_\odot$, $f_{\rm cen} = 0.2$, $\alpha = 0.8$, $\sigma_{\log M} = 0.25$.

For computing the distribution of the dark matter within a halo we assume a NFW density profile (Navarro et al., 1996) with characteristic density ρ_s and scale radius r_s. To calculate the concentration parameter of the dark matter distribution, $c_{\rm dm}(M_h, z)$, we follow Bhattacharya et al. (2013).

In order to calculate the linear matter power spectrum, $P_{\rm m}^{\rm lin}(k, z)$, we make use of accurate fitting functions from Eisenstein & Hu (1998) (EH98 hereafter). These fitting functions are accurate to $\sim 5\%$ and we use them instead of other numerical codes that calculate the power spectrum, such as CAMB (Lewis et al., 2000), to make our numerical code more efficient. We have performed the necessary numerical tests to show that this modeling choice does not affect the final

results. The linear power spectrum, however, poorly describes the power at the small, nonlinear scales. In our modeling we correct for this by using the nonlinear matter power spectrum, $P_{\rm m}^{\rm nl}(k,z)$, by adopting the HALOFIT approximation based on Takahashi et al. (2012) to modify the EH98 linear spectrum. To account for massive neutrinos in the power spectrum, we have modified the base Takahashi et al. (2012) prediction using the corrections from Bird et al. (2012). Note that our method is different from the implementation in CAMB where the Bird et al. (2012) corrections use as base the Takahashi et al. (2012) model. For further discussion on the different HALOFIT versions see also Appendix B in Mead et al. (2021). We also note that more accurate non-linear corrections exist, for example HMCODE[2], but they are not necessary given the required accuracy in our analysis.

3.3 Modelling the observable

Building on Section 3.2, we now describe our model for the galaxy-galaxy lensing signal. We first describe the individual terms in the matter-cross-galaxy power spectrum $P_{\rm gm}(k,z)$ (Section 3.3.1), then we project the 3D $P_{\rm gm}(k,z)$ into the 2D lensing power spectrum $C_{\rm gm}(\ell)$ and finally into the observable, the tangential shear $\gamma_t(\theta)$ (Section 3.3.2). In Sections 3.3.3 through 3.3.6 we describe additional astrophysical components that are considered in our model. In Appendix A, we perform a series of tests on our model with simulations and external codes to check for the validity of our code.

Throughout this paper we fix the cosmological parameters to the σ_8 and Ω_m values from the DES Y3 analysis, and use *Planck 2018* (Planck Collaboration, 2020) for remaining parameters. The cosmological analyses on the two lens samples in DES Y3 give consistent results (DES Collaboration, 2021), albeit slightly different, with Ω_m and σ_8 being the best constrained parameters. For this reason, we choose to only use the DES Y3 results for these two cosmological parameters and use the values as constrained for each lens galaxy

2. `https://github.com/alexander-mead/HMcode`

sample separately. For REDMAGIC we use $\Omega_{\rm m} = 0.341$ and $\sigma_8 = 0.735$, while for MAGLIM we use $\Omega_{\rm m} = 0.339$ and $\sigma_8 = 0.733$. For the remaining cosmological parameters we set $\Omega_{\rm b} = 0.0486$, $H_0 = 67.37$, $n_s = 0.9649$, $\Omega_\nu h^2 = 0.0006$, where h is the Hubble constant in units of 100 km/s/Mpc. Since we consider the Λ-Cold Dark Matter (ΛCDM) cosmological model, we set $w = -1$ for the dark energy equation of state parameter. In addition, all the halo masses use the definition of $M_{200\rm c}$, based on the mass enclosed by radius $R_{200\rm c}$ so that the mean density of a halo is 200 times the critical density at the redshift of the halo. We note that the choice of cosmological parameters mostly affects the inferred large-scale galaxy bias, as we show in Section 3.7.3.

In the DES Y3 3×2pt cosmological analysis (DES Collaboration, 2021) using the REDMAGIC lens sample, it was found that the best-fit galaxy clustering amplitude, b_w, is systematically higher than that of galaxy-galaxy lensing, namely b_{γ_t}. To account for this a de-correlation parameter $X_{\rm lens}$ was introduced, that is defined as the ratio of the two biases, $X_{\rm lens} \equiv b_{\gamma_t}/b_w$. This parameter varies from 0 to 1 and allows for the two biases to vary independently, thus enabling the model to achieve simultaneously good fits to both γ_t and w. Nevertheless, the impact of $X_{\rm lens}$ on the main 3×2pt cosmological constraints, especially on $S_8 \equiv \sigma_8(\Omega_m/0.3)^{1/2}$, were negligible. The exact origin of this inconsistency in REDMAGIC, caused by some measurable unknown systematic effect, is still an open question. Given that we do not know if this systematic is affecting the galaxy clustering or galaxy-galaxy lensing signal, or both to some degree, in our galaxy-galaxy lensing analysis we choose to use the fiducial cosmological results from the 3×2pt analysis and assume $X_{\rm lens} = 1$ throughout. However, we briefly discuss the impact on our derived halo properties from changing to the 3×2pt best-fit value of roughly $X_{\rm lens} \approx 0.877$ when we present our results in Section 3.7.2. We do note, however, that this is the most pessimistic case where the systematic is completely found in γ_t. Given that γ_t is a cross-correlation, while e.g. w is an auto-correlation of the lenses, it is likely that clustering is the most affected by the systematic and not galaxy-

galaxy lensing. In our case, this means that the shift in constraints we quote later would not be as dramatic in reality.

3.3.1 Correlations between galaxy positions and the dark matter distribution

The galaxy-cross-matter power spectrum, $P_{\rm gm}(k,z)$, is composed two terms. The 1-halo term, $P^{\rm 1h}_{\rm gm}(k,z)$, quantifies correlations between dark matter and galaxies inside the halo. The 2-halo term, $P^{\rm 2h}_{\rm gm}(k,z)$, quantifies correlations between the halo and neighboring halos. Each of these terms receives a contribution from central and satellite galaxies. Below we summarise the formalism for these four terms separately. The modeling we follow below is similar to what is being commonly used in the literature; for example, see Seljak (2000); Mandelbaum et al. (2005); Park et al. (2015).

The central 1-halo term describes how the dark matter density distribution inside the halo correlates with the central galaxy, and is thus written as

$$P^{\rm c1h}_{\rm gm}(k,z) = \frac{1}{\rho_m \bar{n}_g} \int dM_h \frac{dn}{dM_h} M_h \langle N_c(M_h)\rangle u_{\rm dm}(k|M_h) \,, \tag{3.4}$$

where $u_{\rm dm}(k|M_h)$ is the Fourier transform of the dark matter density distribution as a function of wavenumber k given a halo of mass M_h.

The satellite 1-halo term describes how the satellite galaxies are spatially distributed within the dark matter host halo, and can be written as:

$$P^{\rm s1h}_{\rm gm}(k,z) = \frac{1}{\rho_m \bar{n}_g} \int dM_h \frac{dn}{dM_h} M_h \langle N_s(M_h)\rangle u_{\rm dm}(k|M_h) u_{\rm sat}(k|M_h) \tag{3.5}$$

with $u_{\rm sat}$ being the Fourier transform of the satellite distribution in the halo. For both $u_{\rm dm}$ and $u_{\rm s}$ we assume NFW profiles with concentration parameters $c_{\rm dm}$ and $c_{\rm sat}$, respectively. The distribution of satellite galaxies is typically less concentrated than that of the dark matter (Carlberg et al., 1997; Nagai & Kravtsov, 2005; Hansen et al., 2005; Lin et al., 2004).

To account for this we allow $c_{\rm sat}$ to be smaller than $c_{\rm dm}$ by introducing the free parameter $a = c_{\rm sat}/c_{\rm dm}$, which is allowed to take values between 0 and 1. The total 1-halo power spectrum is then given by

$$P_{\rm gm}^{\rm 1h}(k,z) = P_{\rm gm}^{\rm c1h}(k,z) + P_{\rm gm}^{\rm s1h}(k,z) \ . \tag{3.6}$$

To introduce the 2-halo terms, we define the following quantities: the average linear galaxy bias and the average satellite fraction of our sample.

The average linear galaxy bias is given by:

$$\bar{b}_g = \int dM_h \frac{dn}{dM_h} b_h(M_h) \frac{\langle N(M_h)\rangle}{\bar{n}_g} \ . \tag{3.7}$$

The halo bias relation $b_h(M_h)$ quantifies the dark matter clustering with respect to the linear dark matter power spectrum, and we adopt the functions in Tinker et al. (2010) for it. In the above equation we define the average number density of galaxies as

$$\bar{n}_g = \int dM_h \frac{dn}{dM_h} \langle N(M_h)\rangle \ , \tag{3.8}$$

and is thus also determined by the HOD.

The satellite galaxy fraction is expressed as:

$$\alpha_{\rm sat} = \int dM_h \frac{dn}{dM_h} \frac{\langle N_s(M_h)\rangle}{\bar{n}_g} \ . \tag{3.9}$$

With $\bar{b}_g$ and $\alpha_{\rm sat}$ defined, the 2-halo central galaxy-dark matter cross power spectrum is

then given by:

$$\begin{aligned} P_{\rm gm}^{\rm c2h}(k,z) =& P_{\rm m}^{\rm nl}(k,z) \\ & \times \int dM_h \, \frac{dn}{dM_h} \frac{M_h}{\rho_m} b_h(M_h) u_{\rm dm}(k|M_h) \\ & \times \int dM_h' \, \frac{dn}{dM_h'} \frac{\langle N_c(M_h') \rangle}{\bar{n}_g} b_h(M_h') \; . \end{aligned} \tag{3.10}$$

At large scales, where $u_{\rm dm}(k|M_h) \to 1$, the first integral in the above equation must go to unity, which implies that the halo bias relation must satisfy the consistency relation that the dark matter is unbiased with respect to itself (Scoccimarro et al., 2001). Furthermore, at the same limit, the second integral approaches $(1 - \alpha_{\rm sat})\bar{b}_g$. Therefore, the $k \to 0$ limit of Equation (3.10) reduces to $P_{\rm gm}^{\rm c2h}(k \to 0, z) \approx (1 - \alpha_{\rm sat})\bar{b}_g P_{\rm m}^{\rm lin}(k,z)$.

Similarly, we can express the 2-halo matter-cross-satellite power spectrum as:

$$\begin{aligned} P_{\rm gm}^{\rm s2h}(k,z) =& P_{\rm m}^{\rm nl}(k,z) \\ & \times \int dM_h \, \frac{dn}{dM_h} \frac{M_h}{\rho_m} b_h(M_h) u_{\rm dm}(k|M_h) \\ & \times \int dM_h' \, \frac{dn}{dM_h'} \frac{\langle N_s(M_h') \rangle}{\bar{n}_g} b_h(M_h') u_{\rm sat}(k|M_h') \; . \end{aligned} \tag{3.11}$$

Similar as above, Equation (3.11) reduces to $P_{\rm gm}^{\rm s2h}(k \to 0, z) \approx \alpha_{\rm sat}\bar{b}_g P_{\rm m}^{\rm lin}(k,z)$. Therefore, putting it all together, at the large-scale limit the 2-halo galaxy-dark matter cross power spectrum reduces to

$$P_{\rm gm}^{\rm 2h}(k,z) = P_{\rm gm}^{\rm c2h}(k,z) + P_{\rm gm}^{\rm s2h}(k,z) \approx \bar{b}_g P_{\rm m}^{\rm lin}(k,z) \; , \tag{3.12}$$

which is what is used in cosmological analyses.

In the 2-halo central galaxy-dark matter cross power spectrum of Equation (3.10), in order to avoid double-counting of halos sometimes the *halo exclusion* (HE) technique is used. Based

on the HE principle (see, e.g. Tinker et al. (2005)), given a halo of mass M_{h1} we only consider nearby halos of mass M_{h2} that satisfy the relation $R_{200c}(M_{h1}) + R_{200c}(M_{h2}) \leq r_{12}$, where $R_{200c}(M_h)$ is the radius of a halo of mass M_h, and r_{12} represents the distance between the centers of the two halos. However, accounting for halo exclusion this way is computationally expensive. For this reason, many effective descriptions have been suggested in the literature to bypass this restriction. After performing tests using a simplified HE model in Appendix C, we find that in our case HE has little to no impact on our model, and we thus decide to neglect it in our fiducial framework.

Finally, in order to get the total power spectrum, $P_{\rm gm}(k, z)$, we combine the 1-halo and 2-halo components. We do so by taking the largest of the two contributions at each k. We perform this operation in real space by transforming the power spectrum to its corresponding 3D correlation function $\xi(r, z)$ and taking the maximum:

$$\xi_{\rm gm}(r, z) = \begin{cases} \xi^{\rm 1h}_{\rm gm}(r, z) & \text{if } \xi^{\rm 1h}_{\rm gm} \geq \xi^{\rm 2h}_{\rm gm} \\ \xi^{\rm 2h}_{\rm gm}(r, z) & \text{if } \xi^{\rm 1h}_{\rm gm} < \xi^{\rm 2h}_{\rm gm} \end{cases} . \tag{3.13}$$

We then transform $\xi_{\rm gm}(r, z)$ back to the total galaxy-cross-matter power spectrum $P_{\rm gm}(k, z)$. This is the same approach followed by Hayashi & White (2008); Zu et al. (2014) and is also utilised by Clampitt et al. (2017). We note here that modeling the transition regime from 1-halo to 2-halo scales is not straightforward, and different prescriptions of how to combine the 1-halo and 2-halo components have been suggested. Furthermore, we note that having adopted the common way of modeling the 2-halo component, we have made the assumption that halos are linearly biased tracers of the underlying dark matter distribution, and we make use of a scale-independent halo bias model. As stressed by Mead et al. (2021), a linear halo bias is not necessarily a good description of the clustering relation between the halos and matter, especially on the transition scales. It could thus be important to incorporate a non-linear halo bias model into the halo model. Implementing such a "beyond-linear" halo

bias model, as described in that paper, into our framework would change the shape of the 2-halo component as a function of k, especially around the scales corresponding to the size of individual dark matter halos. We leave this aspect of the model to be investigated in future work.

3.3.2 Modeling the tangential shear γ_t

Armed with the HOD-dependent galaxy-cross-matter power spectrum, we can now follow the standard procedure in deriving the tangential shear γ_t as done in other large-scale cosmological analyses (Cacciato et al., 2009; Mandelbaum et al., 2013; Clampitt et al., 2017; Prat et al., 2017; Prat et al., 2022). We first construct the lensing angular power spectrum, $C_{\rm gm}$, and then transform it to real space. Under the Limber approximation we define the projected, two-dimensional lensing power spectrum as

$$C_{\rm gm}(\ell|z_\ell, z_s) = \frac{\rho_m \Sigma_c^{-1}(z_\ell, z_s)}{\chi^2(z_\ell)} P_{\rm gm}\left(\frac{\ell + 1/2}{\chi(z_\ell)}, z_\ell\right), \tag{3.14}$$

where the critical surface density at lens redshift z_ℓ and source redshift z_s is given by:

$$\Sigma_c(z_\ell, z_s) = \frac{c^2}{4\pi G} \frac{a(z_\ell)\chi(z_s)}{\chi(z_\ell)\chi(z_\ell, z_s)} \, . \tag{3.15}$$

Here $a(z)$ is the scale factor of the universe at redshift z. In the above expression, $\chi(z_\ell)$ and $\chi(z_s)$ are the comoving distances to the lens and source galaxies, while $\chi(z_\ell, z_s)$ is the comoving distance between the lens and source redshifts. The $a(z_\ell)$ factor comes from the use of comoving distances, while c and G are the speed of light and Newton's gravitational constant, respectively.

The expressions we have introduced above are for specific lens and source galaxy redshift pairs; however, in practice we are working with distribution of galaxies in redshift. We denote the probability density functions (PDF) of the lens and source redshift by $n_\ell(z_\ell)$ and

$n_s(z_s)$, respectively. The observed lensing spectrum is given by

$$\begin{aligned} C_{\rm gm}(\ell) &= \int dz_\ell \; n_\ell(z_\ell - \Delta z_\ell^i) \int dz_s \; n_s(z_s - \Delta z_s^j) C_{\rm gm}(\ell | z_\ell, z_s) \\ &= \frac{3}{2} \frac{H_0^2 \Omega_m}{c^2} \int dz_\ell \; n_\ell(z_\ell - \Delta z_\ell^i) \frac{g(z_\ell)(1+z_\ell)}{\chi(z_\ell)} P_{\rm gm}\left(\frac{\ell + 1/2}{\chi(z_\ell)}, z_\ell \right) \; , \end{aligned} \tag{3.16}$$

where the projection kernel is

$$g(z) = \int_z^\infty dz' n_s(z' - \Delta z_s) \frac{\chi(z') - \chi(z)}{\chi(z')} \; . \tag{3.17}$$

The parameters Δz_ℓ and Δz_s in this equation represent the bias of the mean of the lens and source redshift distributions, similar to that used in Krause et al. (2021b).

The tangential shear, under the flat-sky approximation, then becomes:

$$\gamma_t(\theta) = (1+m) \int \frac{\ell d\ell}{2\pi} C_{\rm gm}(\ell) J_2(\ell\theta) \; , \tag{3.18}$$

where $J_2(x)$ is the second-order Bessel function of the first kind. Again following Krause et al. (2021b), the multiplicative bias parameter m in this expression quantifies uncertainties in the shear estimation. We note here that, our analysis differs from that of Krause et al. (2021b), as well as Prat et al. (2022), which does no make the flat-sky approximation. We have checked that this makes a negligible difference in our analysis over the angular scales we use.

3.3.3 Tidal stripping of the satellites

In addition to the four components described in Section 3.3.1, corresponding to the 1- and 2-halo, satellite and central component of $P_{\rm gm}$, as we get to higher accuracy in the measurements higher-order terms in the halo model could become important. The next-order term in the Halo Model is commonly referred to as the *satellite strip* component, which we denote

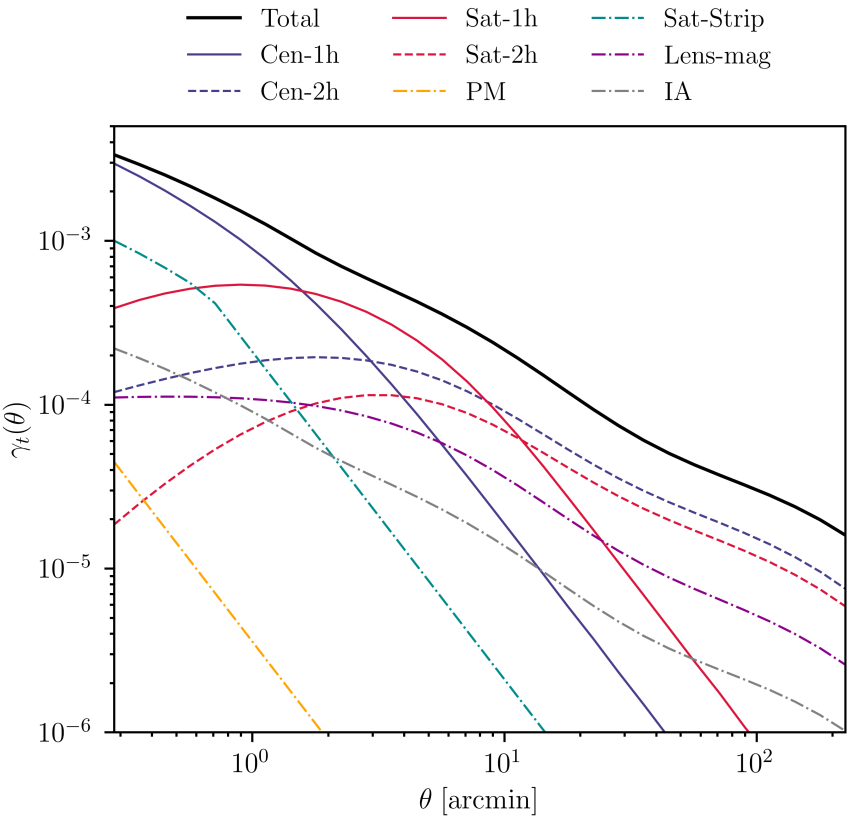


Figure 3.2: This plots illustrates the theory prediction for the shear (solid black) and how the various components contribute to it. The 1- and 2-halo components from the central and satellite galaxies are labeled 'Cen-1h', 'Cen-2h', 'Sat-1h' and 'Sat-2h', respectively. We also show the contribution from IA, lens magnification ('Lens-mag'), satellite strip ('Sat-Strip') and point mass ('PM'). The HOD parameters used are the same as in Figure 3.1; the stellar mass we used is $M_\star = 2 \times 10^{10}$ $M_\odot$; for IA we used the amplitude and power-law parameters $A_{\rm IA} = 0.1$ and $\eta_{\rm IA} = -0.5$, respectively; for the lens magnification coefficient we set the value to $\alpha_{\rm lmag} = 1.3$.

by $\gamma_t^{\rm strip}$. This term is effectively a 1-halo term correlating the satellite galaxies and its own subhalo. As tidal disruptions in the outskirts of the host halo strips off the dark matter content of the satellite subhalo, the density profile of the subhalos drops off at large scales. Therefore, we model this term as a truncated NFW profile which is similar to that of the central 1-halo, $\gamma_t^{\rm c1h}$, out to the truncation radius R and falls off as $\propto r^{-2}$ at larger radii r. The truncation radius is set to $R = 0.4R_{200c}$ and thus does not introduce free parameters to our model. Additionally, since this is a satellite term, it needs to be multiplied by $\alpha_{\rm sat}$, therefore resulting in

$$\gamma_t^{\rm strip}(\theta) = \alpha_{\rm sat} \times \begin{cases} \gamma_t^{\rm c1h}(\theta) & \text{if } r \leq R \\ \gamma_t^{\rm c1h}(R) \left(\frac{R}{r}\right)^2 & \text{if } r > R \end{cases} , \tag{3.19}$$

where $r = r(\theta; z_\ell)$ is the radius from the center of the (sub-)halo at redshift z_ℓ that corresponds to angular scale θ. Note that this is similar to what is used in Mandelbaum et al. (2005); Velander et al. (2013), but is using a mass definition based on $\rho_{200c} = 200\rho_c$ for the halos.

3.3.4 Point-mass contribution

An additional term to γ_t is the contribution to lensing by the baryonic content of the central galaxy (e.g. Velander et al., 2013). This term is simply modelled as a point-source term given by

$$\gamma_t^{\rm PM}(\theta) = \int dz_\ell n_\ell(z_\ell) \frac{M_\star}{\pi r^2(\theta, z_\ell)} \int dz_s n_s(z_s) \Sigma_c^{-1}(z_\ell, z_s) \, . \tag{3.20}$$

Here, $M_\star$ is an effective mass parameter that quantifies the amplitude of the point mass component.

In practice, the amplitude parameter would be allowed to vary as a free parameter or

be set to the average stellar mass inside the redshift bin of interest. When let to vary, it accounts for any imperfect modeling of the galaxy-matter cross-correlation on scales smaller than the smallest measured scale used in the model fit. This is similar to the *point-mass* term derived in MacCrann et al. (2020a) and used in Krause et al. (2021b).

3.3.5 Lens magnification

We now consider the effects of weak lensing magnification on the estimation of our observable. In addition to the distortion (shear) of galaxy shapes, weak lensing also changes the observed flux and number density of galaxies – this effect is referred to as magnification. Following Prat et al. (2022), here we only consider the magnification in flux for the lens galaxies, as that is the dominant effect for galaxy-galaxy lensing.

Similar to shear, magnification is expected to be an increasing function of redshift. In the weak lensing regime, the magnification power spectrum involves an integration of the intervening matter up to the lens redshift and is given by (Unruh et al., 2020)

$$C_{\rm gm}^{\rm lmag}(\ell) = \frac{9H_0^3\Omega_m^2}{4c^3}\int dz_\ell n_\ell(z_\ell)\int_0^{z_\ell} dz \frac{\chi(z,z_\ell)g_{\rm lmag}(z)}{\chi(z)a^2(z)} P_{\rm m}^{\rm nl}\left(\frac{\ell+1/2}{\chi(z)}, z\right) , \tag{3.21}$$

where we have defined

$$g_{\rm lmag}(z) = \int dz_s n_s(z_s)\frac{\chi(z,z_s)}{\chi(z_s)} . \tag{3.22}$$

The contribution to the tangential shear can then be written as

$$\gamma_t^{\rm lmag}(\theta) = 2(\alpha_{\rm lmag}-1)\int \frac{\ell d\ell}{2\pi} C_{\rm gm}^{\rm lmag}(\ell) J_2(\ell\theta) , \tag{3.23}$$

where $\alpha_{\rm lmag}$ is a constant that can be estimated from simulations (Elvin-Poole et al., 2021) and $C_{\rm gm}^{\rm lmag}(\ell)$ is the average of (3.21) over the redshift distributions of the lenses and sources.

In this work we fix $\alpha_{\rm lmag}$ following the Y3 3×2pt analysis and use the values computed in Elvin-Poole et al. (2021), which are $\alpha_{\rm lmag} = \{1.31, -0.52, 0.34, 2.25\}$ for our REDMAGIC and $\alpha_{\rm lmag} = \{1.21, 1.15, 1.88, 1.97\}$ for our MAGLIM lens redshift bins.

3.3.6 Intrinsic alignment

Galaxies are not randomly oriented even in the absence of lensing. On large scales, galaxies can be stretched in a preferable direction by the tidal field of the large scale structure. On small scales, other effects such as the radial orbit of a galaxy in a cluster can affect their orientation. This phenomenon, where the shape of the galaxies is correlated with the density field, is known as *intrinsic alignment* (IA); for a review see Troxel & Ishak (2015).

The contamination of shear by IA can become important in some cases, especially when the source galaxies are physically close to the lenses and gravitational interactions can modify the shape of the galaxies. IA is commonly modeled using the non-linear linear alignment (NLA) model proposed by Hirata & Seljak (2004); Bridle & King (2007); Joachimi et al. (2013). In NLA, the galaxy-cross-matter power spectrum receives an additional term

$$P_{\rm NLA}(k,z) = -A_{\rm IA} C_1 \rho_c \Omega_m D_+^{-1}(z) b P_{\rm m}^{\rm lin}(k,z) \left(\frac{1+z}{1+z_0}\right)^{\eta_{\rm IA}} . \tag{3.24}$$

In the above equation $D_+(z)$ is the linear structure-growth factor at redshift z normalised to unity at $z = 0$, b is the linear bias, $A_{\rm IA}$ determines the overall amplitude, $C_1 = 5 \times 10^{-14} h^{-2} {\rm M}_\odot^{-1} {\rm Mpc}^3$ is a constant, and the power-law index $\eta_{\rm IA}$ models the redshift evolution defined so that the pivot redshift is set to $z_0 = 0.62$.

The IA contribution to galaxy-galaxy lensing simply depends on the galaxy density and has a different projection kernel than Equation (3.16). The projected 2D power spectrum

for NLA is then given in the Limber approximation by

$$C_{\rm NLA}(\ell) = \int dz_\ell \frac{n_\ell(z_\ell) n_s(z_\ell)}{\chi^2(z_\ell)(d\chi/dz)|_{z_\ell}} P_{\rm NLA}\left(\frac{\ell+1/2}{\chi(z_\ell)}, z(\chi_\ell)\right) , \tag{3.25}$$

where $(d\chi/dz)|_{z_\ell}$ is the derivative of the comoving distance with respect to redshift at $z = z_\ell$. To obtain the NLA contribution to the tangential shear, we perform a Hankel transform on $C_{\rm NLA}(\ell)$ using $J_2(\ell\theta)$, as in Equation (3.18).

A simple extension of NLA in our HOD framework will be to use our HOD-based $P_{\rm gm}$ instead of $b_g P_{\rm m}^{\rm nl}$ in Equation (3.24). However, the IA modeling near the 1-halo term is likely more complex and would warrant more detailed studies such as those carried out in Blazek et al. (2015). In this paper, we avoid the complex modeling by choosing redshift bin pairs that are sufficiently separated so that they have significantly low IA contribution (see Section 3.5.1) and we thus choose not to include this component in our fiducial model. However, in Section 3.7.3 we test the full model that includes this IA contribution and show that the results are consistent with our fiducial which does not include IA. We show an example of what all the γ_t components look like in Figure 3.2.

Although we have ignored IA in this paper, given that it is negligible for our purposes, we emphasize that its contribution to lensing can be of high importance to future cosmological studies, as it can produce biases in the inference of the cosmological parameters (e.g. Samuroff et al., 2019). In addition, if not properly accounted for, IA can affect the inference of the lens halo properties in lensing analyses. In this case, a halo-model description of IA would be necessary to capture its sample dependence. Fortuna et al. (2021) described a halo model for IA on small and large scales from central and satellite galaxies which is capable of incorporating the galaxy sample characteristics. We leave the further investigation of IA and its modeling for future work.

3.4 Data

For this work we make use of data from the Dark Energy Survey (DES, Flaugher, 2005). DES is a photometric survey, with a footprint of about 5000 deg^2 of the southern sky, that has imaged hundreds of millions of galaxies. It employs the 570-megapixel Dark Energy Camera (DECam, Flaugher et al., 2015) on the Cerro Tololo Inter-American Observatory (CTIO) 4m Blanco telescope in Chile. We use data from the first three years (Y3) of DES observations. The basic DES Y3 data products are described in Abbott et al. (2018); Sevilla-Noarbe et al. (2021). Below we briefly describe the source and galaxy samples used in this work. By construction, all the samples are the same as that used in Prat et al. (2022) and in the DES Y3 3×2pt cosmological analysis (DES Collaboration, 2021).

3.4.1 Lens galaxies - REDMAGIC

For our first lens sample we use REDMAGIC galaxies. These are red luminous galaxies which provide the advantage of having small photometric redshift errors. The algorithm used to extract this sample of luminous red galaxies is based on how well they fit a red sequence template, calibrated using the red-sequence Matched-filter Probabilistic Percolation cluster-finding algorithm (REDMAPPER, Rykoff et al., 2014, 2016).

To maintain sufficient separation between lenses and sources, we only use the lower four redshift bins used in Prat et al. (2022). The first three bins at $z = \{[0.15, 0.35], [0.35, 0.5], [0.5, 0.65]\}$ consist of the so-called "high-density sample". This is a sub-sample which corresponds to luminosity threshold of $L_{\rm min} = 0.5L_\star$, where $L_\star$ is the characteristic luminosity of the luminosity function, and comoving number density of approximately $\bar{n} \sim 10^{-3}$ $(h/{\rm Mpc})^3$. The fourth redshift bin of $z = [0.65, 0.8]$ is characterised by $L_{\rm min} = L_\star$ and $\bar{n} \sim 4 \times 10^{-4}$ $(h/{\rm Mpc})^3$, and is referred to as the "high-luminosity sample". The redshift distributions for all these bins are shown in Figure 3.3. As we will discuss in Section 3.6 we use the number density values as an additional data point in our fits, which helps constrain the $f_{\rm cen}$

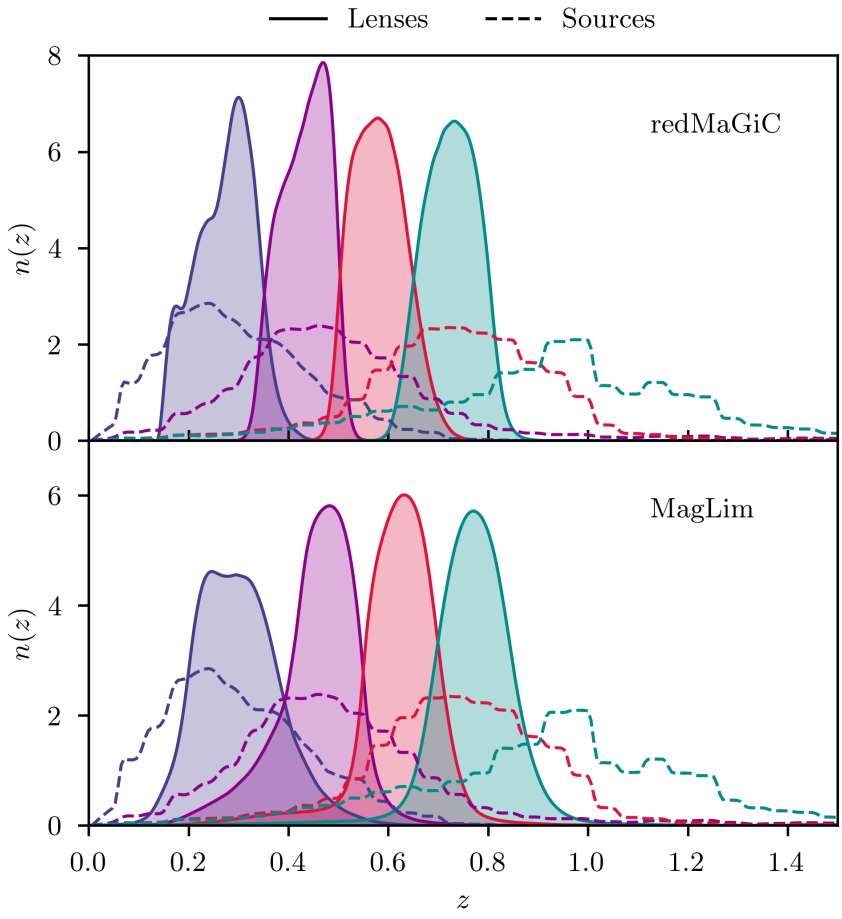


Figure 3.3: Redshift distribution of the lenses (solid filled) and of the source (dashed) galaxies, for redMaGiC (upper) and MagLim (lower).

HOD parameter. The data we used to derive the mean of $\bar{n}_g^i$ and its variance in each lens bin i is the same as what is used in Pandey et al. (2022), and the specific values we used are the following: $\bar{n}_g^i \approx \{9.8 \pm 0.6, 9.6 \pm 0.3, 9.6 \pm 0.2, 3.8 \pm 0.02\} \times 10^{-4}$ $(h/\mathrm{Mpc})^3$, respectively for $i = 1, 2, 3, 4$. We note here that we have also fit our data without the addition of $\bar{n}_g^i$ and our main conclusions hold, except that f_{cen} becomes unconstrained.

3.4.2 Lens galaxies - MagLim

The second sample we use for lens galaxies is MagLim which is defined with a redshift-dependent magnitude cut in i-band. This results in a sample with ~ 4 times more galaxies compared to redMaGiC and is divided into 6 bins in redshift with $\sim 30\%$ wider redshift distributions, also compared to the redMaGiC sample. In this sample, galaxies are selected with a magnitude cut that evolves linearly with the photometric redshift estimate: $i < az_{\mathrm{phot}} + b$. The optimization of this selection, using the DNF photometric redshift estimates (De Vicente et al., 2016), yields $a = 4.0$ and $b = 18$. This optimization was performed taking into account the trade-off between number density and photometric redshift accuracy, propagating this to its impact in terms of cosmological constraints obtained from galaxy clustering and galaxy-galaxy lensing in Porredon et al. (2022). Effectively this selects brighter galaxies at low redshift while including fainter galaxies as redshift increases. Additionally, we apply a lower cut to remove the most luminous objects, $i > 17.5$. Single-object fitting (SOF) magnitudes (a variant of multiobject fitting (MOF) described in Drlica-Wagner et al. (2018)) from the Y3 Gold Catalog were used for sample selection and as input to the photometric redshift codes. See also Porredon et al. (in prep.) for more details on this sample. The redshift distributions of the MagLim sample are shown in Figure 3.3.

3.4.3 Source galaxies

We use the DES Y3 shear catalog presented in Gatti, Sheldon et al. (2020). The galaxy shapes are estimated using the METACALIBRATION (Huff & Mandelbaum, 2017; Sheldon & Huff, 2017) algorithm. The shear catalog has been thoroughly tested in Gatti, Sheldon et al. (2020), and tests specifically tailored for tangential shear have been presented in Prat et al. (2022). In this paper we perform additional tests on this shear catalog for tangential shear measurement on small scales (Section 3.5.3).

Following Prat et al. (2022) we bin the source galaxies into four redshift bins, where details of the redshift binning and calibration is described in Myles, Alarcon et al. (2020). The redshift distributions for the source samples are shown in Figure 3.3.

3.5 Measurements

Our γ_t measurements are carried out using the fast tree code TREECORR[3] (Jarvis et al., 2004). We use the same measurement pipeline as that used in Prat et al. (2022), where details of the estimator, including the implementation of random-subtraction and METACALIBRATION are described therein. The main difference is we extend to smaller scales and add 10 additional logarithmic bins from 0.25 arcmin to 2.5 arcmin. The full data vector in our analysis contains 30 logarithmic bins from 0.25 arcmin to 250 arcmin.

Figures 3.4 and 3.5 show the final measurements using the REDMAGIC and MAGLIM samples as lenses, respectively. The six panels represent the six lens-source redshift bin pairs. The total signal-to-noise for the six redshift bins [Lens, Source]=$\{[1,3],[1,4],[2,3],[2,4],[3,4],[4,4]\}$ are $\sim \{65.5, 59.9, 58.2, 65.5, 55.2, 36.6\}$ for REDMAGIC and $\sim \{104.4, 100.9, 76.6, 99.2, 60.5, 45.5\}$ for MAGLIM numbers. For comparison, the signal-to-noise for the same bin pairs, only accounting for the scales used in the cosmological analysis in Prat et al. (2022) are $\sim$

3. https://github.com/rmjarvis/TreeCorr

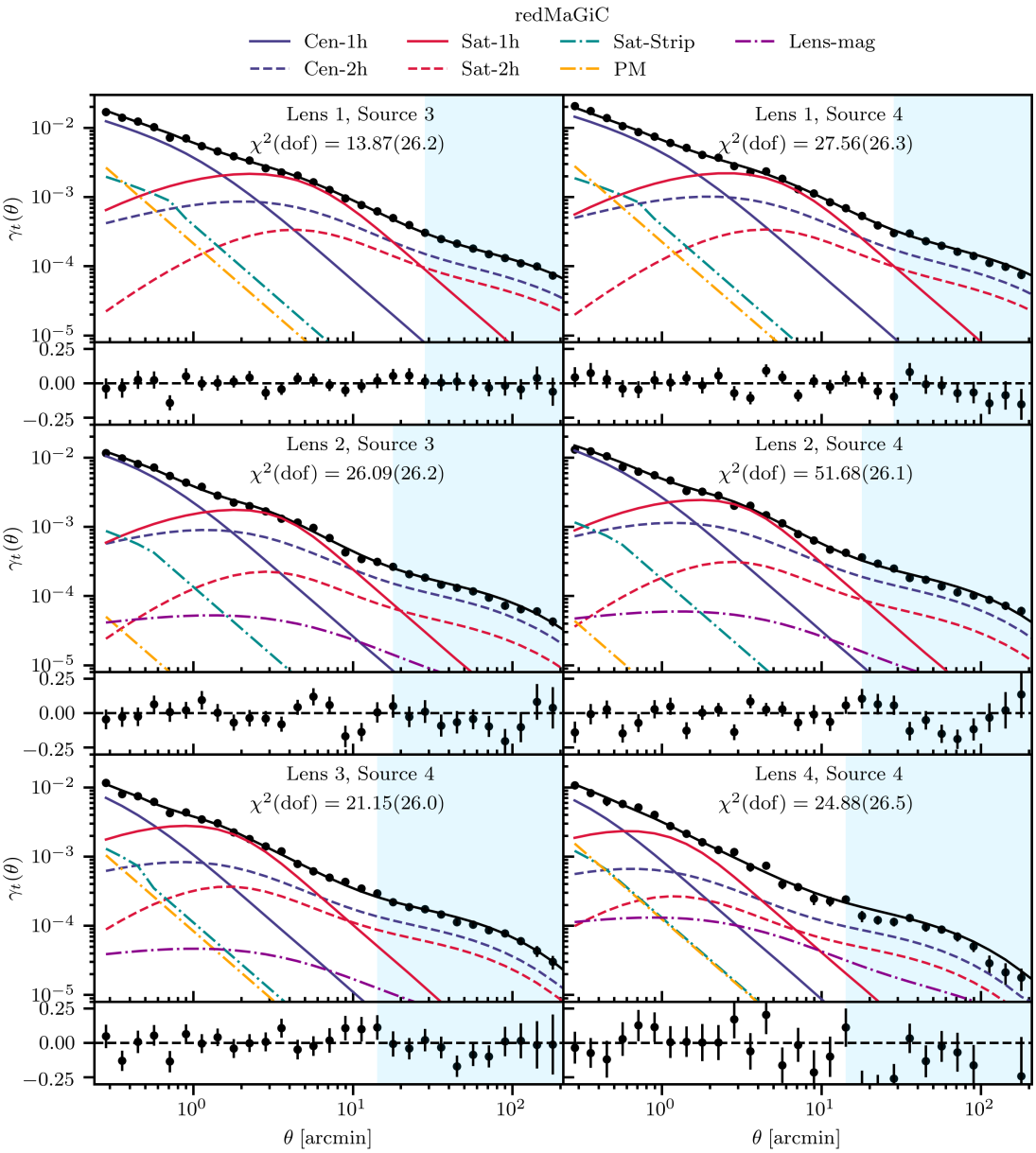


Figure 3.4: Best-fit model (solid black) to REDMAGIC for each lens-source redshift bin combination and the residuals with respect to the data (points) attached below each panel. The various components of the model are also shown: central 1-halo (solid blue) and 2-halo (dashed blue), satellite 1-halo (solid red) and 2-halo (dashed red), satellite strip (dash-dotted orange), point mass (dash-dotted cyan) and lens magnification (dash-dotted green). The blue shaded area marks the scales used in cosmological analyses, while the rest corresponds to the additional small-scale points used in this work. In each panel we also show the total χ^2 of the fit, after applying the Hartlap correction to the inverse covariance matrix, and the number of degrees of freedom.

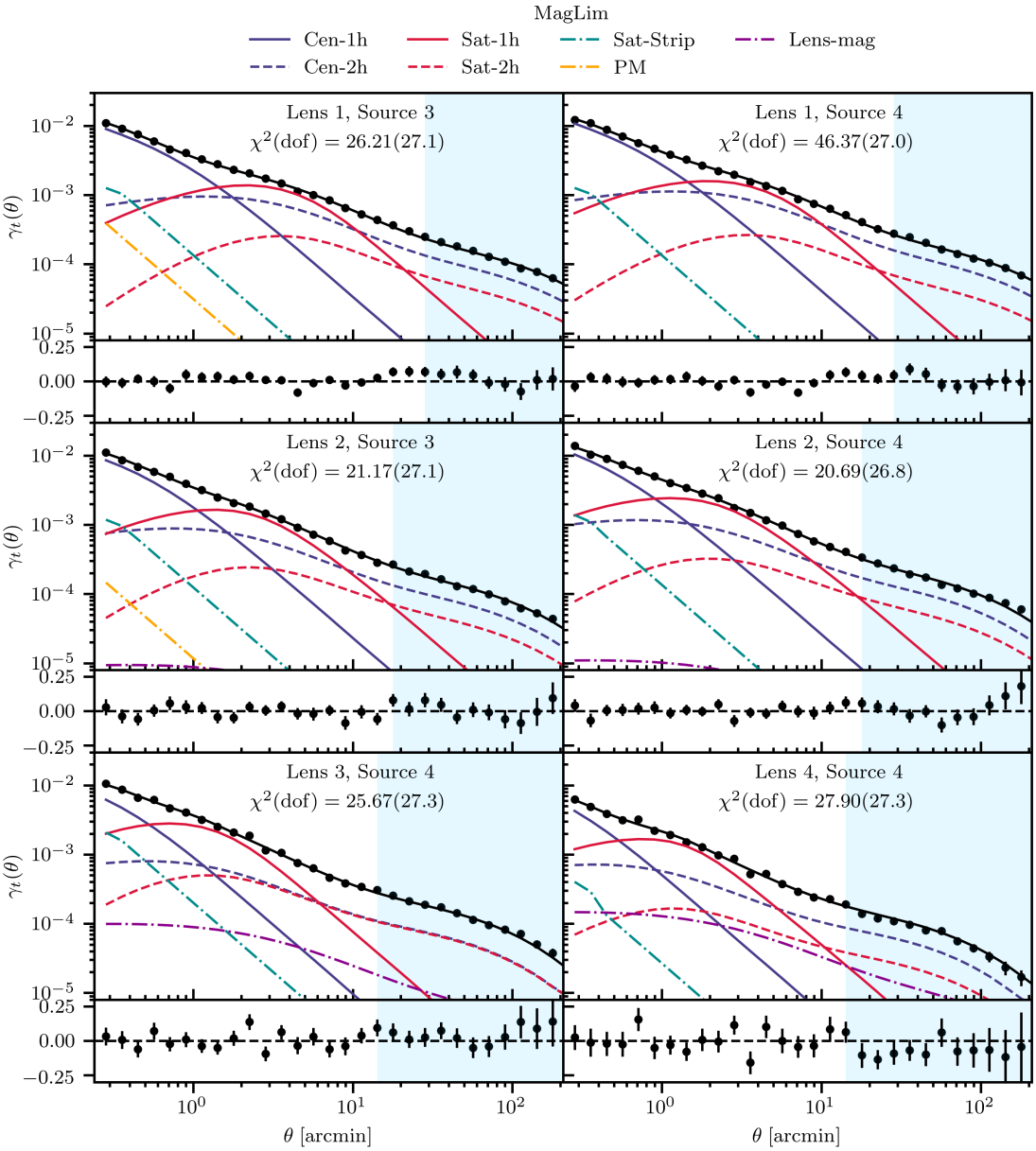


Figure 3.5: Same as Figure 3.4 but for the MagLim sample.

$\{25.1, 26.8, 18.7, 22.1, 18.5, 12.3\}$ for the REDMAGIC sample, and $\sim \{41.2, 35.9, 29.4, 30.4, 21.1, 15.7\}$ for the MAGLIM galaxies. The additional small-scale information from this work increases the signal-to-noise by a factor of 2-3. This again demonstrates that if modelled properly, there is significant statistical power in this data to be harnessed.

Below we briefly describe two elements specifically relevant for this work, the boost factor (Section 3.5.1) and the Jackknife covariance matrix (Section 3.5.2). We also describe briefly the additional data-level tests that we perform to identify any observational systematic effects (Section 3.5.3). Our shear estimator, which includes the boost-factor correction and random-point subtraction (i.e. removing the measured tangential shear measured around isotropically distributed random points in the survey footprint; see Prat et al. (2022) for a more in-depth discussion), is written as (Prat et al., 2022; Pandey et al., 2022):

$$\gamma_t(\theta) = \frac{1}{\langle R \rangle} \left[\frac{\sum_k w_{r_k}}{\sum_i w_{\ell_i}} \frac{\sum_{\rm ij} w_{\ell_i} w_{s_j} e^{\rm LS}_{t,ij}(\theta)}{\sum_{\rm kj} w_{r_k} w_{s_j}} - \frac{\sum_{\rm kj} w_{r_k} w_{s_j} e^{\rm RS}_{t,kj}(\theta)}{\sum_{\rm kj} w_{r_k} w_{s_j}} \right] , \tag{3.26}$$

where w_{ℓ_i}, $w_{r_k} = 1$ and w_{s_j} are the weights associated with the lens galaxy i, random point k and source galaxy j, respectively. Furthermore, the weighted average METACALIBRATION response is $\langle R \rangle = \sum_j w_{s_j} R_{s_j} / \sum_j w_{s_j}$, averaging over the responses R_{s_j} of each source galaxy j, while $e^{\rm LS}_{t,ij}$ and $e^{\rm RS}_{t,kj}$ are, respectively, the measured tangential ellipticity of the source galaxy j around the lens galaxy i and random point k.

3.5.1 Boost factors

While computing the lensing signal we need to take into account that, since galaxies follow a distribution in redshift, namely $n_\ell(z_\ell)$ and $n_s(z_s)$ for lenses and sources respectively, their spatial distributions may overlap. This is something that is naturally accounted for in Equation (3.17) as the lensing efficiency is set to zero when the source is in front of the lens. However, by using fixed $n_\ell(z_\ell)$ and $n_s(z_s)$ in Equation (3.16), we implicitly assume there is no

spatial variation in the lens and source redshift distribution across the footprint. In reality, galaxies are clustered, and the number of sources around a lens can be larger than what we would expect from a uniform distribution. This is usually quantified by the *boost factor* (Sheldon et al., 2004), $B(\theta)$, estimator which is the excess in the number of sources around a lens with respect to randoms. The difference in our γ_t measurements with and without boost factors are shown in Figures B.1 and B.2 (for the full figures, with all lens-source bin combinations, see Prat et al., 2022). As can be seen from the plots, the contribution from this effect can be large at small scales, especially when the bins are more overlapped in redshift. In our analysis we take the boost factors into account by correcting for it before carrying out the model fit. That is, the measurements shown in Figures 3.4 and 3.5 have already been corrected for the boost factor. In addition, since large boost factors will also signal potential failures in parts of our modeling (specifically IA and magnification), we choose to work only with bins that have small boost factors, for which we set a maximum threshold of $\sim 20\%$ deviation from unity, that result in lens and source redshift bin combinations that are largely separated in redshift. We carry out our analysis with 6 lens-source pairs for both lens samples: [Lens 1, Source 3], [Lens 1, Source 4], [Lens 2, Source 3], [Lens 2, Source 4], [Lens 3, Source 4], [Lens 4, Source 4].

3.5.2 Covariance matrix

We use a Jackknife (JK) covariance in this work defined as

$$\mathcal{C}_{ij} \equiv \mathcal{C}(\gamma_t(\theta_i), \gamma_t(\theta_j)) = \frac{N_{\rm JK}-1}{N_{\rm JK}} \sum_{k=1}^{N_{\rm JK}} \Delta\gamma_i^k \Delta\gamma_j^k \ , \tag{3.27}$$

where $\gamma_t^k(\theta_i)$ is the shear in the i'th angular bin for the k'th JK resampling, $\langle\gamma_t(\theta_i)\rangle_k$ is the average over all $N_{\rm JK}$ realizations of the shear for the i'th angular bin and we have defined $\Delta\gamma_i^k \equiv \gamma_t^k(\theta_i) - \langle\gamma_t(\theta_i)\rangle_k$.

We use $N_{\rm JK} = 150$ JK patches for this work defined via the KMEANS[4] algorithm. $N_{\rm JK}$ is chosen so that the individual JK regions are at least as large as the maximum angular scale we need for our measurements. See Prat et al. (2022) for a comparison between the JK diagonal errors and the halo-model covariance errors, which are in good agreement.

When inverting the covariance matrix in the likelihood analysis, a correction factor is needed to account for the bias introduced from the noisy covariance (Friedrich et al., 2016). This correction is often referred to as the Hartlap (Hartlap et al., 2007) correction. When inverting the JK covariance matrix $\mathcal{C}$ we multiply it by a factor H to get the unbiased covariance (Kaufman, 1967)

$$\mathcal{C}_H^{-1} = H\mathcal{C}^{-1} = \left(\frac{N_{\rm JK} - N_\theta - 2}{N_{\rm JK} - 1}\right)\mathcal{C}^{-1}\,, \tag{3.28}$$

where the number of angular bins we use is $N_\theta = 30$, since we analyze each lens-source redshift bin combination independently. As shown in Hartlap et al. (2007), for $N_\theta/N_{\rm JK} < 0.8$ the correction produces an unbiased estimate of the inverse covariance matrix; in our case we find $N_\theta/N_{\rm JK} = 0.2$. However, it is also shown in Hartlap et al. (2007) that as this factor increases, $N_\theta/N_{\rm JK} \to 0.8$, the Bayesian confidence intervals can erroneously grow by up to 30%. Furthermore, it was shown that in order for the confidence intervals to not grow more than 5% the factor $N_\theta/N_{\rm JK} \lesssim 0.12$. For our results this means that, although our covariance matrix gets unbiased, our error bars increase and our constraints can thus look less significant than they actually are.

We finally discuss our choice of a Jackknife covariance matrix in this work. The fiducial covariance used in the 3×2pt analysis in DES Y3 is derived from an analytic halo-model formulation presented in Friedrich et al. (2020). Since our halo model implementation is different from that work (e.g. the modeling of the 1-to-2 halo regime and the HOD parametrization), we cannot use the same framework. Furthermore, since our goal is to model very small

4. `https://github.com/esheldon/kmeans_radec`

scales, where the HOD is needed to model the galaxy bias, using as input to the covariance calculation the HOD would lead to a circular process. Therefore, we opt to use the JK covariance which is not relying on halo-model assumptions.

3.5.3 Systematics diagnostic tests

Similar to Prat et al. (2022), we carry out a series of data-level tests to check for any systematic contamination in the data products. As this work extends from Prat et al. (2022) in terms of the scales used for the analysis, we extend the following tests to the 0.25-2.5 arcmin scales. The tests we performed are the following:

1. *Cross component*: The tangential shear, γ_t, is one of the two components when we decompose a spin-2 shear field. The other component is $\gamma_\times$, which is defined by the projection of the field onto a coordinate system which is rotated by 45° relative to the tangential frame. For isotropically oriented lenses, the average of $\gamma_\times$ due to gravitational lensing alone should be zero. It is thus a useful test to measure this component in the data and make sure that it is consistent with zero for all angular scales. To be able to decide whether this is the case, we report the total χ^2 calculated for $\gamma_\times$ when compared with the null signal.

2. *Responses*: In this work, to measure the shear we make use of the METACALIBRATION algorithm (Sheldon & Huff, 2017; Zuntz et al., 2018b). Based on this, a small known shear is applied to the images and then the galaxy ellipticities $\mathbf{e}$ are re-measure on the sheared images to calculate the response of the estimator to shear. This can be done on every galaxy, and the average response over all galaxies is $\langle \mathbf{R}_\gamma \rangle$. Then, the average shear is $\langle \gamma_t \rangle = \langle \mathbf{R}_\gamma \rangle^{-1} \langle \mathbf{e} \rangle$. Moreover, the METACALIBRATION framework allows us to also correct for selection responses, $\langle \mathbf{R}_S \rangle$, produced due to selection effects (e.g. by applying redshift cuts). The final response would then be the sum of the two effects,

$\langle \mathbf{R} \rangle = \langle \mathbf{R}_\gamma \rangle + \langle \mathbf{R}_S \rangle$. In practice, this procedure can be performed in an exact, scale-dependent way or be approximated by an average scale-independent response, $\langle R_\gamma \rangle$. In this test, we show that this approximation is sufficiently good by comparing the measured shear derived from both of these methods.

3. *LSS weights*: Photometric surveys are subject to galaxy density variations throughout the survey footprint due to time-dependent observing conditions. This variation in the density of the lenses must be accounted for by applying the LSS-weights, which removes this dependence on observing conditions, such as exposure time and air-mass. In galaxy-galaxy lensing, since it is a cross-correlation probe, the impact of observing conditions is small compared to e.g. galaxy clustering. Therefore, in this test we compare the shear measurements with and without the application of the LSS-weighting scheme and report the difference between the two.

We show in Appendix B the results of these tests, where we do not find significant signs of systematic effects in our data vector.

3.6 Model fitting

In this section we discuss how we have performed the fitting of the HOD model introduced in Section 3.2 to our data. We have five HOD parameters ($M_{\rm min}$, $\sigma_{\log M}$, $f_{\rm cen}$, M_1, α), two parameters that correspond to the additional contributions to lensing from point-mass ($M_\star$) and the different satellite spatial distribution compared to that of the dark matter ($a = c_{\rm sat}/c_{\rm dm}$), and three parameters to account for systematic uncertainties (Δz^i_ℓ, Δz^i_s, m^i). For the MagLim sample we have additional parameters (Σ^i_ℓ) that correspond to the stretching factors of the lens redshift distributions, which are further discussed in Porredon et al. (2021).

Our priors on these parameters are shown in Table 3.1. We will discuss in Section 3.7 the

effects of these priors and whether they are appropriate in fitting all redshift bins. The choice of priors on the HOD parameters was based on previous works on red galaxies (Brown et al., 2008; White et al., 2011; Rykoff et al., 2014, 2016), and is similar to the priors in Clampitt et al. (2017) but modified to better suit our HOD parametrization. As for the Δz^i and m^i parameters, our Gaussian priors on them are the same as in Myles, Alarcon et al. (2020) and in MacCrann et al. (2020b). The priors we apply on $M_\star$ and $a = c_{\rm sat}/c_{\rm gm}$ are derived from our tests in Section 3.7.3.

Our full data vector for the REDMAGIC sample consists of the γ_t measurements to which we append the additional data point $\bar{n}_g^i$, the average number density of galaxies in each lens redshift bin i, as mentioned in Section 3.4.1. As we discuss in Section 3.7.1, the addition of this information helps control some of the model parameter constraints. To account for this in the covariance, we formed the full covariance matrix of our data vector by appending to $\mathcal{C}_{ij}$ the variance of $\bar{n}_g^i$ on the diagonal, with zero off-diagonal entries. Our usage of $\bar{n}_g^i$ effectively serves as a prior in our fits. We note here that we do not add $\bar{n}_g^i$ in the data vector of MAGLIM, as we discuss in Section 3.7.1.

Finally, for reasons we will discuss in more detail in Section 3.7.1, we apply a prior on the satellite fraction specifically in the highest-redshift bin we fit, namely [Lens 4, Source 4], for the REDMAGIC sample. In summary, this prior is based on the observation that most of the galaxies in that redshift range are expected to be central and thus we choose to use the flat prior range $[0, 0.2]$ for $\alpha_{\rm sat}$. Note that a similar approach is adopted in van Uitert et al. (2011) (see Appendix C therein) and Velander et al. (2013) for high-redshift red galaxies.

To sample the posterior of each data set we utilise the MULTINEST[5] sampler, which implements a nested sampling algorithm (see for example Feroz et al., 2009). In our analysis we assume that our data is generated by an underlying Gaussian process, thus making its covariance Gaussian in nature. Therefore, for data vector $\mathbf{d}$ of length N_d and model

5. `https://github.com/JohannesBuchner/MultiNest`

prediction vector $\mathbf{m}$ of the same length we express the log-likelihood as

$$\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{d}-\mathbf{m})^T \mathcal{C}_H^{-1}(\mathbf{d}-\mathbf{m}) \equiv -\frac{\chi^2}{2} \,, \tag{3.29}$$

where $\boldsymbol{\theta}$ is the parameter vector of our model $\mathcal{M}$ and $\mathcal{C}_H^{-1}$ is the Hartlap-corrected data covariance matrix (see discussion in Section 3.5.2). Notice that we have neglected the constant factors which are not useful while sampling the likelihood.

For our model fits, we report the total χ^2 of our best-fit model to the data, as a measure of the goodness of fit. Alongside this we report the number of degrees of freedom (dof), which we calculate as the effective number of parameters that are constrained by the data, $N_{\rm eff} = \mathrm{tr}\left[\mathcal{C}_\Pi^{-1}\mathcal{C}_H\right]$, subtracted from the number of data points, N_d:

$$N_{\rm dof} = N_{\rm d} - \mathrm{tr}\left[\mathcal{C}_\Pi^{-1}\mathcal{C}_H\right] \,, \tag{3.30}$$

where the prior covariance is $\mathcal{C}_\Pi$. We should note here that a goodness-of-fit estimation based on finding an effective number of parameters is not always straightforward when the parameters do not enter the model linearly, as discussed in Section 6.3 of Joachimi et al. (2021). Therefore, our approach of calculating a reduced χ^2 using Equation (3.29) based on the $N_{\rm dof}$ from (3.30) yields a conservative answer if model under-fitting is the main concern.

3.7 Results

In this section we present the results from our analysis[6] Before unblinding we performed several validation tests of our pipeline using simulations and simulated data vectors. After the tests were successfully passed, and after unblinding of the data, we applied our full methodology to the unblind measurements to derive our main results. We first present in

6. In what follows we discuss our results after unblinding the data (see Muir et al. (2020) for details on the data blinding procedure).

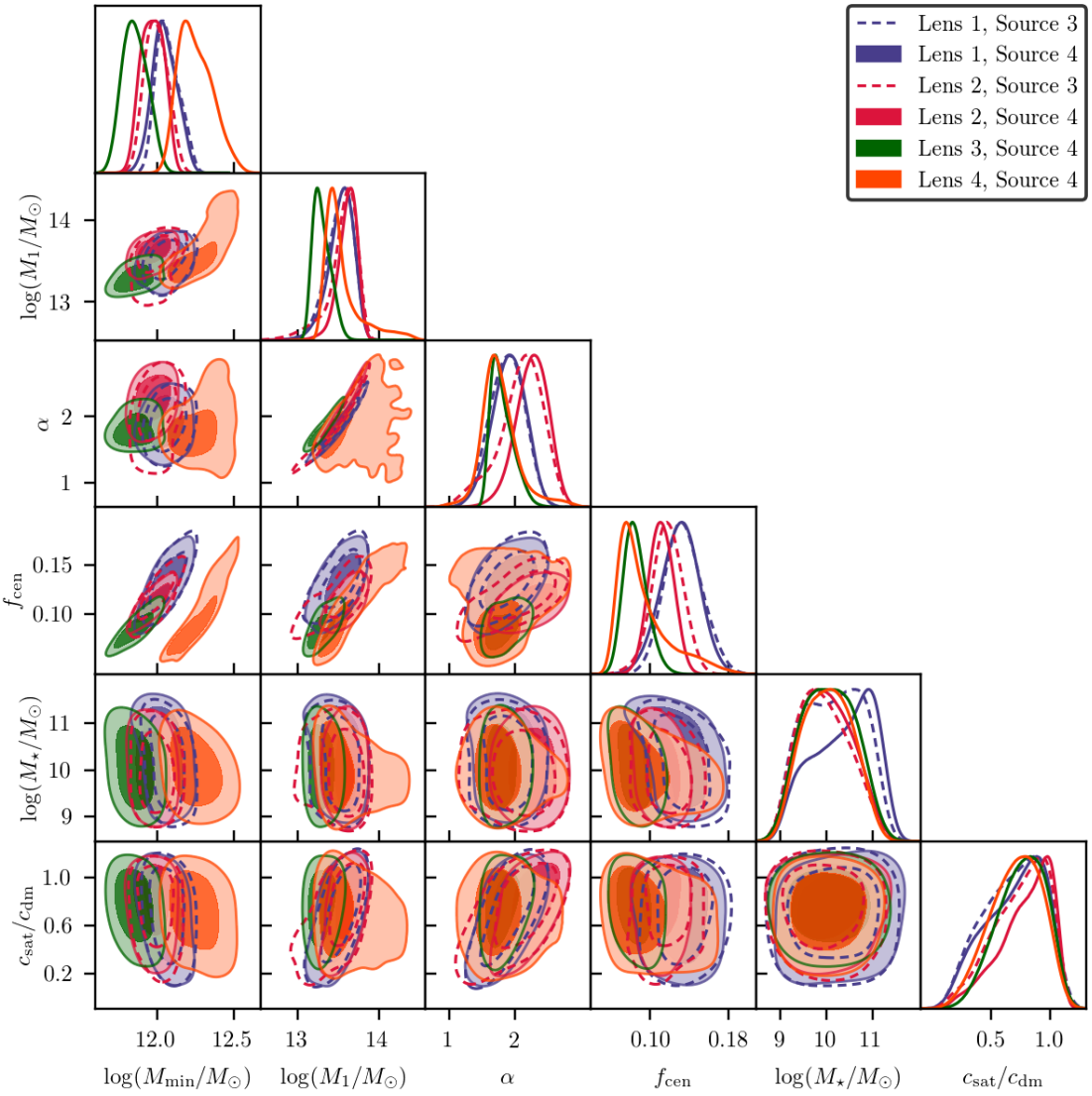


Figure 3.6: Parameter constraints for REDMAGIC using the fiducial cosmology. Combinations with the same lens bin but different source bins are presented with the same colors (solid versus dashed).

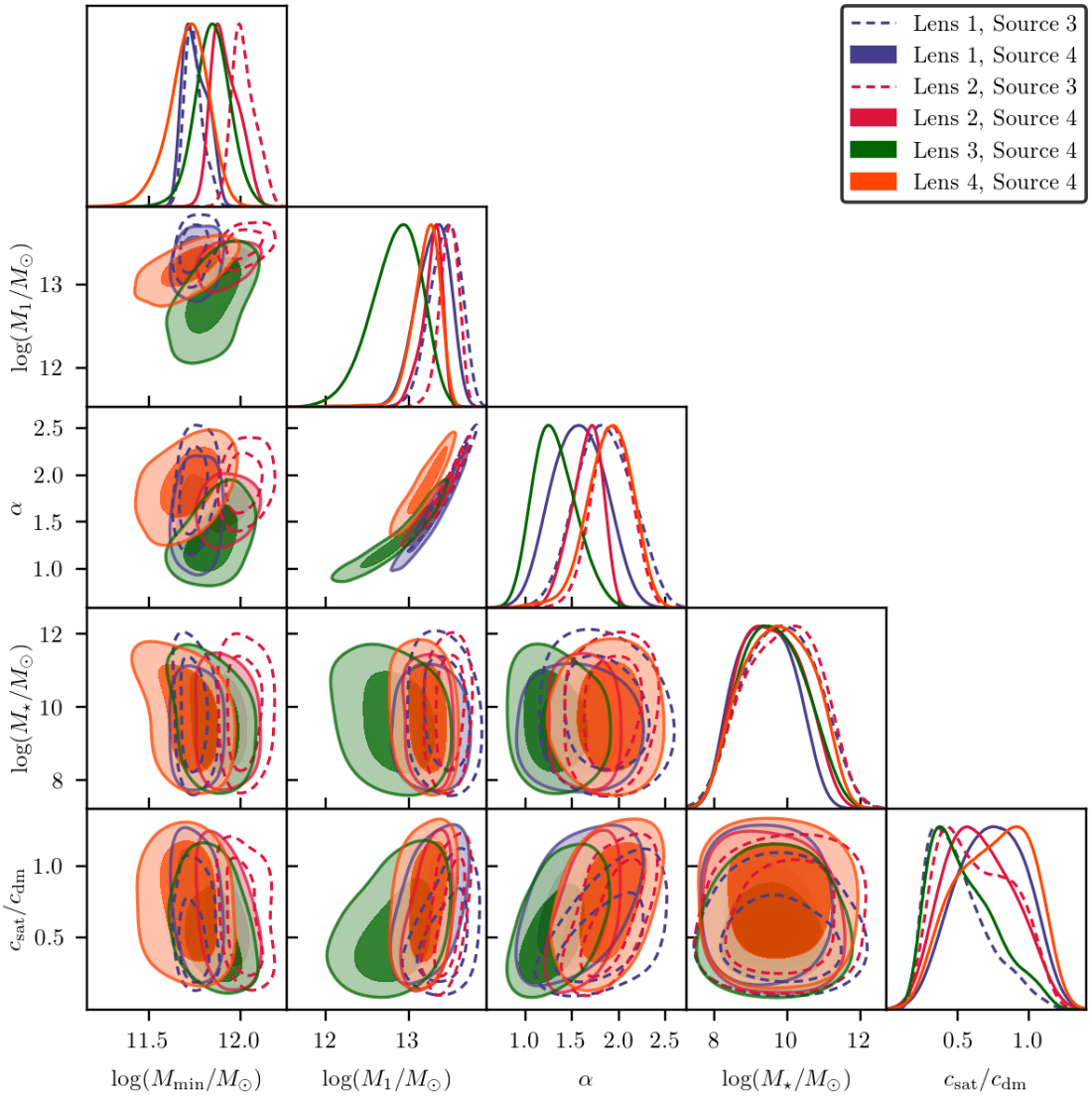


Figure 3.7: Same as Figure 3.6 but for the MagLim sample.

Parameter	**Prior (REDMAGIC)**	**Prior (MAGLIM)**
$\log(M_{\rm min}/M_\odot)$	$\mathcal{U}[11, 13]$	$\mathcal{U}[11, 12.5]$
$\log(M_1/M_\odot)$	$\mathcal{U}[12, 14.5]$	$\mathcal{U}[11.5, 14]$
$\sigma_{\log M}$	$\mathcal{U}[0.01, 0.5]$	$\mathcal{U}[0.01, 0.5]$
$f_{\rm cen}$	$\mathcal{U}[0.0, 0.3]$	–
α	$\mathcal{U}[0.8, 3]$	$\mathcal{U}[0.1, 2.5]$
$\log(M_\star/M_\odot)$	$\mathcal{U}[9, 12]$	$\mathcal{U}[9, 12]$
$a = c_{\rm sat}/c_{\rm dm}$	$\mathcal{U}[0.1, 1.1]$	$\mathcal{U}[0.1, 1.1]$
Δz_ℓ^1	$\mathcal{N}(0.006, 0.004)$	$\mathcal{N}(-0.009, 0.007)$
Δz_ℓ^2	$\mathcal{N}(0.001, 0.003)$	$\mathcal{N}(-0.035, 0.011)$
Δz_ℓ^3	$\mathcal{N}(0.004, 0.003)$	$\mathcal{N}(-0.005, 0.006)$
Δz_ℓ^4	$\mathcal{N}(-0.002, 0.005)$	$\mathcal{N}(-0.007, 0.006)$
Δz_s^3	$\mathcal{N}(0.0, 0.006)$	$\mathcal{N}(0.0, 0.006)$
Δz_s^4	$\mathcal{N}(0.0, 0.013)$	$\mathcal{N}(0.0, 0.013)$
m^3	$\mathcal{N}(-0.0255, 0.0085)$	$\mathcal{N}(-0.0255, 0.0085)$
m^4	$\mathcal{N}(-0.0322, 0.0118)$	$\mathcal{N}(-0.0322, 0.0118)$
Σ_ℓ^1	–	$\mathcal{N}(0.975, 0.062)$
Σ_ℓ^2	–	$\mathcal{N}(1.306, 0.093)$
Σ_ℓ^3	–	$\mathcal{N}(0.870, 0.054)$
Σ_ℓ^4	–	$\mathcal{N}(0.918, 0.051)$
$\alpha_{\rm sat}$	$\mathcal{U}[0, 0.2]$	–

Table 3.1: Priors on model and uncertainty parameters. If the prior is flat we present its range, while for Gaussian priors we list the mean and variance.

Section 3.7.1 the model fits to the data and the parameter constraints. We then show in Section 3.7.2 several derived quantities from our model fits: the average halo mass, galaxy bias and satellite fraction for our samples. We compare these quantities with literature as well as estimations using only the large, cosmological scales. Finally in Section 3.7.3 we perform a series of tests to demonstrate the robustness of our results to various analysis choices.

3.7.1 Model fits

Best-fit models for all the lens-source redshift bin combinations for the redMaGiC and MagLim lens samples are shown in Figures 3.4 and 3.5 respectively, with the χ^2 of the fits and the corresponding number of degrees of freedom listed on the plots. We show the decomposition of the different components that contribute to the final model as described in Section 3.3. The parameter constraints are shown in Figures 3.6 and 3.7, respectively. These plots only show the parameters that are constrained by the data. The best-fit parameters are listed in Tables D.1 and D.2.

From Figures 3.4 and 3.5 we observe that our model generally describes the data well between the measured scales of 0.25–250 arcmin. The χ^2 per degree-of-freedom is close to 1 for most bins, with the largest value ~ 2 for redMaGiC bin [Lens 2, Source 4] and MagLim bin [Lens 1, Source 4], and the smallest value ~ 0.5 for redMaGiC bin [Lens 2, Source 3]. We do not consider this very problematic given that there is no apparent trends in the model residuals and that these datasets are much more constraining compared to previous work. Nevertheless, the slightly high χ^2 values could motivate additional modeling improvements beyond this work. We also note that not all the components in our model are contributing significantly to the fit. For a detailed discussion on how different components contribute to the model see Section 3.7.3.

From Figures 3.6 and 3.7, we observe that the mass parameters $M_{\rm min}$ and M_1 are well-

constrained, with $M_{\rm min}$ for the fourth REDMAGIC bin being higher than the first three as a result of the luminosity threshold being higher in that redshift bin. The satellite power-law index parameter α is also constrained mainly by the inclusion of small scales (see discussion in Section 3.7.3). The tight degeneracy between M_1 and α is expected based on Equation (3.2), since a higher normalization M_1 requires a larger α to keep $\alpha_{\rm sat}$ the same, and vice versa. The point-mass parameter, $M_\star$, is not constrained, which means that it is not needed to improve the χ^2 of the fits. This implies that our current model for the mass distribution below the scales we measure (~ 0.25 arcmin) is not significantly different from what the data prefers.

As a side note, we have found that the inclusion of $\bar{n}_g^i$ values in the REDMAGIC data vector (see Section 3.6) constrains the $f_{\rm cen}$ parameter to low values, which indicates that the model prefers a significant number of centrals not being included in our REDMAGIC lens sample by the selection algorithm. Without this additional information, $f_{\rm cen}$ is not constrained[7]. On the other hand, for MAGLIM since $f_{\rm cen} = 1$ we do not see this effect and there is no need to incorporate $\bar{n}_g^i$ into the data vector of that sample.

3.7.2 Halo properties

Given the model fit, we can derive a number of quantities that describe the properties of the halos hosting the lens galaxies. Specifically, we discuss the average lens halo mass as estimated by:

$$\langle M_h \rangle = \frac{1}{\bar{n}_g} \int dM_h \; M_h \frac{dn}{dM_h} \langle N(M_h) \rangle \; , \tag{3.31}$$

the average satellite fraction using Equation (3.9) and the average galaxy bias calculated from Equation (3.7).

7. To understand this we need to look at Equations (3.7) and (3.9) which define the average galaxy bias and satellite fraction, respectively. Since in our HOD parametrization both the expectation number for centrals and satellites (Equations (3.1) and (3.2)) are proportional to $f_{\rm cen}$, and since $\bar{n}_g \propto f_{\rm cen}$ as well, $f_{\rm cen}$ cancels out in $\bar{b}_g$ and $\alpha_{\rm sat}$. It is, therefore, only through $\bar{n}_g$ that we can constrain $f_{\rm cen}$.

Figures 3.8 and 3.9 show the average halo mass (top panel), the average linear galaxy bias (middle panel), and the satellite fraction (bottom panel) for the REDMAGIC and MAGLIM lens samples in the four redshift bins. The points represent the best-fit maximum posterior and the error bars represent the 68% confidence intervals from the MCMC chain. To derive these constraints, we calculate Equations (3.31), (3.7) and (3.9) at each step of our chains to build the distributions of these three quantities and then estimate the reported constraints.

We first focus on REDMAGIC. For the average halo mass, we compare our results with that derived in the DES Science Verification (SV) data in Clampitt et al. (2017). The SV sample is broadly similar to the first three lens bins in terms of the luminosity selection and number density. Note, however, that there are some differences in the lens samples between SV and our three lower redshift bins. In particular, the photometry pipeline and the REDMAGIC code have both been updated since SV, and the redshift bins are not identical. With these differences in mind, our results appear broadly consistent with Clampitt et al. (2017) in the HOD-inferred halo mass, with roughly ~ 2 times tighter error bars on average. We point out, however, that due to adding more free parameters to our model compared to Clampitt et al. (2017), our error bars should not be directly compared. Rather, we should take into account that our error bars would be roughly an additional factor of ~ 1.5 tighter, had we considered the simplified model in Clampitt et al. (2017), as illustrated in Figure E.2.

The halo mass in the first three redshift bins appears to decrease with redshift. A big part of this is the *pseudo-evolution* of halo mass due to the mass definition we use. This effect is also mentioned in Clampitt et al. (2017) and is studied in Diemer et al. (2013). In short, since we use the critical (or mean in our plots and tables) density of the universe at every redshift to define the halo mass, we observe a pseudo-evolution of our mass constraints over redshift as the reference density evolves. According to Diemer et al. (2013), from $z \sim 0.2$ to $z \sim 0.6$ the pseudo-evolution of the $200\rho_m$ mass, namely M_{200m}, corresponds to $\Delta \log(M_{200m}/M_\odot) \sim 0.11$ for a halo of $200\rho_m$ mass $\sim 10^{13.8}\ M_\odot$ at $z = 0$. This can

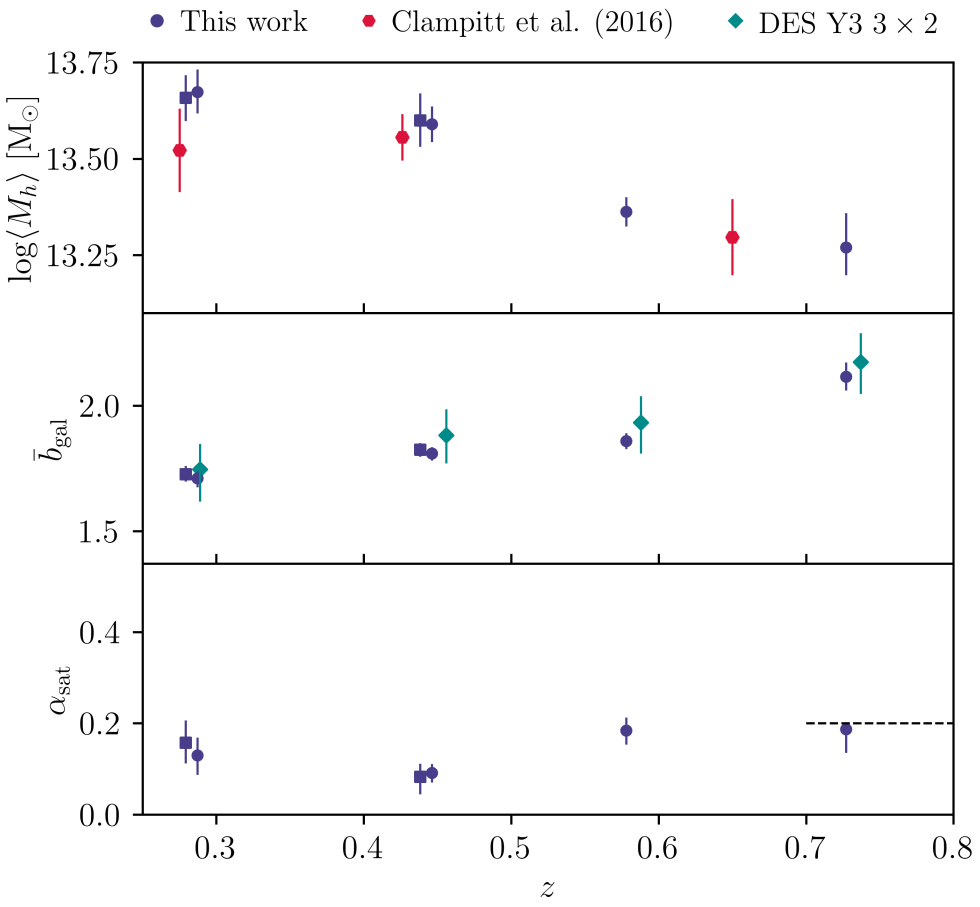


Figure 3.8: Redshift evolution of REDMAGIC properties. Bin combinations with the same lenses but different sources are shown in different markers (square for source bin 3 and circle for source bin 4) and a small offset of 0.005 between the two has been applied in the horizontal axis to make the plot easier to read. As we discuss in Section 3.3, these results assume the de-correlation parameter $X_{\rm lens} = 1$. *Top panel:* The average halo mass, compared with results from Clampitt et al. (2017) (red pentagon). *Middle panel:* The average galaxy bias, compared to constraints from DES Collaboration (2021) (cyan diamond). *Bottom panel:* The average satellite fraction; the dashed horizontal line shows the prior on $\alpha_{\rm sat}$ applied to the last redshift bin.

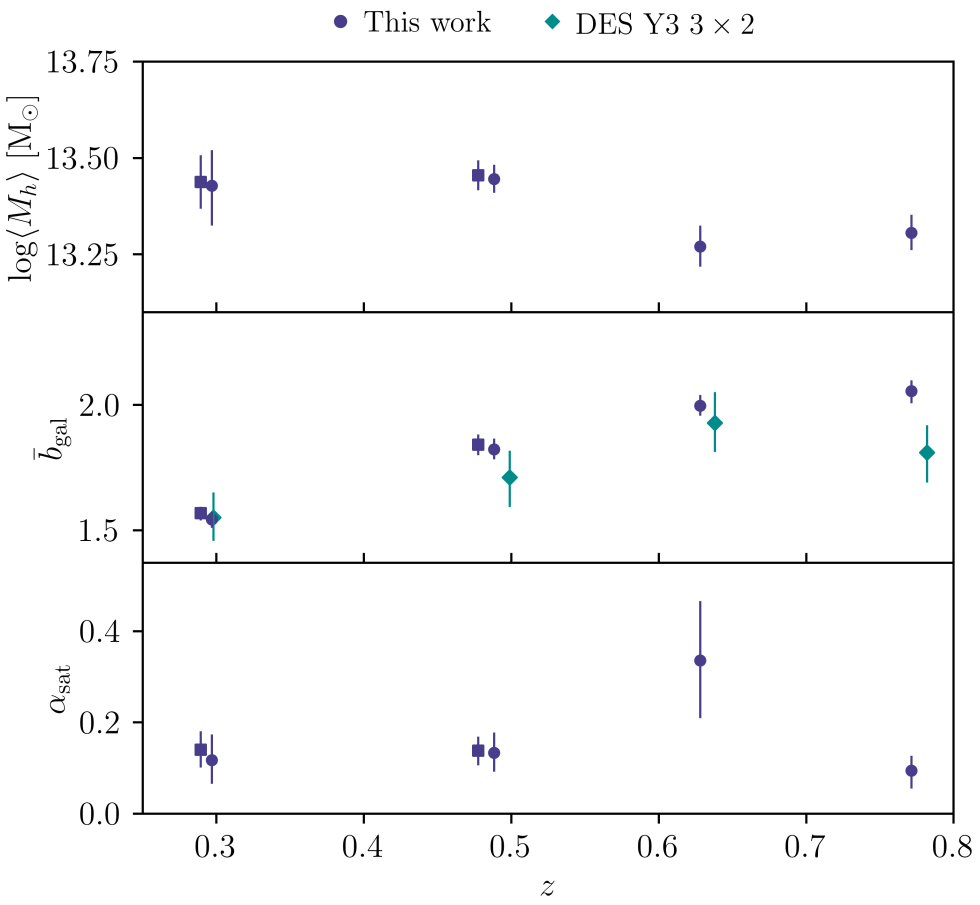


Figure 3.9: Same as Figure 3.8 but for the MagLim sample.

account for most of the difference between the first two bins and the third one. Therefore, we do not find significant change in mass beyond this pseudo-evolution. For the last redshift bin, in addition to the pseudo-evolution in mass, we note that the sample is more luminous (see Section 3.4.1) compared to the first three bins and thus we are looking at more massive halos, which acts opposite to the trend from the pseudo-evolution. We point out here that the overall trend we observe in redshift for the mass is consistent with that seen in simulations (see Appendix A.2). As a further test, we note that we have roughly calculated the ratio of halo mass to stellar mass for the redMaGiC sample and found it to be a few $\times 10^2$. This result is reasonable for $\sim 3 \times 10^{13}$ $M_\odot$-mass galaxies, based on stellar-to-halo mass relation constraints (for a review see Wechsler & Tinker (2018)).

For the average galaxy bias we first compare our results with constraints from large-scale cosmology for the same sample presented in DES Collaboration (2021). The large-scale constraints come from combining galaxy-galaxy lensing and two other two-point functions (galaxy density-galaxy density correlation and shear-shear correlation) to form the so-called 3×2pt probes, so they are not expected to agree trivially. We find that the DES Y3 3×2pt constraints on galaxy bias is quite consistent with our HOD-inferred galaxy bias. The main additional information that our HOD analysis adds to the picture here is the small-scale information, which is consistent with the large-scale information in galaxy-galaxy lensing only (see cyan points in Figure 3.8) – as we will show later in Section 3.7.3, most of the constraining power comes from the 1-halo regime and our galaxy bias constraints does not change whether or not we include the large cosmological scales. The small-scale constraints are tighter than the large-scale only constraints by a factor of roughly 5. In particular, we note that the main improvement is not coming from the increased signal to noise. Rather, it is the wealth of information in the 1-halo regime that improves the constraints. The higher galaxy bias measured for the last redshift bin, compared to the first three bins, is mainly a result of the different selection criteria. We remind the reader here that the galaxies which form the last bin are selected using a higher luminosity threshold, as discussed in Section 3.4.1.

For the satellite fraction, we find that our redMaGiC sample prefers a low (~ 0.2) satellite fraction in all redshift bins we consider. We note that this trend and the values appear quite different from that observed in the MICE simulations (see Appendix A.2). They are, however, in good agreement with the high-resolution Buzzard simulations (discussed also in Appendix A.2) which show an average satellite fraction of redMaGiC which is ~ 0.2 in all three bins. When looking at a red galaxy sample that is likely to share characteristics with redMaGiC, Velander et al. (2013) constrained the satellite fraction to be small and decreasing with redshift to ~ 0.2 or less, which broadly confirms that our constraints on the

REDMAGIC satellite fraction appear reasonable.

As we have discussed in Section 3.3, throughout our analysis we assume the de-correlation parameter $X_{\rm lens} = 1$. If we were to use the best-fit value of $X_{\rm lens} \approx 0.877$ from the 3×2pt analysis with free $X_{\rm lens}$ our constraints would change. Specifically, given that the galaxy-galaxy lensing signal's amplitude, being multiplied by $X_{\rm lens}$, would decrease, our bias constraints would increase by $\sim 10\%$. This would also increase the average lens halo mass by the same factor, and our satellite fractions would increase too as a result. Given our little understanding of what is causing the inconsistency between clustering and galaxy-galaxy lensing in REDMAGIC we choose to keep $X_{\rm lens}$ fixed to 1 and have these results being our fiducial. Further investigating this issue is out of the scope of this paper.

Next we turn our attention to the MAGLIM sample. By construction, the MAGLIM sample is designed to be close to a luminosity-selected sample, while maximizing the cosmological constraints when using it as lenses in galaxy clustering and galaxy-galaxy lensing. Compared to REDMAGIC, this sample does not include additional selection on color or photometric redshift. On the other hand, since it is not exactly a luminosity selection, the physical interpretation of the redshift trends of this sample is not straightforward. There is also no previous literature for comparison.

As shown in Figure 3.9, we find the average halo mass of the MAGLIM sample to be on average lower than that of REDMAGIC, with the lower two redshift bins appear more massive than the higher redshift bins by $\sim 30\%$. Contrary to intuition, the uncertainties on the halo masses are larger compared to REDMAGIC even though the error bars on the measurements are ~ 4 times smaller. This is because the priors in the nuisance parameters for MAGLIM is larger than that of REDMAGIC – this trend has also been seen in DES Collaboration (2021). The galaxy bias appears quite similar to that of REDMAGIC, with the first and last bins somewhat lower. Compared to the 3x2pt constraints we find overall good agreement with our results, with the last bin having a slightly higher bias in our HOD fits. Finally, we find

the satellite fraction for the MAGLIM sample to be $\sim 0.1 - 0.2$ for all bins, except for the third one which is significantly higher at ~ 0.35 and not as well-constrained.

Overall, we also observe that for bin combinations that share the same lens bin, the derived halo properties are consistent when using different source bins. This is assuring and a useful check that our model is indeed capturing properties of the lens samples instead of fitting systematic effects.

3.7.3 Robustness tests

In this section we study the robustness of our results to a number of analysis choices: cosmology, scale cuts, parameter priors, and the addition of higher-order model components. In particular, we are interested in how the average lens halo mass $\langle M_h \rangle$, average galaxy bias $\bar{b}_{\rm gal}$ and average satellite fraction $\alpha_{\rm sat}$ change under the different analysis choices. We show all the tests in this section for REDMAGIC only, but we expect similar results with the MAGLIM sample.

Robustness to cosmology

In this paper we present our main results assuming a specific fixed cosmology, namely our fiducial cosmological values introduced in Section 3.3. We study here the sensitivity to this assumption. The top panel of Figure 3.10 shows how our results change when two alternative assumptions for cosmology: (1) best-fit ΛCDM parameters from *Planck 2018* (Planck Collaboration, 2020) (2) freeing σ_8.

The average mass of REDMAGIC galaxies and the fraction of satellite galaxies are robust to changing the cosmological parameters to *Planck 2018*. Given that these quantities are best constrained by the small-scale information (the points below the 1-halo to 2-halo transition), this implies that varying the cosmology, to a small degree with respect to our fiducial one, leaves the 1-halo central model prediction almost unchanged. We remind the reader here that

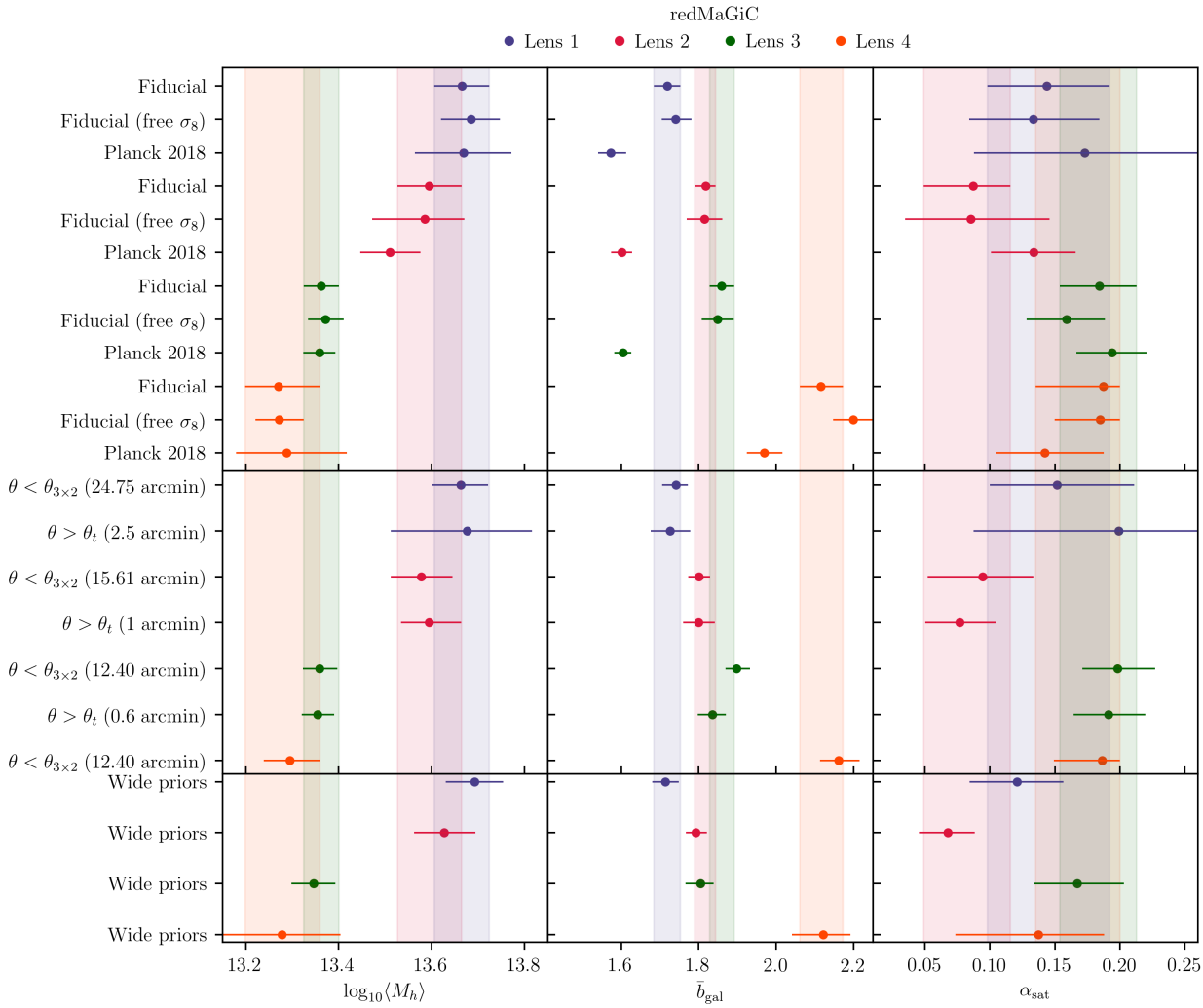


Figure 3.10: Testing the robustness of the halo properties for different cosmologies (*upper panels*), to applying angular scale cuts (*middle panels*), and to changing the prior range on our parameters (*lower panels*) on the REDMAGIC sample . The vertical bands correspond to the fiducial constraints and we added them for an easier comparison with the rest of our results. Note that, to reduce the size of this figure we have combined bins with the same lenses and different sources by presenting the mean of the best-fit values and, to be conservative, the maximum of the error bars.

our fiducial cosmology is similar to *Planck* with the difference that σ_8 we use is slightly lower and our Ω_m is slightly higher compared to *Planck*. The average galaxy bias, on the other hand, is degenerate with σ_8 on the large scales. This means that changing to the *Planck 2018* cosmology directly changes the inferred galaxy bias as seen in Figure 3.10 – using the *Planck 2018* cosmology with a higher σ_8 value results in lower values for the galaxy bias.

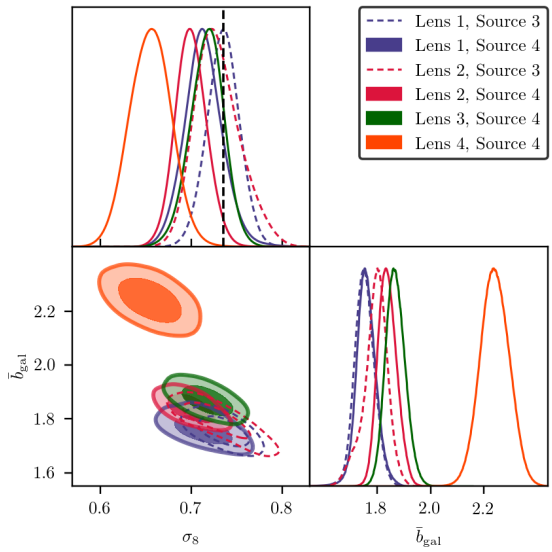


Figure 3.11: Robustness to freeing σ_8 for REDMAGIC galaxies. We present the joint constraints on σ_8 and the derived average galaxy bias for all redshift bins we consider. The vertical dashed line shows the fixed value of σ_8 used in our fiducial cosmology.

Next, we allow for σ_8 to freely vary within the prior range $[0.4, 1.2]$, fixing all other cosmological parameters to our fiducial cosmology. Figure 3.11 presents our results for the σ_8 and galaxy bias constraints from this test for the REDMAGIC galaxy sample. In addition, we have compared the average halo mass, galaxy bias and satellite fraction from these chains in Figure 3.10 to the fiducial results. As we can see, our constraints on σ_8 from the first three lens bins recover the fiducial value of σ_8 quite well. The last bin prefers a lower value of σ_8, and a slightly higher galaxy bias – these are still consistent within 1σ though. Overall, the constraints on all these quantities remain consistent with our fiducial ones. We can, therefore, conclude that freeing the matter power spectrum's amplitude does not alter our constraints in a meaningful way.

Angular scale cuts

Next we study how removing data points on different scales from the fits affects our results. For these tests we first cut out small scales by setting the minimum θ to the *threshold* values $\theta_t = \{2.5, 1, 0.6, -\}$ arcmin for each lens redshift bin, after which we find the data is not constraining enough and this leads to nonphysical constraints and projection effects[8]. This happens because using only the $\theta > \theta_t$ scales in our fits the total central component of γ_t, namely $\gamma_t^{\rm cen}$, becomes identical to $\gamma_t^{\rm sat}$, the total satellite term. These two are then identical to the total shear and, therefore, the fit cannot distinguish between the two. This means the satellite fraction cannot be determined accurately and the other two halo properties suffer too as a result.

To determine the maximum scale cut we can use in each redshift bin without being dominated by projection effects we perform the following analysis using simulated data vectors.

The simulated data vectors are produced with our model using parameters that correspond to the best-fit maximum posterior values from our fiducial runs on the REDMAGIC data, as they are presented in Section 3.7.1. We first fit all angular scales and confirmed our pipeline can recover the input. Next, we remove data points from the smallest scales and repeat the fitting and analysis. We then compared both the constraints on the model parameters and the inferred halo properties from all these runs with different scale cuts. From this comparison we were able to identify the scale cut with the maximum $\theta_{\rm min}$ which was still able to give us results consistent with the full-scale simulated-data runs. At high redshift the threshold θ was found to be lower since the same angular scale corresponds to higher physical scale. This is especially evident in the last redshift bin where we cannot remove any of the scales since they are all needed to constrain the HOD parameters, and

8. Projection effect here means that when we project a multidimensional parameter space to the one-dimensional posterior distributions sometimes the constraints could appear biased.

it even requires the additional prior on the satellite fraction, as discussed in Section 3.6, in order to keep $\alpha_{\rm sat}$ under control.

We also test the case where we remove scales used in the cosmological analysis, derived in Krause et al. (2021b), which we refer to as *cosmological scales* and we denote by $\theta_{3\times2}$. Since small scales are expected to provide most of the constraining power, we put that to test by comparing our constraints from fitting only the small scales, excluding the cosmology scales.

The middle panels of Figure 3.10 present our results for the derived halo properties from applying the above angular scale cuts on REDMAGIC data. For comparison, in the same plots we have included the vertical bands that correspond to the fiducial chains which use the full range of angular scales. As we can see, using the scale cuts discussed above, all our results stay consistent with our fiducial constraints. In addition to this, we can see that the small scales-only fits are also consistent with all other points. Furthermore, these fits, despite using fewer points, can constrain all halo properties almost as well as the full-scale runs, showcasing the rich information contained in the small scales.

Effect of the priors

In our main analysis we have performed various tests on how and whether the priors on our model parameters can have an impact on our results. Here we demonstrate that our parameter priors are not too restrictive and informing the constrained parameters. For our tests in this section, we test the sensitivity of our results when we use roughly 2 times wider priors than that used in the fiducial analysis for all model parameters, keeping the prior center the same.

The bottom row of panels of Figure 3.10 shows the inferred halo property constraints with the widened priors compared to the fiducial, for the REDMAGIC sample. We see that the derived parameters appear consistent. We note here, however, that during our tests we

found that small shifts in the best-fit points can occur if the prior range changes or if it is kept the same but the sampler starts at a different position in parameter space. These effects are not significant, though, in our runs and thus our results stay robust, as discussed above.

Model complexity

In Section 3.3 we described the details of the various model components. In this section, we explain the process we have used to decide whether or not a component has been included in our fiducial model based on how each of them affects the fits and the inferred halo properties.

Our fiducial framework starts with the basic HOD modeling where γ_t is composed of the following four terms: the 1-halo central and satellite contributions $\gamma_t^{\rm c1h}$ and $\gamma_t^{\rm s1h}$, respectively, and their 2-halo counterparts $\gamma_t^{\rm c2h}$ and $\gamma_t^{\rm s2h}$. We will refer to the combination of these four components as the *HOD-only* model. As a first step we like to see if HOD-only can describe our data well. For the six bin combinations, we find that the HOD-only model achieves reduced χ^2 of $\{0.585, 1.144, 1.019, 2.101, 0.879, 1.119\}$. These fits are already good, but there is room for improvement on bin [Lens 2, Source 4] which has noticeably the worst χ^2. Our fiducial model improves the reduced χ^2 over the HOD-only model by $\{0.055, 0.094, 0.023, 0.030, 0.066, 0.181\}$ for reMaGiC.

The procedure we use to determine our fiducial framework is discussed in detail in Appendix E and goes as follows: Using the HOD-only model as a baseline we systematically include additional components and test whether the fits to the data improve, by calculating and comparing the reduced χ^2 of the corresponding data fits. In addition to a change in the reduced χ^2, we also check in each case if the inferred halo properties change significantly as a result of adding a contribution to γ_t. This step is intended to check if omitting a term would introduce a bias in our constraints. Finally, we consider whether it makes physical sense to include a component. If a component is physically well-motivated, we may decide to keep it even if it does not significantly improve the fit. On the other hand, if a contribution is not

well motivated and its modeling is uncertain, we may decide to discard it even if it makes a difference in the goodness-of-fit.

From Appendix E we decide to include the following additional modeling components to γ_t from the HOD-only model: (1) Point-mass contribution; (2) Tidal stripping of the satellites; (3) A concentration parameter for the satellites which is different from that of the dark matter's distribution; (4) Magnification of the lenses. This is the fiducial model which we used to derive the main results in Section 3.7.

As a further note, the particular choice of the HOD model itself is another aspect of the full model that can be much more complex than, or different from, what we used in this work as described in Section 3.2.1. To that end, we experimented with various treatments of the galaxy-halo connection and did not find that adding additional parameters to it or modifying its parametrization made a significant difference to our results. Specifically, we have tested the following modifications to our fiducial HOD. We modified the satellite HOD, $\langle N_s(M_h)\rangle$ of Equation (3.2), by multiplying it by an exponential cutoff $\exp(M_h/M_{\rm cut})$, with mass cutoff $M_{\rm cut}$, following, for example, Leauthaud et al. (2011); Zu & Mandelbaum (2015) where the authors expanded the standard HOD to include the stellar mass function in a robust framework to study the galaxy-halo connection. Another similar variance of the HOD model we tested was to modify the satellite terms by replacing $(M_h/M_1)^\alpha$ by $[(M_h - M_0)/M_1]^\alpha$, as in Guo et al. (2016) for instance where the HOD was compared to subhalo matching in order to determine which describes better the clustering statistics in SDSS DR7, where we introduce the additional mass cutoff parameter M_0, setting $\langle N_s(M_h)\rangle$ to zero if $M_h/M_0 < 1$. We, furthermore, tested altering the satellite term by not multiplying $\langle N_s(M_h)\rangle$ by $f_{\rm cen}$, considering this parameter only through $\langle N_s(M_h)\rangle$, as in Clampitt et al. (2017). Finally, we modified our model by decoupling the satellites from the central galaxies, setting $\langle N_s(M_h)\rangle = (M_h/M_1)^\alpha$, thus not multiplying the satellite term by the number of central galaxies. These variants of the HOD framework we tested did not significantly alter

our results.

We also compare our HOD modeling choices to previous literature. For instance, Clampitt et al. (2017), which performed an HOD study on REDMAGIC galaxies from the DES SV data, used a basic HOD model that was sufficient to fit their data, given that their statistical uncertainties were much larger compared to this work and the range of scales used was narrower. In another study, Velander et al. (2013) used 154 $\deg^2$ of CFHTLenS lensing data, splitting galaxies into blue and red, and considered a more complex model where they included the effects from baryons as a point-mass source and satellite stripping, similarly to our work, although they did not use the full five-parameter HOD model we employ here but rather one similar to Mandelbaum et al. (2005) that fixes the satellite power-law index. Therefore, compared to both Velander et al. (2013) and Clampitt et al. (2017) we have used a more complex model which, although increased our error bars on the parameter constrains, was required to capture the features of our more constraining data. In addition to that, we have taken into account systematic uncertainties by introducing the Δz^i and m^i parameters (discussed in Section 3.6) which further increased our error bars.

3.8 Summary and Discussion

In this work, we have carried out a detailed analysis on modelling the small-scale galaxy-galaxy lensing measurements for the two lens samples REDMAGIC and MAGLIM using a Halo Occupation Distribution (HOD) framework. Our lens samples were divided into four tomographic bins each spanning a redshift range about 0.2–0.9. In this work we have extended the measurements in Prat et al. (2022) to smaller scales, totalling 30 logarithmic bins in angular scales from 0.25 to 250 arcmin (physical scales from ~ 70 kpc in the lowest redshift bin to ~ 110 Mpc in the highest redshift bin). Our main findings are:

- These measurements increase the signal-to-noise of our measurements by a factor of 2-3 compared to the signal-to-noise from scales used by cosmology analyses.

- We constrain the average halo mass of our REDMAGIC (MAGLIM) sample to $\sim 10^{13.6}$ $M_\odot$ ($10^{13.4}$ $M_\odot$) in the lowest redshift bin and $\sim$ $10^{13.3}$ $M_\odot$ ($10^{13.3}$ $M_\odot$) at the highest redshift bin. The uncertainty on these mass constraints are about $\sim 15\%$. The REDMAGIC constraints are consistent with previous work in Clampitt et al. (2017). The halo masses of MAGLIM are overall lower compared to REDMAGIC, especially at lower redshift.

- We constrain the average linear galaxy bias for the REDMAGIC (MAGLIM) sample to be ~ 1.7 (1.5) at low redshift and ~ 2.1 (2) at high redshift. Our results are consistent with those inferred only from the large scales from DES Collaboration (2021), but with about 5 times smaller uncertainties due to the small-scale information.

- We constrain the satellite fraction for the REDMAGIC (MAGLIM) sample to be $0.1-0.2$ $(0.1 - 0.3)$ with no clear redshift trend. Our REDMAGIC results appear to be in agreement with other studies which measured the satellite fraction of red galaxies, e.g. in Velander et al. (2013). Our results for MAGLIM, which consists of a more wide variety of galaxies than REDMAGIC, also appear reasonable and in agreement with studies like Mandelbaum et al. (2006b); Coupon et al. (2012); Velander et al. (2013). In these studies, the authors concluded that the fraction of satellite galaxies is reducing with increasing halo mass and that $\alpha_{\rm sat}$ is roughly what our constraints point to.

Motivated by the increased signal-to-noise, we consider additional model complexity on top of the basic HOD framework: a point-mass component, stripping of the satellites of their outer dark matter, magnification of the lenses, and modifying the spatial distribution of the satellite galaxies by varying its concentration parameter with respect to the distribution of dark matter in the lens halos. Using this model we were able to obtain good fits to the measurements over all angular scales and for all redshift bins we considered. We note that two out of twelve bin combinations show a best-fit χ^2 per degree-of-freedom ~ 2, which

could motivate additional modeling developments for the future, or indicate some residual systematic effect that is not well understood.

To further test our analysis, we have preformed various tests where we vary parts of our modeling and fitting procedure to make sure that our results remain robust under small changes around the fiducial framework. We tested the sensitivity of our results to the assumption of cosmology, the angular scales used in the model fit, and the width of our priors – we find that our results are robust to these changes.

There are a number of limitations in our analyses that we point out here for the readers to appropriately interpret our findings. First, in Appendix A.2 we showed a series of tests that we performed using available simulations. However, the resolution of these simulations were insufficient for us to conclusively validate our model and methodology on scales deep in the 1-halo regime. That is, it is plausible that our fiducial model, although well-fitted to the data, is not the true description of the galaxy-halo connection. Higher resolution simulations exist (Nelson et al., 2019, Illustris TNG), but the simulation volume is much smaller and to exactly match our sample we would require running the redMaPPer algorithm on the simulations. We point these out both as caveats for interpreting our results and as inspirations for future studies. The second element that would benefit from future advances is the modeling of the covariance matrix. An analytic covariance model on this large range of scales is possible to calculate, but there are differences in the halo-model assumptions and HOD parametrization between the existing covariance modeling (Friedrich et al., 2020) and our assumptions. Furthermore, we would need to find a sophisticated way of treating the HOD in the analytic covariance calculations, given that the HOD is what we are constraining in our analysis and we thus should avoid the resulting circularity. As a result, we have adopted a data-based Jackknife covariance, which has its own issues of being noisy and often overestimated (Friedrich et al., 2016). This is an area of active research and it would be interesting for future studies to re-analyze this data using a more advanced covariance

model. Finally, as we mention in Section 3.3.6, in our analysis an accurate and tested model for IA is missing in the 1-halo regime. Therefore, although we found that our simple IA model to contribute insignificantly in our analysis (see relevant discussion in Section 3.7.3), it is plausible that a more accurate IA model could have a larger effect on the full model fit. This again can serve as a starting point for exploration of better IA models in the 1-halo regime, now that our data is becoming sufficiently constraining.

In this work we established a framework to systematically explore a number of modeling choices in the galaxy-galaxy lensing signal from deep in the 1-halo regime to the cosmological 2-halo regime. Many of these effects were ignored in earlier work as the statistical uncertainties were large relative to these effect. In the final DES Y6 dataset we expect 1.5-2 times more source galaxies and a reach to higher redshift for the lens sample, which will allow us to further test the different model components. What we learn will feed into future analyses with the Rubin Observatory's Legacy Survey of Space and Time, the Nancy Roman Space Telescope and the ESA's Euclid mission. We expect these future datasets to be qualitatively different in terms of data quantity and quality, and a combination of modeling techniques (HOD models like what we studied here, hydrodynamical simulations and emulator approaches) will be needed to understand how galaxies and dark matter halos are connected at the very small scales.

CHAPTER 4

DES Y3 STELLAR MASS SAMPLE

In this chapter we present an updated and improved model of the galaxy-halo connection compared to Chapter 3. The new model incorporates small-scale modeling of the two-point statistics of galaxy clustering and it also accounts for the stellar-mass information of the galaxies. This work is based on the upcoming publication Zacharegkas et al. (2023), in prep that is scheduled to be released in the near future.

4.1 Introduction

The connection between galaxies and dark matter is one of the most fundamental aspects of modern Cosmology. It is thus essential to understand this connection in order to form a comprehensive interpretation of the observed Universe. In the era of precision cosmology, analyses of Large-scale Structure (LSS) from galaxy surveys can lead to biased results if the details of the galaxy-halo connection are ignored (McDonald & Roy, 2009; Baldauf et al., 2012; Krause et al., 2017). To avoid modeling the highly complicated physics on the smallest scales, those data points are usually ignored and only the data in the linear to quasilinear regime, where a simple prescription of the galaxy-halo connection such as *linear galaxy bias* is sufficient (see e.g. DES Collaboration, 2021), are considered. In some cases more complicated galaxy bias models on small scales (such as Heymans et al., 2021) are used and marginalise over the model parameters. For either approach, a data-driven model of the galaxy-halo connection on scales below a few Mpc could allow us to significantly improve the cosmological constraints achievable by a given dataset. Galaxy bias on small scales is non-linear by nature (Dvornik et al., 2018) and accurate galaxy-halo connection models provide a wealth of crucial information. In addition, understanding the connection between different galaxy samples and their host halos also has implications for galaxy evolution (see Wechsler

& Tinker, 2018, for a review of studies for galaxy-halo connection).

One of the most powerful probe of the galaxy-halo connection is *galaxy-galaxy lensing* which directly probes the dark matter halos of galaxies (Tyson et al., 1984; McKay et al., 2001; Hoekstra et al., 2004; Sheldon et al., 2004; Mandelbaum et al., 2005, 2006a; Seljak et al., 2005; Johnston et al., 2007; Heymans et al., 2006; Leauthaud et al., 2011). Galaxy-galaxy lensing refers to the measurement of the cross-correlation between the positions of foreground galaxies and shapes of background galaxies. Due to gravitational lensing, the images of background galaxies appear distorted due to the deflection of light as it passes by foreground galaxies and the dark matter halos they are in. As a result, this measurement effectively maps the average mass profile of the dark matter halos hosting the foreground galaxy sample. A common approach to modeling this measurement is via the *Halo Model* (Seljak, 2000; Ma & Fry, 2000; Cooray & Sheth, 2002) and the *Halo Occupation Distribution* (HOD) framework (Peacock & Smith, 2000b; Scoccimarro et al., 2001; Berlind & Weinberg, 2002; Tinker et al., 2007; Zheng et al., 2007; Zehavi et al., 2011; White et al., 2011). In this framework, a large luminous *central galaxy* is considered to exist in the centers of the host halos and the smaller *satellite galaxies* are distributed within the halo (Kravtsov et al., 2004).

Another powerful probe of the galaxy-halo connection is *galaxy clustering*, which refers to the measure of spatial clustering of galaxies. However, small-scale galaxy clustering alone cannot yield cosmological constraints because unless combined with probes that are sensitive to the mass scales of dark matter halos, such as galaxy-galaxy lensing (van den Bosch et al., 2003; Tinker et al., 2005; Seljak et al., 2005). Combining the two, has been proved (Yoo et al., 2006; Cacciato et al., 2009) to be sensitive to cosmological information, and particularly to σ_8 and Ω_m, since this combination probes the shape and amplitude of the halo mass function on small scales and the overall matter density and the bias of the galaxy sample at large scales.

Various flavors of the HOD framework have been proposed in the literature. For example, the *conditional luminosity function* $\Phi(L|M_h)$ that provides the average number of galaxies of luminosity $L \pm dL$ that reside in a halo of mass M_h (Tasitsiomi et al., 2004; Yang et al., 2004; Van Den Bosch et al., 2004; Vale & Ostriker, 2006; Cooray, 2006) is one, or similarly the *conditional stellar mass function* $\Phi(M_\star|M_h)dM_\star$ which describes the average number density of galaxies in a halo of mass M_h with stellar masses in the range of $M_\star \pm dM_\star$ (Yang et al., 2007; Moster et al., 2010; Behroozi et al., 2013). Moreover, *abundance matching*, i.e. assuming that there is there is a monotonic correspondence between halo mass (or circular velocity) and galaxy stellar mass (or luminosity) (e.g., Conroy et al., 2006; Kravtsov et al., 2004; Vale & Ostriker, 2006; Conroy & Wechsler, 2009; Drory et al., 2009; Moster et al., 2010; Behroozi et al., 2013; Guo et al., 2010) has been used to constrain the galaxy *stellar mass function*. However, even though this technique is easy to use in practice, prior knowledge about the mass distribution of halos from cosmological N-body simulations is needed and the assumption that field halos and subhalos of the same halo mass contain galaxies of the same stellar mass.

In this chapter we develop a framework based on the HOD to model small scales in galaxy-galaxy lensing and galaxy clustering. We do so by utilizing an HOD model that is mostly based on our previous work presented in Zacharegkas et al. (2022) for lensing. In that paper we analyzed the two lens galaxy samples that were used by the Dark Energy Survey (DEC) Collaboration Year 3 (Y3) for the main cosmological studies (e.g. Abbott et al., 2022; Prat et al., 2022). Modeling the non-linear scales, we were able to constrain the galaxy-halo connection of both lens samples and we quoted constraints on the average halo mass, galaxy bias and satellite fraction. Our HOD model was widely based on Zheng et al. (2007); Zehavi et al. (2011) and the halo occupation was predicted as a function of the halo mass. In this work we expand upon this framework in two main ways: 1) We also incorporate theory predictions for galaxy clustering on small scales; 2) We develop a

small-scale model that also accounts for the stellar mass of the lens galaxies. The latter is done by direct parametrization of the HOD, but is equivalent to using, e.g. the conditional stellar mass function mentioned before.

We apply our model to data from DES Y3. Specifically, we develop a new lens sample that includes the galaxy stellar mass and color. Moreover, our catalog provides the full two-dimensional joint distribution of redshift and stellar mass for each galaxy, which is produced using a Machine Learning approach. Our model is constructed such that it is capable of accounting for the aforementioned full joint redshift-stellar mass distribution. We describe this new catalog in detail in this study. This sample can be further used in future studies in various analyses. Our goal is to fit measurements of galaxy-galaxy lensing and galaxy clustering of galaxies binned in redshift and stellar mass in order to constrain their galaxy-halo connection via the *stellar-to-halo mass* relation which is a key component in our new HOD model.

The structure of the chapter is as follows. In Section 4.2 we introduce the new lens sample by describing in detail the Machine Learning algorithm (training, validation and its products) and we present the binning our the lens galaxies in redshift and stellar mass. In Section 4.3 we discuss the source galaxy sample that we use in the galaxy-galaxy lensing part of the data vector. Section 4.4 describes the measurements of lensing and clustering, the covariance matrix, and also focuses on the systematics tests we performed on our measurements and how they inform our analyses choices such as scale cuts and bin selection. We then proceed to describe in depth our HOD model in Section 4.5. After we do so, we are able to discuss in Section 4.6 how we model the two observables of choice: galaxy-galaxy lensing and galaxy clustering. Equipped with the data vector and model, in Section 4.7 we present the likelihood analysis we follow to fit the measurements and present the results from the Marcov Chain Monde Carlo chains in Section 4.8. We conclude in Section 4.9 by summarizing and discussing the implications of this work.

Throughout this study we fix the Ω_m and σ_8 cosmological parameters to the best-fit for the `MagLim` sample from DES Y3 (Abbott et al., 2022), while we use the best-fit cosmology from *Planck 2018* (Planck Collaboration, 2020) for the rest of them: $\Omega_{\rm m} = 0.339$, $\sigma_8 = 0.733$, $\Omega_{\rm b} = 0.0486$, $H_0 = 67.37$, $n_s = 0.9649$, $\Omega_\nu h^2 = 0.0006$, where h is the Hubble constant in units of 100 km/s/Mpc. Since we consider the Λ-Cold Dark Matter (ΛCDM) cosmological model, we set $w = -1$ for the dark energy equation of state parameter. In addition, all the halo masses use the definition of $M_{200\rm c}$, based on the mass enclosed by radius $R_{200\rm c}$ so that the mean density of a halo is 200 times the critical density at the redshift of the halo. However, our results are valid for any common choice of mass definition.

4.2 Lens galaxies: The stellar-mass catalog

For our lens sample, we want to generate a catalog of galaxies that is a subset of the `Y3 GOLD` with estimates of redshift, stellar mass, and color. For this purpose, we utilize the ML-based software called `GalPro` (Mucesh et al., 2020). Specifically, we train the ML model on a sample of galaxies whose properties (redshift, stellar mass, color) are known – training sample – and then apply this model to the full lens sample from the `Y3 GOLD`. The only cuts we apply to the GOLD catalog are: 1) We select galaxies with i-band magnitude $i < 23$; 2) We select galaxies with `EXTENDED_CLASS`$\geq$2, which provides a good star-galaxy separation criterion to ensure that almost all objects in our sample are galaxies; 3) A lower stellar-mass limit per redshift bin so that our samples are complete. We also apply a depth mask in i-band with $i < 23$ to the sample to remove parts of the footprint that might be susceptible to bad observing conditions.

4.2.1 Constructing the training sample

A big part of the process is then to construct the training sample. We choose to train our model on galaxies in the DES `Deep Fields` (DF, Hartley et al., 2022), matched to an over-

lapping subset of galaxies in the Cosmic Evolution Survey `COSMOS2020` (COSMOS, Weaver et al., 2022). Unfortunately, there are complications in doing so due to some inconsistencies between the two, which we will very briefly discuss here; however, we conclude that we can use this procedure to produce a reliable training sample for our purposes. The main caveat in treating the DES DF as our training sample is that there is a difference in its photometry compared to COSMOS. Therefore, before we produce the training sample we have to adjust accordingly the stellar masses.

To explain why differences in photometry can lead to strong biases in estimating stellar mass and redshift, we first give a rough overview of the nuances of template fitting and matching of DES DF to COSMOS.

In general, we do not have accurate SED templates for almost any galaxy in the Universe. We only have an average SED for some sub-population at some redshift z, but even that does not represent the average of that population at a different redshift. Furthermore, these average SED's are not able to handle the variance in the intrinsic galaxy SED's that come from, e.g., emission line strengths, the stochastic nature of the Inter-Galactic Medium (IGM) absorption or the full variety of dust attenuation in galaxies, which can depend on location within a given galaxy image, as well as orientation and many other things. Instead, we have SED's based on theoretical models, but these are almost definitely inaccurate. However, in order to measure stellar masses, redshifts and other rest-frame physical properties, we need to use synthetic galaxy SED's of this type – either directly or indirectly. And to make things worse, we also have to consider additional factors such as: inaccuracies in filter response curves, imperfect background subtraction, systematic errors in flux zero-point computation, and systematic biases in photometric measurements (blending, crowding, simplified light profile prescriptions, etc), to name a few. As a result, we do not expect to be able to use the reduced χ^2 of a model fit to the photometric data as a reliable goodness of fit indicator directly. Therefore, the redshift PDF's are also called into question and so various tweaks

are used by the photo-z community, some of which are fairly principled, while others are more like tricks, such as:

- Multiply the photometric errors by some factor to compensate for the fact that the most common photometric measurement software, Source Extractor[1] (SExtractor) has no way of knowing the extent to which the noise in the image pixels is correlated, and so tends to underestimate the uncertainty from the background noise.

- Include a systematic a percentage of the object's flux to the flux error in quadrature. This is the correct thing to do if the photometric zero point has a systematic uncertainty, though it is rare that one actually tries to estimate that uncertainty rigorously.

- Include a template error function to capture the variance in galaxy SED's that is not already represented in the template set. This includes things like the IGM absorption and internal dust reddening mentioned before, which is typically much larger in the rest-frame UV than in the rest-frame optical. Because the SED's are scaled in amplitude during fitting, the template error function is implemented as a fraction of the model flux. This is slightly undesirable, but only really impacts high-z galaxies ($z \geq 3$) and is thus not relevant for this work.

In using DF as a training sample, we match galaxies to COSMOS and as a result have individual galaxies each populated by a single position-matched counterpart from the latter catalogue. So, the question now becomes, why do we have to pause to consider the photometric consistency between DES DF and COSMOS? The two main reasons are following:

1. Some objects may have bad fits to their photometry due to blending, image defects, etc, which can be either in the DES DF or COSMOS. In both cases, however, this implies that we will have a mismatch between the training sample's photometry and the physical properties that it represents.

1. https://www.astromatic.net/software/sextractor/

2. While redshift depends almost entirely on the shape of the SED (i.e. observed colours), the stellar mass depends on both the shape and the amplitude of the SED. Roughly, the shape of the SED encodes the redshift and mass-to-light ratio, while from the amplitude we can infer the luminosity information. Note, however, that the luminosity also depends on redshift, i.e. the SED shape. Moreover, while a weak or strong 4000 Å break will usually result in the same redshift (depending on whether it is suitably represented in the SED set), the stellar mass that it implies can be quite different. Thus, a small bias in one photometric band might bias the mass to a greater extent than it does the redshift.

Therefore, we need to be careful in how we match the two catalogs for the purpose of this work in order to avoid large biases in our inferences.

The optical DF and COSMOS data provide independent measurements of the same galaxies. The near-IR (NIR) data are from UltraVISTA[2] in both cases, but are different data releases and have been handled differently for the two catalogs. Therefore, they are not independent but also not exactly the same. The first step is to compute the χ^2 goodness-of-fit, effectively treating the COSMOS photometry as a model, and the DF as our data. Under the assumption that both data sets are well calibrated and the errors correctly estimated, the distribution of χ^2 values should follow the analytic distribution for the appropriate number of degrees of freedom, with the likely addition of a longer tail or some fraction of outliers that indicate objects that have bad photometry. We could then devise a cut to remove them and use the rest for building the training sample, after also masking out bad regions in both catalogs.

However, after computing the χ^2 between the COSMOS $grizJHKs$ and Deep Fields $grizJHKs$ photometry for each galaxy, removing u because we know that the DF $u-$band calibration is poor, and the χ^2 for just the optical set ($griz$), the resulting χ^2 is poor. This

2. https://ultravista.org/

can be due to a the following reasons:

- The contribution of photon flux noise is often neglected in these catalogs because ground-based images are background dominated in terms of uncertainty for the vast majority of sources. This means that the flux errors for bright objects are frequently underestimated.

- The COSMOS catalog does not report systematic zero-point uncertainties. Thus, we have not included a possible zero-point error from the COSMOS catalog.

It is worth highlighting here that COSMOS is not a single catalog, but two catalogs with a total of four different options for the main photometry. The two catalogues were extracted in different ways – one is a fairly standard method, and the other a model fitting method similar to what is used in DES. In addition, the zero-points are calibrated using a photo-z code by fitting templates at true redshift and deriving common factors of flux difference between the catalogue fluxes and those predicted by the SED templates used by the photo-z algorithm. They use two photo-z codes, and apply them both to the two catalogs, resulting in two different sets of zero-point corrections per catalogue. While fiddling with the zero-points in this way is standard, this process shows how misleading it can be.

To address the above issue, we added a systematic flux error to the COSMOS catalog as a percentage of object flux. We can think of this as reverse-engineering the average systematic zero-point error in COSMOS, under the assumption that we've measured our uncertainties correctly in the DF. We found that a 3% error was sufficient to produce a good χ^2 distribution for bright objects. This is a typical value in this sort of investigations, and we thus conclude that for these bright objects the photometry seems consistent under typical assumptions of data quality.

However, for fainter objects the χ^2 distribution remained poor, even after we introduced a parameter to allow the relative SED's to scale in amplitude, as we discuss shortly below, to improve the χ^2 (while accounting for the loss of one degree of freedom in doing so). After

we plotted the magnitude difference between the catalogs against magnitude, we observed a "curved" dependence between the two, which is typical of background level of one of the images being wrong – it impacts faint objects to a far greater degree than bright objects. But the magnitude of the background offset implied by these plots is much larger than we expected or what is measured in the DF.

There are various approaches to addressing this behaviour. The major hypotheses one could test are the following:

1. There is a systematic in the DES data (e.g. over-subtracted background) that makes the flux of a DES object smaller than it should be. The pipeline is much the same for the Wide Survey and Deep Fields, so a systematic such as this would probably affect both data sets.

2. The background is under-estimated and under-subtracted in the COSMOS data leading to a universal background flux on top of which each galaxy resides. Cases like this tend to be rarer than over-subtraction, but still exist in galaxy surveys today. In this case, the COSMOS stellar mass is biased to high values because there is more light than there should be. If we simply tag the Deep Fields objects with this higher stellar mass then we propagate the bias through to the posteriors of the main survey objects.

In both cases the correct thing to do would be to fix the problem with the images and re-test the agreement. However, that would be beyond the scope of this work. Thus, we instead choose to account for the biases by computing a multiplicative factor that best matches the COSMOS and DF photometry to one another, and then adjust the COSMOS stellar mass so that it reflects the amplitude of the Deep Fields SED. A caveat in this method is that any background problem would vary among different bands, and thus distort the SED rather than just globally scale it. However, the most critical bands are *griz* in DES, and given that they are all from the same pipeline and measurement process, any background defects are likely to be similar between bands. We, therefore, ignore any colour dependencies.

A different scenario, too, would be that the model fitting in the photometric measurement process produces a magnitude-dependent bias in one or both of the catalogs. There is no particular reason, however, why we would believe one above the other, and if it amounts to an overall change in the SED amplitude, then the most pragmatic course of action is to aim for internal consistency in so far as that is possible.

After all this, we are still left with the question of whether the SED shape is reasonable enough that we can use the COSMOS masses (after scaling). Keeping in mind that we don't expect a "good" fit from the χ^2 perspective unless we are able to account properly for all the systematic uncertainties involved, we concluded that this is the case and we can thus proceed with constructing our stellar-mass lens catalog. As a final note, for the purposes of this HOD analysis, the posterior distributions in mass and redshift will be broad, and will most likely subsume the remaining COSMOS to DES bias.

4.2.2 The ML model

After the investigations of Section 4.2.1, we can use the DF catalog, matched to COSMOS after re-scaling, as our training sample in `GalPro`, keeping the default basic setting of the algorithm as described in Mucesh et al. (2020), with the goal to apply the ML model to the whole galaxy sample from the `Y3 GOLD` catalog, as mentioned in the beginning of this section. The input quantities that `GalPro` accepts are *luptitudes* (Lupton et al., 1999), μ, and their uncertainties σ_μ. Thus, starting from the *griz* fluxes f_ρ, with $\rho \in \{g, r, i, z\}$, and their corresponding uncertainties σ_ρ, from the GOLD catalog, we define the corresponding luptitudes as:

$$\mu_\rho = \mu_0 - a \sinh^{-1}\left(\frac{f_\rho}{2b_\rho}\right) , \tag{4.1}$$

where $\mu_0 = m_0 - 2.5 \log b$, $a = 2.5 \log e$ and $b_\rho = \sigma_{\mu,\rho}\sqrt{a}$. In these definitions, e is the base of natural logarithms, $m_0 = 2.5 \log f_0$, and f_0 is a normalization that we set to $f_0 = 10^{12}$.

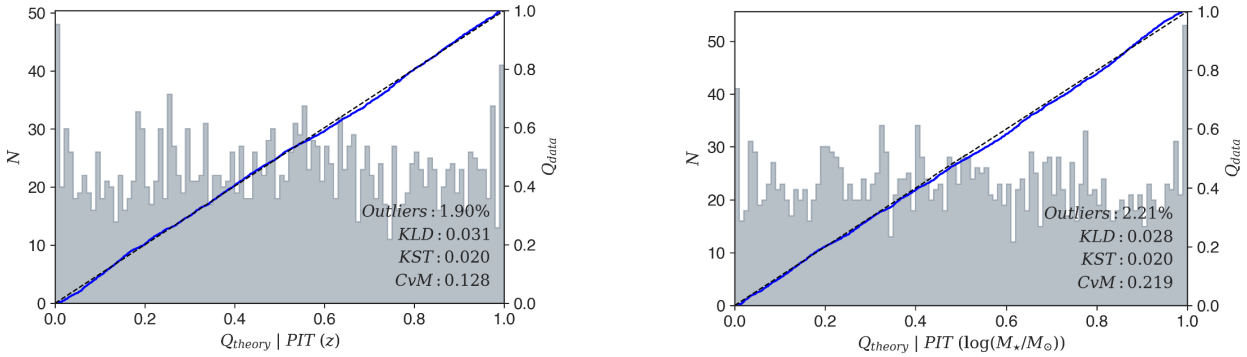


Figure 4.1: Redshift (left) and stellar-mass (right) PIT distributions for the ML model trained to produce the DES Y3 stellar-mass sample. The quantile-quantile (Q–Q) plots are also shown in each plot to highlight deviations from uniformity. The black-dashed and solid blue lines represent the quantiles of the uniform and PIT distributions, respectively. The percentage of catastrophic outliers along with the values of the Kullback–Leibler (KL) divergence, Kolmogorov–Smirnov (KS) test, and Cramér-von Mises (CvM) metrics are also stated to quantify uniformity of the PIT distributions. We define a catastrophic outlier to be any galaxy with a redshift or stellar mass completely outside the support of its marginal PDF. For more on these sort of plots see Mucesh et al. (2020).

We also transform the flux error, σ_ρ, into errors in luptitude by:

$$\sigma_{\mu,\rho} = \sqrt{\frac{a^2\sigma_\rho^2}{4b_\rho^2 + f_\rho^2}} \approx \frac{a\sigma_\rho}{2b_\rho} \,. \tag{4.2}$$

By training the ML model, we obtain for the four bands *griz* the following values of σ: $(\sigma_g, \sigma_r, \sigma_i, \sigma_z) = (30.13635719, 42.84311709, 71.74077268, 124.80833046)$. We note here that, the reason for performing the above transformation is some desired properties of the luptitudes, as described in Lupton et al. (1999); Mucesh et al. (2020). Namely, luptitudes behave like magnitudes for bright photometry and like fluxes for faint photometry, with the turning point in the behaviour determined by the softening parameter. Thus, by converting to luptitudes, we avoid introducing an additional selection effect by not discarding galaxies with negative fluxes. In addition, we need to provide the corresponding $g-r$, $r-i$ and $i-z$ *lupti-colors* and the corresponding errors. These are simply calculated as $\mu_\rho - \mu_\nu$ for two bands $\rho, \nu \in \{g, r, i, z\}$ and $\sqrt{\sigma_{\mu,\rho}^2 + \sigma_{\mu,\nu}^2}$, respectively.

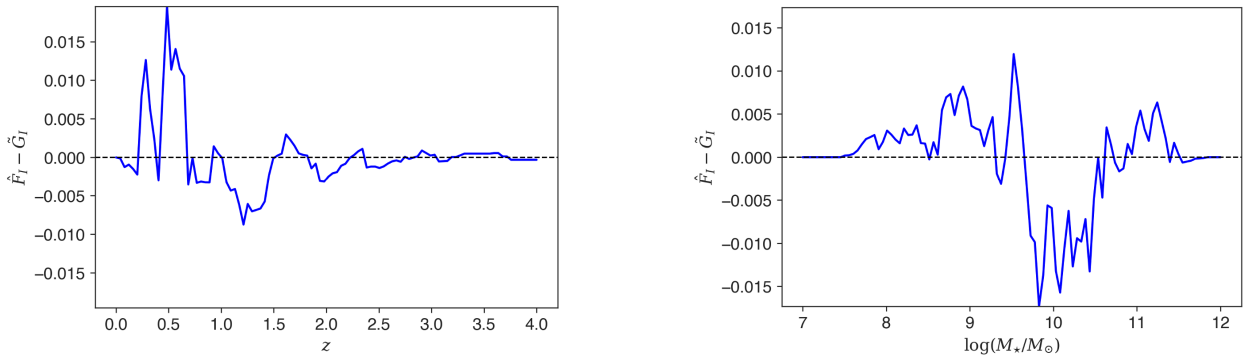


Figure 4.2: The difference between the average predictive CDF ($\hat{F}_I$) and the true empirical CDF ($\tilde{G}_I$) of redshift (left) and stellar-mass (right) plotted at different intervals in their respective ranges, for the ML models used to create the DES Y3 stellar-mass sample. For more on these kinds of plots see Mucesh et al. (2020).

After the model is trained and before we run it on the GOLD catalog, we need to run a few validation tests, based on Gneiting et al. (2007), which we summarize below and are also discussed in Mucesh et al. (2020) in more detail. There are two validation methods that we use in this work: a *probabilistic* and a *marginal* method.

- The probabilistic method to validating the PDF's that `GalPro` produces relies on the *Probability Integral Transform (PIT)*, which is defined as:

$$PIT \equiv \int_{-\infty}^{\tilde{y}} f(y)dy \ , \tag{4.3}$$

where $\tilde{y}$ is the "true" redshift or stellar mass and $f(y)$ is the marginal PDF. If $f(y)$ is probabilistically calibrated, then at the true values $\tilde{y}$ its Cumulative Distributions Function (CDF) should not have a preferred values. In this case, for an ensemble of galaxies, the PIT would follow the *Uniform distribution* $U(0,1)$. Note, however, that although this is necessary condition for the marginal PDF to be valid, a biased distribution could still produce a uniform PIT distribution; this test is thus not sufficient and we need more tests to be performed on the marginal PDF's. The results of this test on our model are shown in figure 4.1. As we can see, the PIT distributions follow

closely the uniform and we thus conclude that our model passes this test.

- The marginal calibration method is comparing the *average predictive CDF*, $\hat{F}_I(y)$, with the *true empirical CDF*, $\tilde{G}_I(y)$, respectively defined as:

$$\hat{F}_I(y) \equiv \frac{1}{n}\sum_{i=1}^{n} F_i(y) \tag{4.4}$$

and

$$\tilde{G}_I(y) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{y}_i \leq y\} \ , \tag{4.5}$$

where we have defined the predictive CDF per galaxy as F_i for all n test galaxies $i = 1, 2, ..., n$, and where

$$\mathbb{1}\{\tilde{y}_i \leq y\} = \begin{cases} 1, & \tilde{y}_i \leq y \\ 0, & \tilde{y}_i > y \end{cases}$$

is the *indicator function*. If the PDF's are marginally calibrated, then $\hat{F}_I$ should be the same as $\tilde{G}_I$. Figure 4.2 shows the results from this test for our model, where it can be seen that the maximum deviation between $\hat{F}_I$ and $\tilde{G}_I$ is ~ 0.015; thus, our model passed this test as well.

Therefore, we conclude in this section that we have successfully trained our ML model and applied it to the GOLD subsample in this work. Hence, we then discuss how we can utilize the products from this section to create the final lens sample.

4.2.3 The stellar-mass catalog

As we mentioned before, `GalPro` returns the full 2D joint distribution between redshift and stellar mass for each galaxy in our sample. More specifically, ML produces multiple predictions for the two properties in the $z - M_\star$ plane which form the aforementioned 2D distribution. And example of this for a galaxy in the `Y3 GOLD` is shown in figure 4.3. The

average point-mass estimate is also shown in that figure and it is what we use to define our bins later below.

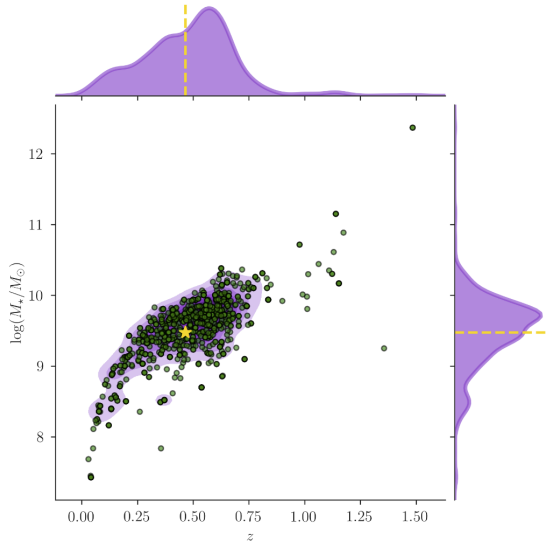


Figure 4.3: The 2D joint distribution of redshift (z) and stellar-mass ($M_\star$) for a single galaxy in the DES Y3 GOLD catalog produced by `GalPro`. The 1D projected distributions are also shown on the top and right sides. Each green point represents a prediction from `GalPro`, while the gold star is the mean point-mass estimate.

For this work we decided to define the bins in redshift to be the same as in DES Y3 (e.g. Prat et al., 2022; Abbott et al., 2022), and we thus consider four tomographic redshift bins, as defined in table 4.1. Per redshift bin, we then split galaxies into bins in stellar-mass so that, if N is the number of stellar-mass bins, the first $N-1$ contain 3×10^6 objects and the last bin contains the number of objects left in that redshift bin. The number of objects per bin, as well as the limits in stellar-mass per bin are shown in table 4.1. In that same table, we also report the exact number of galaxies per bin and their number density per square arcmin. From now on, we will denote a single bin with the indices (ℓ, m), the first of which refers to the reshift bin and the latter to the stellar-mass bins in that ℓ. We also show figure 4.4 as a visual illustration of our bins overall.

Stacking galaxies within each bin, then, can give us an estimate of how they are distributed in both redshift, $n(z)$, and stellar mass, $n(M_\star)$. The procedure to derive these distributions is the following: we iterate over all galaxies per $z-M_\star$ bin and produce their posteriors, from which we sample randomly one point. After we have collected all samples we bin them in order to construct $n(z)$ and $n(M_\star)$, or even the 2D joint distributions, $p(z, M_\star)$

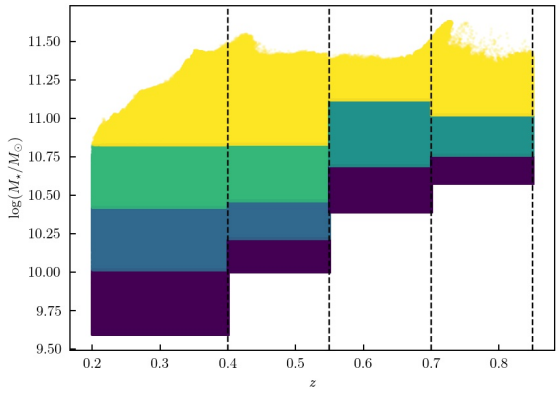


Figure 4.4: The bins for the stellar-mass sample from DES Y3. This is a scatter plot of redshift and stellar-mass. The vertical dashed lines separate the four bins in redshift, while we represent different stellar-mass bins with different colors. The "step-function" appearance of the low end of the bins is due to the completeness limit that we have applied to the samples per redshift bin.

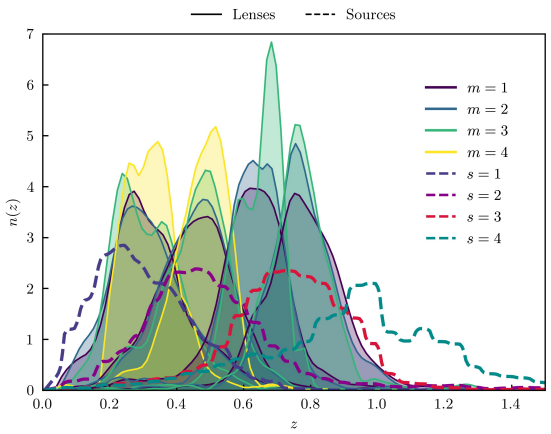


Figure 4.5: Redshift distributions, $n(z)$, of the lenses (solid filled; different mass bins are plotted with different color per redshift bin) and of the source (dashed) galaxies.

per (ℓ, m) pair. Note that we do not use the mean point estimate per galaxy during this procedure so that we sample the variance, i.e. the width, of the distributions appropriately. The redshift distributions are shown in figure 4.5.

Using the same methodology as for the redshift distributions above, we can generate the 1D distributions in stellar mass for our lenses, i.e. by sampling for the PDF's from running `GalPro`. The resulting distributions, $n(M_\star)$, are shown in figure 4.6 for all bins. Furthermore, in figure 4.7, using bin $(\ell = 1, m = 1)$ as a example (all bins exhibit similar trends), a joint 2D $z - M_\star$ distribution $p(z, M_\star)$. Given that we do not see strong correlations between z and $M_\star$, from this point on we will be assuming independence of the two quantities, such that $p(z_\ell, M_\star) \approx n_\ell(z_\ell) n_\star(M_\star)$.

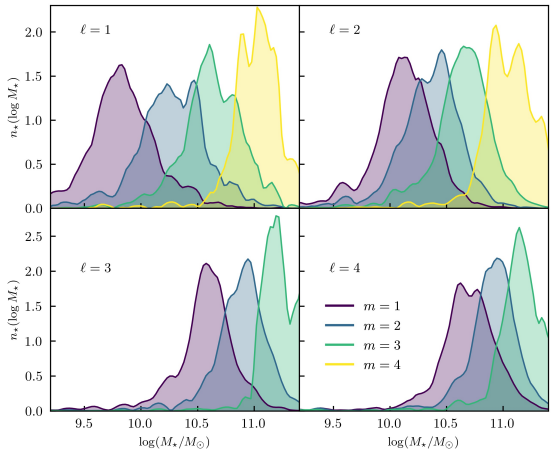


Figure 4.6: Stellar-mass distributions, $n_\star(\log M_\star)$, of our lenses. Each panel corresponds to a different redshift bin ℓ, and in each panel all stellar-mass bins m are shown.

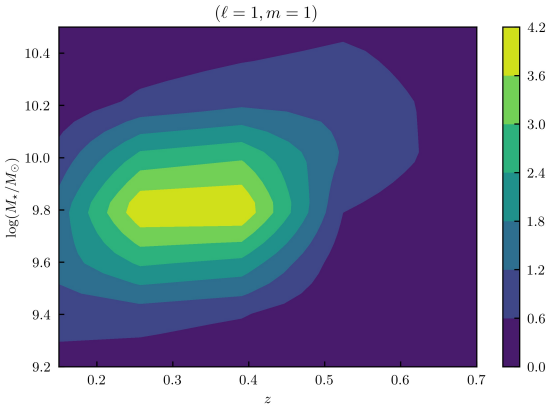


Figure 4.7: Joint 2D redshift-stellar mass distribution, $p(\log M_\star, z)$, for lens bin $(\ell, m) = (1, 1)$ galaxies.

4.3 Source galaxies

We use the DES Y3 shear catalog presented in Gatti, Sheldon et al. (2020). The galaxy shapes are estimated using the METACALIBRATION (Huff & Mandelbaum, 2017; Sheldon & Huff, 2017) algorithm. The shear catalog has been thoroughly tested in Gatti, Sheldon et al. (2020), and tests specifically tailored for tangential shear have been presented in Prat et al. (2022). In this paper we perform additional tests on this shear catalog for tangential shear measurement on small scales (Section 4.4.3).

Following Prat et al. (2022) we bin the source galaxies into four redshift bins, where details of the redshift binning and calibration is described in Myles, Alarcon et al. (2020). The redshift distributions of the sources is shown in figure 4.5, where they are plotted together with the lens distributions. Moreover, the number of galaxies per source bin and the galaxy number density per square arcmin are both shown in table 4.1.

4.4 Measurements

In this section we discuss in detail how we measure the observables of interest, namely galaxy-galaxy lensing and galaxy clustering, the covariance matrices for both, and the systematic tests that we have performed on them in order to validate them. We then proceed to talk about how some of these informed our decision on what data points and bins we needed to remove due to systematics out of our control in this work. The measurements that we choose we work with are shown, together with the best-fit model predictions, in figure 4.15 for galaxy-galaxy lensing and in figure 4.16 for galaxy clustering. We will go back to discussing these figures in Section 4.8 when we go over our model fits to these measurements.

4.4.1 Data vectors

Our shear estimator is the same as in Prat et al. (2022) and includes the boost-factor correction and random-point subtraction, which we discuss more below in this section. The full estimator is then written as

$$\gamma_t(\theta) = \frac{1}{\langle R \rangle} \left[\frac{\sum_k w_{r_k}}{\sum_i w_{\ell_i}} \frac{\sum_{ij} w_{\ell_i} w_{s_j} e^{\mathrm{LS}}_{t,ij}(\theta)}{\sum_{kj} w_{r_k} w_{s_j}} - \frac{\sum_{kj} w_{r_k} w_{s_j} e^{\mathrm{RS}}_{t,kj}(\theta)}{\sum_{kj} w_{r_k} w_{s_j}} \right] , \tag{4.6}$$

where w_{ℓ_i}, $w_{r_k} = 1$ and w_{s_j} are the weights associated with the lens galaxy i, random point k and source galaxy j, respectively. Furthermore, the weighted average METACALIBRATION response is $\langle R \rangle = \sum_j w_{s_j} R_{s_j} / \sum_j w_{s_j}$, averaging over the responses R_{s_j} of each source galaxy j, while $e^{\mathrm{LS}}_{t,ij}$ and $e^{\mathrm{RS}}_{t,kj}$ are, respectively, the measured tangential ellipticity of the source galaxy j around the lens galaxy i and random point k.

To measure galaxy clustering we use the Landy-Szalay (Landy & Szalay, 1993) estimator which is written as:

$$w(\theta) = \frac{(\mathrm{D} - \mathrm{R})^2}{\mathrm{RR}} = 1 + \frac{\mathrm{DD} - 2\mathrm{DR}}{\mathrm{RR}} \tag{4.7}$$

where DD, DR and RR are normalized weighted counts of galaxy-galaxy, galaxy-random

Redshift bin (z)	**Mass bin ($\log M_\star/M_\odot$)**	$\mathbf{N_{gal}}$ $[\times\mathbf{10^6}]$	$\mathbf{n_{gal}}$ $[\mathbf{arcmin^{-2}}]$
$\ell = 1$ $(0.20, 0.40)$	$m = 1$ $(9.5950, 10.018)$	3.00	0.218
$\ell = 1$ $(0.20, 0.40)$	$m = 2$ $(10.018, 10.424)$	3.00	0.218
$\ell = 1$ $(0.20, 0.40)$	$m = 3$ $(10.424, 10.830)$	3.00	0.218
$\ell = 1$ $(0.20, 0.40)$	$m = 4$ $(10.830, 11.504)$	1.33	0.097
$\ell = 2$ $(0.40, 0.55)$	$m = 1$ $(10.003, 10.220)$	3.00	0.218
$\ell = 2$ $(0.40, 0.55)$	$m = 2$ $(10.220, 10.467)$	3.00	0.218
$\ell = 2$ $(0.40, 0.55)$	$m = 3$ $(10.467, 10.835)$	3.00	0.218
$\ell = 2$ $(0.40, 0.55)$	$m = 4$ $(10.835, 11.557)$	1.87	0.136
$\ell = 3$ $(0.55, 0.70)$	$m = 1$ $(10.393, 10.695)$	3.00	0.218
$\ell = 3$ $(0.55, 0.70)$	$m = 2$ $(10.695, 11.124)$	3.00	0.218
$\ell = 3$ $(0.55, 0.70)$	$m = 3$ $(11.124, 11.527)$	0.53	0.039
$\ell = 4$ $(0.70, 0.85)$	$m = 1$ $(10.581, 10.765)$	3.00	0.218
$\ell = 4$ $(0.70, 0.85)$	$m = 2$ $(10.765, 11.027)$	3.00	0.218
$\ell = 4$ $(0.70, 0.85)$	$m = 3$ $(11.027, 11.634)$	1.46	0.106
$s = 1$ $(0.20, 0.43)$		24.94	1.476
$s = 2$ $(0.43, 0.63)$		25.28	1.479
$s = 3$ $(0.63, 0.90)$		24.89	1.484
$s = 4$ $(0.90, 1.30)$		25.09	1.461

Table 4.1: Lens (indexed by (ℓ, m)) and source (indexed by s) galaxy bin information. The first column corresponds to the redshift bins, while the second one corresponds to the stellar-mass bins of our lens galaxies within each redshift bin. The third and fourth columns show, respectively, the number of galaxies in each of the bins and their number density in units of gal/arcmin^2.

point and random point-random point pairs within the angular bin θ.

For both $\gamma_t(\theta)$ and $w(\theta)$ we use $N_\theta = 30$ angular bins in the range of $[0.25, 250]$ arcmin. Each galaxy i in our sample is weighted by $w_{g,i}$ in order to correct for large-scale fluctuations due to observing conditions. We train a Neural Network in order to produce our weights and the procedure we follow in delineated in Section 4.4.3. Our catalog of random points contains 30 times the number of galaxies in our lens galaxy catalog. In total we have 4 redshift bins for our lens galaxies and the number of stellar-mass bins inside each of these redshift bins is $[4, 4, 3, 3]$, respectively, with the details being presented in Table 4.1.

4.4.2 Covariance matrix

In this work we use the Jackknife (JK) covariance matrix, which we calculate as:

$$\mathcal{C}_{ij} \equiv \mathcal{C}(\gamma_t(\theta_i), \gamma_t(\theta_j)) = \frac{N_{\rm JK}-1}{N_{\rm JK}} \sum_{k=1}^{N_{\rm JK}} \Delta\gamma_i^k \Delta\gamma_j^k \ , \tag{4.8}$$

where $\gamma_t^k(\theta_i)$ is the shear in the i'th angular bin for the k'th JK resampling, $\langle\gamma_t(\theta_i)\rangle_k$ is the average over all $N_{\rm JK}$ realizations of the shear for the i'th angular bin and we have defined $\Delta\gamma_i^k \equiv \gamma_t^k(\theta_i) - \langle\gamma_t(\theta_i)\rangle_k$. We use $N_{\rm JK} = 200$ JK patches defined via the KMEANS[3] algorithm. $N_{\rm JK}$ is chosen so that the individual JK regions are at least as large as the maximum angular scale we need for our measurements.

When inverting the covariance matrix in the likelihood analysis, a correction factor is needed to account for the bias introduced from the noisy covariance (Friedrich et al., 2016). This correction is often referred to as the Hartlap (Hartlap et al., 2007) correction. When inverting the JK covariance matrix $\mathcal{C}$ we multiply it by a factor H to get the unbiased covariance (Kaufman, 1967)

$$\mathcal{C}_H^{-1} = H\mathcal{C}^{-1} = \left(\frac{N_{\rm JK} - N_\theta - 2}{N_{\rm JK} - 1}\right)\mathcal{C}^{-1} \ , \tag{4.9}$$

where the number of angular bins we use is $N_\theta = 30$, since we analyze each lens-source redshift bin combination independently. As shown in Hartlap et al. (2007), for $N_\theta/N_{\rm JK} < 0.8$ the correction produces an unbiased estimate of the inverse covariance matrix; in our case we find $N_\theta/N_{\rm JK} = 0.2$. However, it is also shown in Hartlap et al. (2007) that as this factor increases, $N_\theta/N_{\rm JK} \to 0.8$, the Bayesian confidence intervals can erroneously grow by up to 30%. Furthermore, it was shown that in order for the confidence intervals to not grow more than 5% the factor $N_\theta/N_{\rm JK} \lesssim 0.12$. For our results this means that, although our

3. `https://github.com/esheldon/kmeans_radec`

covariance matrix gets unbiased, our error bars increase and our constraints can thus look less significant than they actually are.

We compare this covariance matrix with a theoretical covariance matrix calculated using the `CosmoCov` (Krause & Eifler, 2017; Fang et al., 2020) software. Figures 4.8 and 4.9 show the comparison between the Jackknife and `CosmoCov` error bars (the square root of the diagonal of the covariance matrix) in the case of lens-source bin combination bin $(\ell, m, s) = (2, 1, 4)$, for γ_t and w respectively; we use that bin for demonstration here but we see similar trends in all bins. In figure 4.8 we also show the line $\propto 1/\theta$ which shows the expected behavior of the covariance if it were *shape noise* dominated. As we see in that plot, γ_t is dominated by shape noise over most scales. This also implies that the cross-covariance between γ_t and w in a single bin and across different redshift and stellar-mass bins is weak. Therefore, we choose to ignore the cross-covariance in this work.

The agreement between Jackknife and theory is not perfect. However, this is not very surprising given the differences in the HOD theory used by `CosmoCov` compared to this study, as the one implemented in the former is based on Zheng et al. (2005); Zehavi et al. (2011). Nevertheless, the differences we see in the covariance are not expected to have a significant effect on our data fits and we proceed to use the Jackknife covariace matrix in Section 4.7 while performing the likelihood analysis.

4.4.3 Systematics tests

The tests we have performed on the data and measurements for this work are summarized below. The relevant plots are shown in figures 4.10 and 4.11 and the results are discussed in what follows:

1. *Cross component*: The tangential shear, γ_t, is one of the two components when we decompose a spin-2 shear field. The other component is $\gamma_\times$, which is defined by the projection of the field onto a coordinate system which is rotated by 45° relative

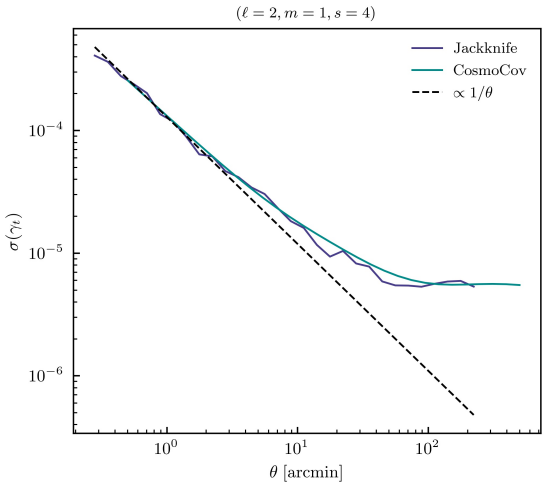


Figure 4.8: Square root of the diagonal of the Jackknife covariance matrix that we use in this work (dark blue solid line) and corresponding theory estimation from `CosmoCov` (liht blue solid line). The straight line (dashed black) $\propto 1/\theta$ shows the trend we would expect if the covariance was shape noise dominated. For this plot we used the lens-source bin combination $(\ell, m, s) = (2, 1, 4)$ for demonstration purposes.

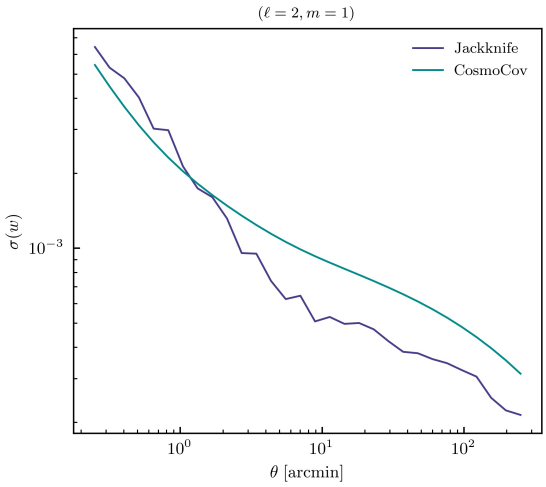


Figure 4.9: Similar plot to figure 4.8 on the left but for w for lens bin $(\ell, m) = (2, 1)$. The solid dark blue shows the size of the Jackknife error bars we use in this work and the solid light blue line is the prediction from `CosmoCov`.

to the tangential frame. For isotropically oriented lenses, the average of $\gamma_\times$ due to gravitational lensing alone should be zero. It is thus a useful test to measure this component in the data and make sure that it is consistent with zero for all angular scales. To be able to decide whether this is the case, we report the total χ^2 calculated for $\gamma_\times$ when compared with the null signal. This test on our shear measurements is presented in figure 4.10. The measurements of $\gamma_\times$ at large scales are consistent with zero. At smaller scales, below a few arcmin, $\gamma_\times$ fluctuates around zero, roughly within the error bars. Considering that at small scales the level of noise increases, we do not find the behavior of $\gamma_\times$ worrisome. When the reduced χ^2 is examined per bin we see that some bins have a somewhat large value, but the χ^2 distribution is still largely consistent with the theoretical expectation.

2. *Random points*: The mean tangential shear around random points tests the importance of geometrical and mask effects in the signal. Although our estimator of galaxy-galaxy lensing includes the subtraction of tangential shear measurement around random points, it is useful to check that this correction is small, especially for the bins with the highest signal. For this work we use a number of random points that is 30 times the number of lenses in each bin. We note here, however, that we have tested different numbers of random points, both higher and lower, and we have verified that this number is sufficiently large for our purposes. See also Prat et al. (2022) for a discussion on random-point subtraction and its effects on the measured shear. The effect from including the random-point subtraction is shown on figure 4.10, where we can see that it is a small effect, and within the error bars in all bins, as also indicated by the reduced χ^2 values we obtain.

3. *LSS weights*: Photometric surveys are subject to galaxy density variations throughout the survey footprint due to time-dependent observing conditions. This variation in the density of the lenses must be accounted for by applying the LSS-weights, which

removes this dependence on observing conditions, such as exposure time and air-mass. In galaxy-galaxy lensing, since it is a cross-correlation probe, the impact of observing conditions is small compared to e.g. galaxy clustering. Therefore, in this test we compare the measurements with and without the application of the LSS-weighting scheme and report the difference between the two. We discuss in detail how we generate the weights in Section 4.4.4. This is a test performed both on γ_t and w, and the results from performing the test are shown, respectively for the two observables, in figures 4.10 and 4.11. Comparing the measured shear with and without applying the LSS weights leads to no practically no differences, as also indicated by the very small reduced χ^2 of each panel. For clustering, the effects are larger, especially in small scales and we thus conclude that it is far more important to correct for this effect in w, as we do for both in our fiducial measurements.

4. *Boost factors*: Galaxies are clustered, and the number of sources around a lens can be larger than what we would expect from a uniform distribution. This is usually quantified by the *boost factor* (Sheldon et al., 2004), $B(\theta)$, estimator which is the excess in the number of sources around a lens with respect to randoms. This is again a test only relevant for γ_t and the results are plotted in figure 4.10. As can be seen from the plots, the contribution from this effect can be large at small scales, especially when the bins are more overlapped in redshift. In our analysis we take the boost factors into account by correcting for it before carrying out the model fit. For a further discussion on how boost factors help us with bin selection and scale cuts see Section 4.4.5.

4.4.4 Neural-network weights

In this section we discuss how we produce the LSS weights, which are heavily mentioned in Section 4.4.3, in order to correct for systematics on large angular scales in our measurements of galaxy clustering and galaxy-galaxy lensing. The methodology we implement in this work

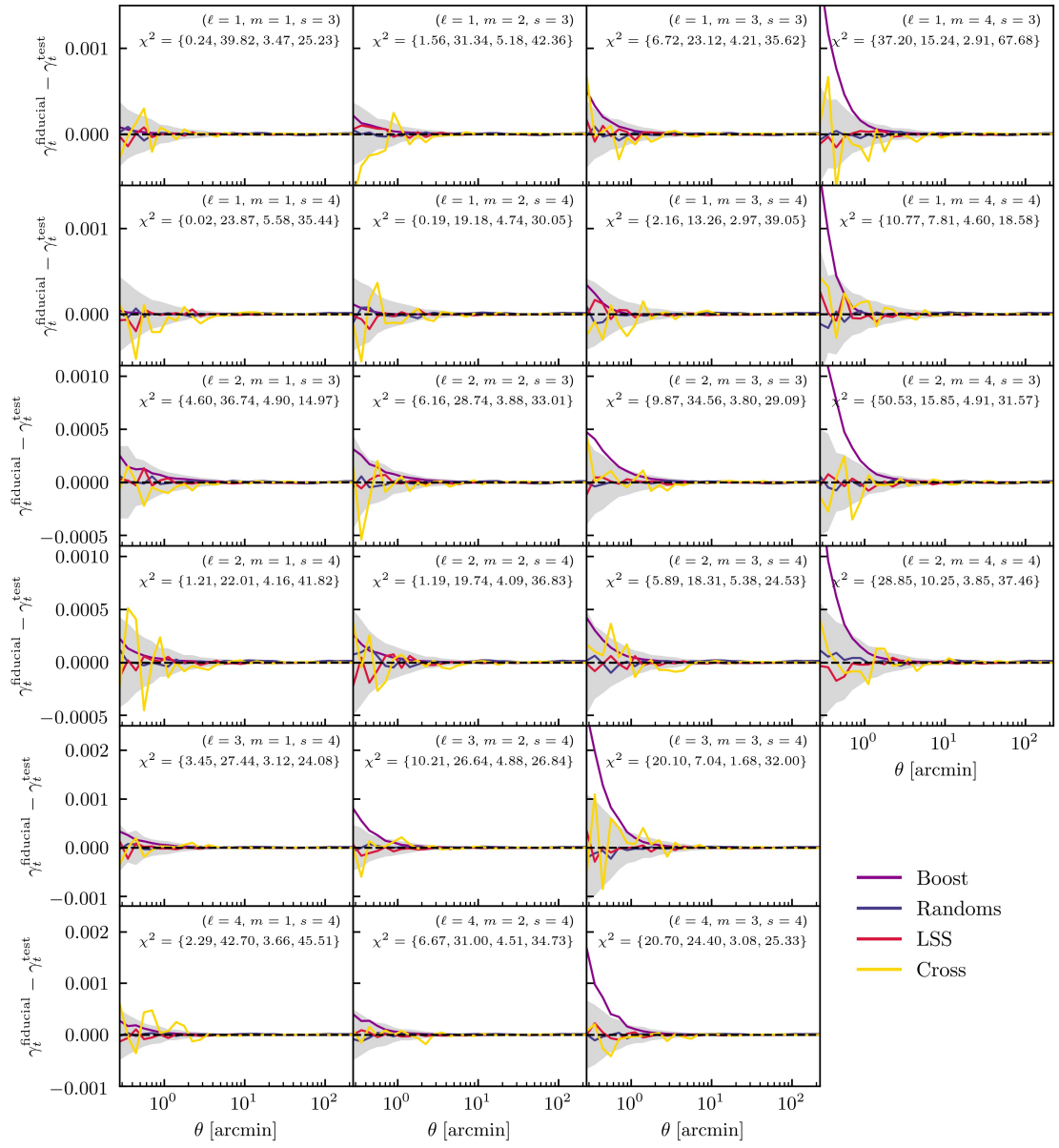


Figure 4.10: Systematics tests for the γ_t measurements, as discussed in Section 4.4.3, for the stellar-mass sample. *Boosts:* Comparison of γ_t with and without applying the boost factor correction; *Cross component:* The cross-component of shear; *Random points*: Effect from applying the random-point subtraction; *No LSS weights*: Effect from not applying the LSS weights to correct for observing conditions; *Gray area:* The error bars on the shear measurement. In each panel we also list the χ^2 (the order of which follows the order in the legend) between each test and the null, using the covariance of our γ_t measurements. The number of points for each of the lines is 30.

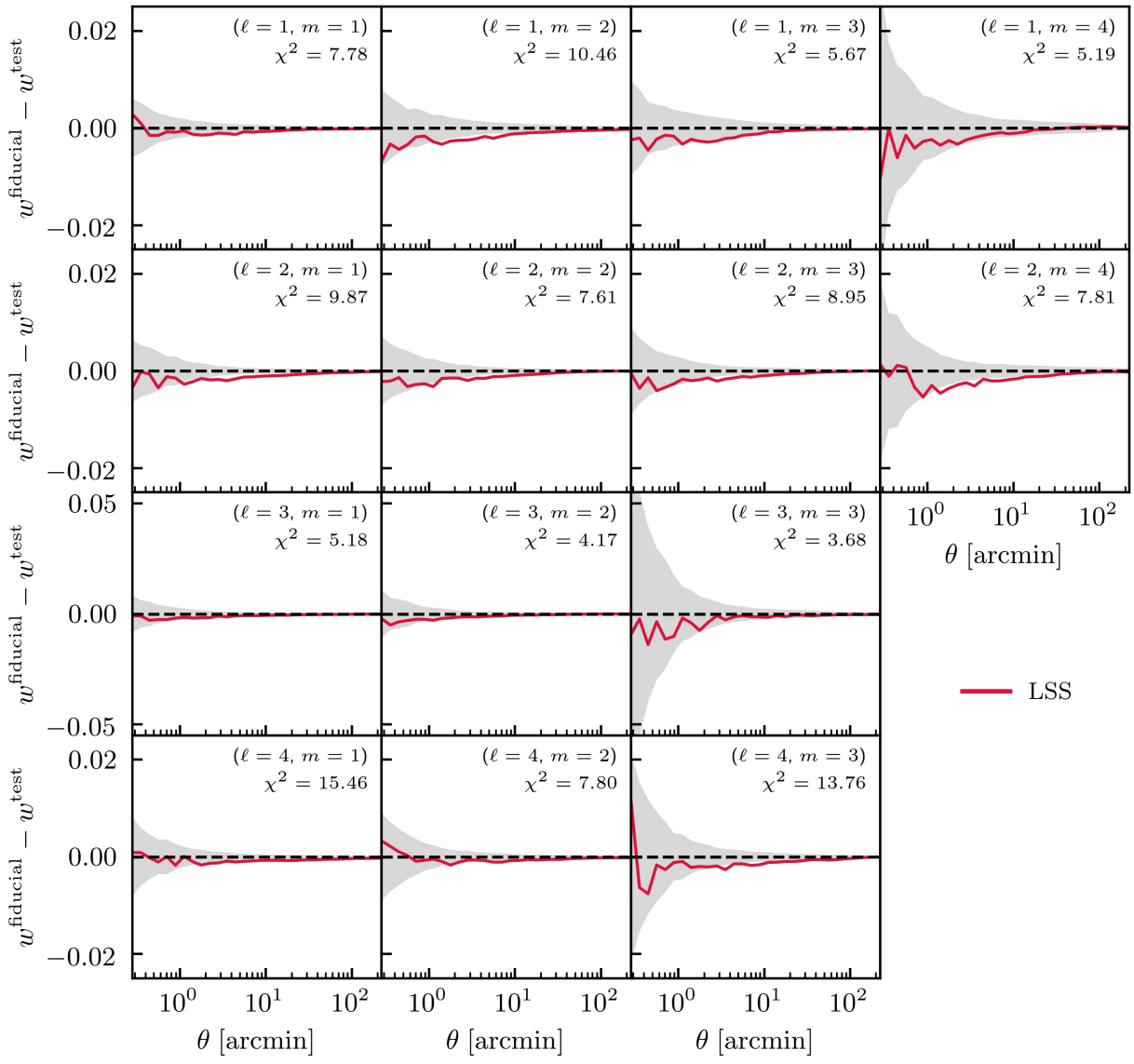


Figure 4.11: Systematics tests for w, as discussed in Section 4.4.3, for the stellar-mass sample. *No LSS weights*: Effect from not applying the LSS weights to correct for observing conditions; *Gray area:* The error bars on the shear measurement. In each panel we also list the χ^2 between each test and the null, using the covariance of our w measurements. The number of points for each of the lines is 30.

is very similar to what is done in Sánchez et al. (2022b), so we refer the reader to that paper for a more extended discussion on what follows. Given that our galaxy sample is generic, due to the faint nature of a large fraction the galaxies in it, it is expected that their selection function will fluctuate within the survey footprint due to observing conditions. These fluctuations induce correlations between the galaxy density in the footprint and maps of survey properties. These fluctuations therefore produce signal that mimics galaxy clustering and contaminates our measurements; it is thus crucial to correct for this effect. We do so by utilizing a Neural Network (NN) which is trained so that it models the correlations between galaxy density and the Survey Property (SP) maps. The SP maps, in all four *griz* bands, we utilize are the following:

- **Depth:** Mean survey depth, computed as the mean magnitude for which galaxies are detected at S/N = 10.
- **Sky Brightness:** Estimated sky brightness, or more precisely, the standard deviation of sky pixels due to shot noise and read noise, measured in units of electrons/second/pixel.
- **Exposure time:** Total exposure time at a given point in the survey footprint, measured in seconds.
- **Airmass:** Mean airmass, computed as the optical path length for light from a celestial object through Earth's atmosphere (in the secant approximation), relative to that at the zenith for the altitude of the telescope site.
- **Seeing:** Mean seeing, measured in arcseconds, computed as the full width at half maximum of the flux profile.
- **FWHM:** Full width at half maximum of seeing.

- **Galactic extinction:** We use the SFD dust extinction map from Schlegel et al. (1998), which measures the $E(B-V)$ reddening, in magnitudes.

- **Stellar density:** We use a map of stellar density, in $\deg^{-2}$, using stellar sources from Gaia EDR3 (Gaia Collaboration, 2021).

Therefore, we account for in total 26 SP maps per lens redshift-stellar mass bin as the *features* and we train our NN to model the correlations between those maps and galaxy count maps, which play the role of the *labels*. The NN is able to model nonlinear relations between the SP maps and the galaxy count maps. We note here that in order to avoid the NN learning about the clustering of galaxies in the count maps, each pixel in the latter maps have a label of either 0 (empty pixel) of 1 (occupied by ≥ 1 galaxy pixel). The *loss function* for the network is the binary *cross-entropy* between the predicted pixel occupancy and the occupancy of the training set.

The architecture of the network is based on our guess that the selection function scales primarily as some power law combination of the SP's. To this end, the input SP values are all logarithmically scaled (except those, such as depth, which are already logarithmic quantities), and the output of the network is exponentiated to form the selection probability. The network output is a sum of two branches: the first branch is a simple linear combination of the 26 scaled SP's, since we expect this to capture most of the functional variation. The second branch is intended to capture departures from a simple power law: it takes the input layer of 22 dimensions through 3 hidden layers of 64, 32 and 4 fully connected neurons, respectively, and a single neuron on the output layer, each with *relu activation*. The output of the network, for each tomographic bin, consists of a single value for each `Healpixel` within our mask, which will be used to weigh the galaxies accordingly.

To prevent the network from overfitting, it is constructed with *k-fold cross-validation*, which works in the following way: The `NSIDE` $= 4096$ maps are re-binned into a coarser grid of `NSIDE` $= 16$ (with a resolution of about 4 degrees). We then randomly divide these cells

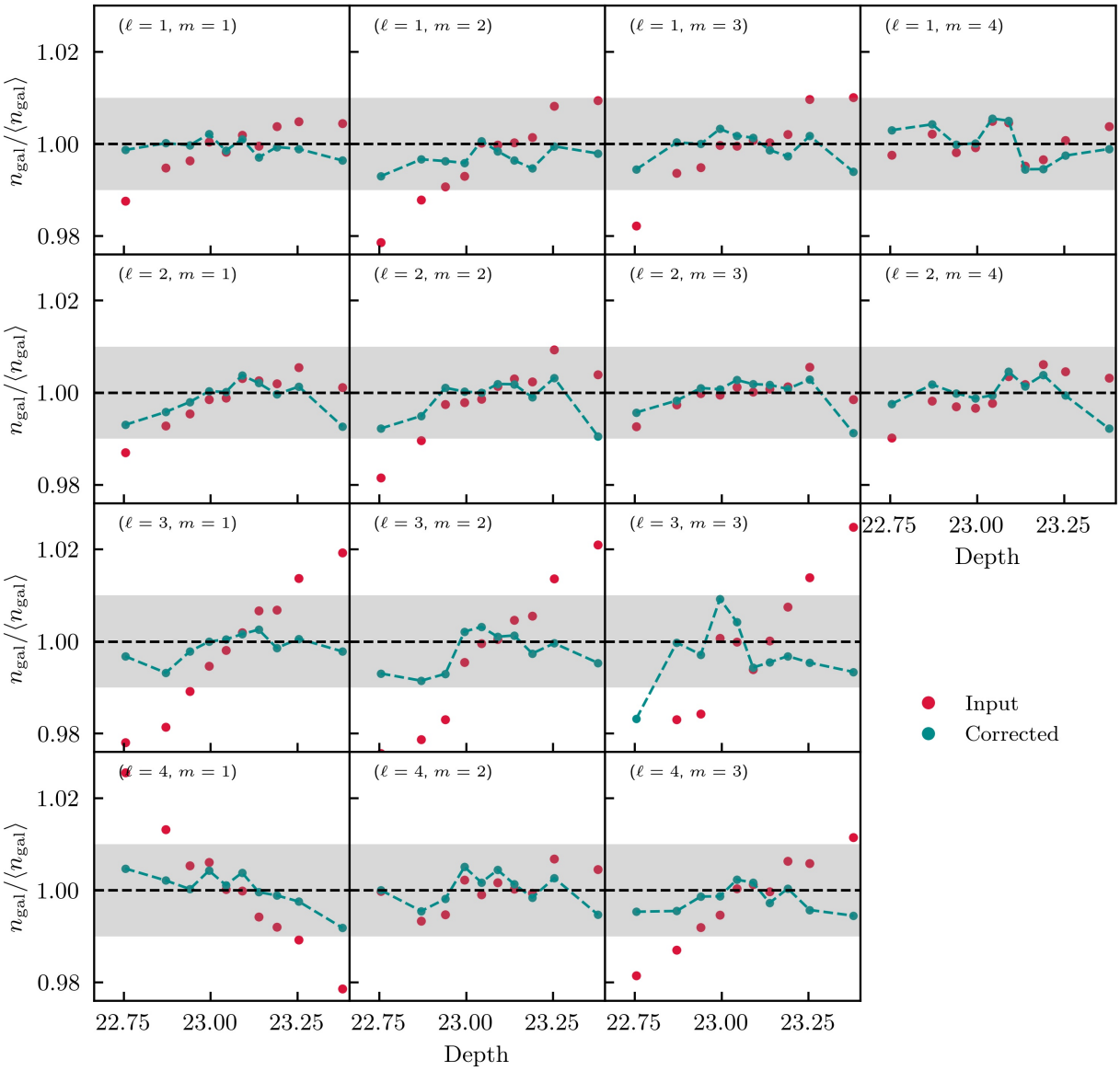


Figure 4.12: Correlations between the i-band depth survey property and observed galaxy density relative to the mean density over the full footprint, before (red) and after (blue) the correction using the Neural-Network weights from Section 4.4.4. We show this relationship for all (ℓ, m) lens bins as noted in each panel. The gray shaded region in every panel corresponds to 1% deviation from unity.

into k equal-area groups. To derive the weights for a given fold k, we train the NN on the other folds, using fold k as a validation sample (the training halts when the training metric no longer improves on the validation set). This cross-validation scheme will only work to prevent overfitting on scales below the resolution defined by `NSIDE` split, in this case around 4 degrees. As also mentioned in Sánchez et al. (2022b), testing the corrected and uncorrected galaxy clustering on log-normal mock catalogs demonstrated no overfitting from this method at scales below 1 degree, and an impact of around 5% overfitting at scales of 2 degrees. We thus keep the galaxy clustering analysis in this work to angular scales below one degree, to be conservative.

We present the correction from applying the NN weights to the correlation between the galaxy number density $n_{\rm gal}$, scaled to the average density over the full footprint, $\langle n_{\rm gal}\rangle$, and the i-band depth SP in figure 4.12. We can see how the NN weights are able to capture the non-linear relation between $n_{\rm gal}/\langle n_{\rm gal}\rangle$ and that SP and produce weights that correct for it: The input points deviate from unity significantly, especially in some bins, but the corrected ones are within the 1% error band. Due to the large number of SP maps and lens bins that we used we are not showing all of the plots here but rather note that they all exhibit a similar behavior. This validates the weights and we thus conclude that we can safely use them in this work.

4.4.5 Bin selection and scale cuts

In this section we discuss the procedure we follow in order to determine two things in our analysis: 1) What are the bins we can most reliably model and use? 2) What scales do we need to remove from the data vectors due to systematics we cannot control in other ways? Overall, we base our conclusions on two aspects of the data processing when measuring γ_t and w. These are the boost factors (for lensing) and the LSS weights (for both, but mostly for clustering). We address the two questions above for both lensing and clustering below.

As far as the boost factors are concerned, large values suggest potential failures in the modeling part, especially due to IA and magnification. We, therefore, choose to work only with those bins that have small boost factors, by setting a maximum threshold of $\sim 10\%$ deviation from unity. This results in lens and source redshift bin combinations that are largely separated in redshift. For this work, therefore, we select redshift bin combinations in γ_t for the following pairs: $(\ell, s) = \{(1,3),(1,4),(2,3),(2,4),(3,4),(4,4)\}$, which is based on the results from figure 4.10. In those bins we further remove the small-scale data points with boost factor values above the threshold. In addition, since we have excluded bins where IA is expected to have a significant contribution we can ignore this term in our fiducial model fits to the data. For all (ℓ, s) pairs we consider all stellar-mass bins, as listed in table 4.1.

On the galaxy clustering part, we do not remove bins, but we determine the scale cuts based on the LSS weights that are discussed in Section 4.4.4. The effects from correcting for the LSS weights in w are shown in figure 4.11. Due to how our method of producing the weights works, we need to remove large angular scales in our data fits above the size of about one degree, and thus angular scales of $\theta \gtrsim 60$ arcmin in $w(\theta)$ are removed. See Section 4.4.4 for the details on why these scales are excluded.

4.5 The Halo Occupation Distribution model

In this section we discuss how we model the Halo Occupation Distribution (HOD) for this work. This model is based on Zacharegkas et al. (2022) and it is modified so that the stellar-mass information of the galaxies is accounted for. This is done by incorporating a Stellar-to-Halo-mass relation (SHMR) into the HOD framework. Below we introduce the SHMR we adopt in this study and we then proceed to describe our full HOD model. For an extensive review on the galaxy-halo connection and stellar-to-halo-mass relations commonly used see Wechsler & Tinker (2018).

4.5.1 The Stellar-to-Halo-Mass relation

Henceforth, we will be using the notation $f_{\rm SHMR}(M_h) \equiv M_\star(M_h)$ for the SHMR. There are a lot of variants of the this relation used in the literature and, for the most part, they are in good agreement. However, especially at large masses, there can exist significant differences.

At redshift z, we choose to parameterize the SHMR in this work following Moster et al. (2013) as:

$$f_{\rm SHMR}(M_h) \equiv M_\star(M_h) = 2AM_h \left[\left(\frac{M_h}{M_1}\right)^{-\beta} + \left(\frac{M_h}{M_1}\right)^{\gamma} \right]^{-1} , \tag{4.10}$$

where the overall normalization is defined by A, M_1 is a characteristic mass scale where the SHMR shows a break from one power-law to another of different index, and the two power-law parameters β and γ control the low- and high-mass ends of the ratio $M_\star/M_h$. We have therefore four parameters in our SHMR. However, by adding a redshift dependence to each of them, we end up with eight parameters to tune:

$$A = A_0 + A_1(1-a) , \tag{4.11}$$

$$\log M_1 = M_{1,0} + M_{1,1}(1-a) , \tag{4.12}$$

$$\beta = \beta_0 + \beta_1(1-a) , \tag{4.13}$$

$$\gamma = \gamma_0 + \gamma_1(1-a) , \tag{4.14}$$

where $a(z) = 1/(z+1)$. Namely, the parameter vector from the SHMR alone is

$$\boldsymbol{\theta}_{\rm SHMR} = \{A_0, A_1, M_{1,0}, M_{1,1}, \beta_0, \beta_1, \gamma_0, \gamma_1\} .$$

As we discuss later in Section 4.7, depending on the data vectors we want to fit our model to, that many parameters can introduce too much freedom which results in issues with the fitting process. Mostly, we are interested in constraining the "0" part of the SHMR, usually fixing the redshift-dependent part to the best-fit values from the literature (Moster et al.,

2010).

4.5.2 The Halo Occupation Distribution conditioned on stellar-mass

For a galaxy sample selected by a stellar-mass threshold $M_\star^t$, i.e. $M_\star \geq M_\star^t$, the central HOD at redshift z is similar to the most basic form of HOD first presented in Zheng et al. (2007):

$$\langle N_c(M_h|M_\star^t)\rangle = \frac{f_{\rm cen}}{2}\left[1+{\rm erf}\left(\frac{\log M_h - \log f_{\rm SHMR}^{-1}(M_\star^t)}{\sigma_{\log M_h}}\right)\right] , \tag{4.15}$$

where the parameter $\sigma_{\log M_\star}$ describes the scatter in stellar mass for a given halo mass. The parameters that characterize the centrals are then: (1) $f_{\rm cen}$ which represents the fraction of central galaxies that made it into the sample after we apply the various cuts; (2) the parameter $\sigma_{\log M_h}$ which quantifies the scatter in $\log M_h$ and is related to the scatter in stellar mass via the relation $\sigma_{\log M_\star}/\sigma_{M_h} = d\log M_\star/d\log M_h$; (3) the parameters that enter the SHMR as mentioned in Section 4.5.1. As a useful note, if we map the threshold stellar mass to the more familiar HOD parameter $M_{\rm min}$ (see the parametrization in, e.g., Zacharegkas et al., 2022) via the relation $M_{\rm min} = f_{\rm SHMR}^{-1}(M_\star^t)$, we can immediately see how the two parametrizations relate to each other in a direct way.

We are then interested in converting the above HOD to a central occupation *conditioned on a range of stellar masses* $[M_{\star,1}, M_{\star,2}]$. The simplest way of doing this is to consider the two limits as two thresholds, i.e. $M_{\star,1} \equiv M_\star^{t1}$ and $M_{\star,2} \equiv M_\star^{t2}$, and subtract the two $M_\star$-thresholded HOD's, which yields:

$$\langle N_c(M_h)\rangle = \langle N_c(M_h|M_\star^{t1})\rangle - \langle N_c(M_h|M_\star^{t2})\rangle \, . \tag{4.16}$$

The above approach has the advantage of being straightforward and easy to implement in practice, and can work well in fitting the data (see e.g. Leauthaud et al., 2011, and section 4.7). However, for a more accurate description of the data, especially when the

measurements are as precise as in this work, we may need to use a HOD model that accounts for the exact distribution that the galaxies follow in stellar mass. This is the subject of Section 4.5.3.

For the satellite galaxies, a similar approach can be used when the same range of stellar masses is considered. We can thus write:

$$\langle N_s(M_h)\rangle = \langle N_s(M_h|M_\star^{t1})\rangle - \langle N_s(M_h|M_\star^{t2})\rangle \ , \tag{4.17}$$

where we model the satellite occupation for a $M_\star$-thresholded sample as a re-scaled version of the central HOD multiplied by a power-law:

$$\langle N_s(M_h|M_\star^{t})\rangle = \frac{f_{\rm sat}}{f_{\rm cen}}\langle N_c(M_h|M_\star^{t})\rangle \left(\frac{M_h}{M_{\rm sat}}\right)^{\alpha_{\rm sat}} \ , \tag{4.18}$$

where $f_{\rm sat}$, $M_{\rm sat}$ and $\alpha_{\rm sat}$ are free parameters. Specifically, $f_{\rm sat}$ is the equivalent of $f_{\rm cen}$ in the central HOD of equation (4.15), $\alpha_{\rm sat}$ controls the power-law behaviour of the satellite occupation at large halo masses, and the stellar-mass dependence enters implicitly into the satellite modeling via the parameter $M_{\rm sat}$ (see, e.g. Leauthaud et al., 2011; Zu & Mandelbaum, 2015), as:

$$\frac{M_{\rm sat}}{10^{12}\ M_\odot} = B_{\rm sat}\left(\frac{f^{-1}_{\rm SHMR}(M_\star^t)}{10^{12}\ M_\odot}\right)^{\beta_{\rm sat}} = B_{\rm sat}\left(\frac{M_{\rm min}}{10^{12}\ M_\odot}\right)^{\beta_{\rm sat}} \ , \tag{4.19}$$

where the new free parameters $B_{\rm sat}$ and $\beta_{\rm sat}$ have been introduced. It is worth explicitly noting here that the set of parameters $(B_{\rm sat}, \beta_{\rm sat})$ modifies the SHMR of the satellite galaxies from that of the centrals, i.e. allowing for the satellites to follow a different behaviour. This is important as we do not expect the satellites and centrals to both follow the same SHMR, but it is reasonable to assume that the overall behaviour of them is similar, i.e. a similar functional form might be utilized for both. The above also demonstrates that

there is a connection between the $M_{\rm min}$ parameter and the satellite normalization $M_{\rm sat}$; this comes from studies (see, e.g. Kravtsov et al., 2004; Zheng et al., 2007; Tinker et al., 2007; Zheng et al., 2009; Abbas et al., 2010; Zehavi et al., 2011) that find a self-similarity in the occupation numbers, so that $M_{\rm min}/M_{\rm sat} \sim$ constant for luminosity-defined samples, but by introducing $B_{\rm sat}$ and $\alpha_{\rm sat}$ we give the model more flexibility. Note that, Zheng et al. (2007) find $M_{\rm sat}/M_{\rm min} \sim 16 - 18$. As a further step in the satellite modeling, following e.g. Kravtsov et al. (2004); Conroy et al. (2006); Leauthaud et al. (2011), it might be better to introduce an exponential cutoff to the high-mass end of the satellites, where $\langle N_c \rangle \sim 1$ and they follow a power-law behavior, given that a pure power-law fit might yield artificially steep slopes $\alpha_{\rm sat}$ (Conroy et al., 2006); we thus write:

$$\langle N_s(M_h|M_\star^t)\rangle = \frac{f_{\rm sat}}{f_{\rm cen}}\langle N_c(M_h|M_\star^t)\rangle \left(\frac{M_h}{M_{\rm sat}}\right)^{\alpha_{\rm sat}} \exp\left(-\frac{M_{\rm cut}}{M_h}\right) , \tag{4.20}$$

where the $M_{\rm cut}$ parameter is again related to $M_{\rm min}$ in a similar manner as $M_{\rm sat}$ does, and thus:

$$\frac{M_{\rm cut}}{10^{12}\ M_\odot} = B_{\rm cut}\left(\frac{f^{-1}_{\rm SHMR}(M_\star^t)}{10^{12}\ M_\odot}\right)^{\beta_{\rm cut}} = B_{\rm cut}\left(\frac{M_{\rm min}}{10^{12}\ M_\odot}\right)^{\beta_{\rm cut}} , \tag{4.21}$$

with the two new parameters $B_{\rm cut}$ and $\beta_{\rm cut}$.

The total HOD prediction in the thresholded case is then given by:

$$\langle N(M_h|M_\star^t)\rangle = \langle N_c(M_h|M_\star^t)\rangle + \langle N_s(M_h|M_\star^t)\rangle \ . \tag{4.22}$$

The above HOD is only exactly correct when a *top-hat*, i.e. uniform, distribution of stellar masses is assumed. The full expressions that take into account the exact stellar-mass distributions are described in Section 4.5.3 below.

4.5.3 HOD given joint redshift-stellar mass distribution

In this section we derive the theory of our new HOD model from simple principles and basic assumptions. For more, e.g. also Leauthaud et al. (2011); Zu & Mandelbaum (2015) where a similar framework is developed. The goal is to map the stellar-mass information to the dark matter content of a halo, given its mass. Moreover, we would like to incorporate into this model the distribution $p(M_\star, z_\ell)$ which is the probability density function (PDF) to find a galaxy of mass $M_\star$ at redshift z_ℓ. The normalization of $p(M_\star, z_\ell)$ is such that:

$$\int_{M_{\star,1}}^{M_{\star,2}} dM_\star \int_{z_{\ell,1}}^{z_{\ell,2}} dz_\ell \; p(M_\star, z) = 1, \tag{4.23}$$

given a stellar mass range $[M_{\star,1}, M_{\star,2}]$ and a redshift range $[z_{\ell,1}, z_{\ell,2}]$.

We start from considering the probability $p(M_h, M_\star, z)$ that a galaxy of stellar mass $M_\star$ exists in a halo of mass M_h at redshift z, which is normalized to unity:

$$\int\int\int dM_h dM_\star dz \; p(M_h, M_\star, z) = 1 \; . \tag{4.24}$$

The goal would then be: 1) to relate this to the HOD $\langle N(M_h) \rangle$ of the previous Section 4.5.2, and 2) to model this PDF so that we can obtain a parametrized theory prediction. For the remainder of this section we assume that the HOD does not evolve with redshift within the interval $[z_{\ell,1}, z_{\ell,2}]$, and account only for the correlation between stellar mass and redshift.

We can use some very basic arguments to do so, starting from computing the probability to find a galaxy in a halo of some mass. Given the HOD notation we have used so far, we

can write:

$$\left(\begin{array}{c}\text{prob. to find gal. of } \textit{any } M_\star \\ \text{in halo of mass } M_h\end{array}\right) = \left(\frac{\text{num. den. of gal. given } M_h}{\text{total num. den. of galaxies}}\right) = \frac{\langle N(M_h)\rangle \times (dn/dM_h)}{\bar{n}_g}\ , \tag{4.25}$$

where dn/dM_h is the Halo Mass Function (HMF) and $\bar{n}_g$ is the anerage number density of galaxies. But also, the same probability can be expressed in terms of $p(M_h, M_\star)$ as:

$$\left(\begin{array}{c}\text{prob. to find gal. of } \textit{any } M_\star \\ \text{in halo of mass } M_h\end{array}\right) = \int_{z_{\ell,1}}^{z_{\ell,2}} \int dM_\star dz_\ell\ p(M_h, M_\star) \times n_\star(\log M_\star|z_\ell)\ , \tag{4.26}$$

where we take into account the stellar-mass distribution of a galaxy at redshift z_ℓ by multiplying by the conditional probability function $n_\star(\log M_\star|z_\ell) = p(\log M_\star, z_\ell)/n_\ell(z_\ell)$, with $n_\ell(z_\ell)$ being the redshift distribution and where we have defined $n_\star(\log M_\star) \equiv [M_\star/\log(e)]n_\star(M_\star)$. Then, $p(\log M_\star, z) = [M_\star/\log(e)]p(M_\star, z)$. Therefore, combining equations (4.25) and (4.26) we have:

$$\langle N(M_h)\rangle = \bar{n}_g \left(\frac{dn}{dM_h}\right)^{-1} \int dM_\star\ p(M_h, M_\star) n_\star(\log M_\star|z_\ell)\ . \tag{4.27}$$

What is missing in equation (4.27) is a model for $p(M_h, M_\star)$. We can, again, use basic

arguments as above to model this PDF. Specifically, by definition, it is:

$$
\begin{aligned}
p(M_h, M_\star) &:= \left(\begin{array}{c} \text{prob. to find gal. of } M_\star \\ \text{in halo of mass } M_h \end{array} \right) \\
&= \left(\frac{\text{num. den. of gal. in } \delta M_\star \text{ given } M_h}{\text{total num. den. of galaxies}} \right) \times p(M_h) \\
&= \frac{(\delta \langle \tilde{N}(M_\star|M_h) \rangle / \delta M_\star) \times (dn/dM_h)}{\bar{n}_g} \\
&= \frac{(\delta \langle \tilde{N}(M_\star|M_h) \rangle / \delta \log M_\star) \times (dn/dM_h)}{\ln(10) M_\star \bar{n}_g} \\
&\equiv \frac{\langle N(M_\star|M_h) \rangle}{\ln(10) M_\star \bar{n}_g} \left(\frac{dn}{dM_h} \right) \, . \qquad (4.28)
\end{aligned}
$$

In the above, we defined $\delta \langle \tilde{N}(M_\star|M_h) \rangle / \delta M_\star$ to be the number of galaxies, $\delta \langle \tilde{N}(M_\star|M_h) \rangle$, that have stellar mass in an infinitesimally small range of $\delta M_\star$ and exist in a halo of mass M_h. We then rewrote it in terms of logarithmic stellar-mass bins, i.e. in $\delta \log M_\star$ in order to convert our notations to the "usual" HOD one of $\delta \langle \tilde{N}(M_\star|M_h) \rangle / \delta \log M_\star \equiv \langle N(M_\star|M_h) \rangle$. The latter is directly parametrized, e.g. similarly to the $M_\star$-thresholded HOD of Section 4.5.2. But before we discuss that, this is the final equation for our HOD:

$$
\begin{aligned}
\langle N(M_h) \rangle &= \int_{M_{\star,1}}^{M_{\star,2}} d \log M_\star \, \langle N(M_\star|M_h) \rangle n_\star(\log M_\star|z) \\
&= \int_{M_{\star,1}}^{M_{\star,2}} d \log M_\star \, \langle N(M_\star|M_h) \rangle \frac{p(\log M_\star, z)}{n_\ell(z_\ell)} \, , \qquad (4.29)
\end{aligned}
$$

after plugging equation (4.28) into (4.27). In a redshift bin $z_\ell \in [z_{\ell,1}, z_{\ell,2}]$ we can also

calculate the average HOD as:

$$\langle \bar{N}(M_h)\rangle = \int_{z_{\ell,1}}^{z_{\ell,2}} dz_\ell \; n_\ell(z_\ell)\langle N(M_h)\rangle$$
$$= \int_{z_{\ell,1}}^{z_{\ell,2}} \int_{M_{\star,1}}^{M_{\star,2}} dz_\ell d\log M_\star \; \langle N(M_\star|M_h)\rangle p(\log M_\star, z) \; . \tag{4.30}$$

Notice here how, if we assume a top-hat distribution $U(M_{\star,1}, M_{\star,2})$ in stellar-mass, and if $p(\log M_\star, z_\ell) = [M_\star/\log(e)]n_\ell(z_\ell)/(\log M_{\star,2} - \log M_{\star,1})$, equation (4.29) reduces to the tophat HOD that we described in Section 4.5.2.

4.5.4 Model for $\langle N(M_\star|M_h)\rangle$

To complete the modeling of the galaxy occupation statistics we developed above in this section, we need to model $\langle N(M_\star|M_h)\rangle$ which is needed in equation (4.29). We do this by separately modeling the central and satellite occupations, similarly to Zacharegkas et al. (2022).

For the central occupation it is straightforward to do this, as in Zacharegkas et al. (2022). Specifically, we consider a log-normal distribution in $\log M_\star$ with mean provided by the SHMR (see Section 4.5.1) as $f_{\rm SHMR}(M_h)$ and a scatter around the mean denoted by $\sigma_{\log M_\star}$. We then write

$$\langle N_c(M_\star|M_h)\rangle = \frac{f_{\rm cen}}{\sqrt{\pi}\sigma_{\log M_\star}} \exp\left[-\left(\frac{\log M_\star - \log f_{\rm SHMR}(M_h)}{\sigma_{\log M_\star}}\right)^2\right] \; , \tag{4.31}$$

and thus the final model prediction for the expected number of central galaxies is

$$\langle N_c(M_h)\rangle = \int_{M_\star^{t1}}^{M_\star^{t2}} d\log M_\star \; \langle N_c(M_\star|M_h)\rangle \frac{p(\log M_\star, z)}{n_\ell(z_\ell)} \; . \tag{4.32}$$

For the satellites galaxies we have to follow a different approach. Instead of parametrizing

$\langle N_s(M_\star|M_h)\rangle$ directly, we model it as the derivative of the thresholded occupation distribution $\langle N_s(M_h|M_\star)\rangle$, which we have already modeled as equation (4.20) in Section 4.5.2, with respect to $\log M_\star$. Therefore, we have:

$$\langle N_s(M_\star|M_h)\rangle = \left.\frac{d\langle N_s(M_h|\tilde{M}_\star)\rangle}{d\log\tilde{M}_\star}\right|_{\tilde{M}_\star=M_\star} . \tag{4.33}$$

Moreover, averaging over the stellar-mass distribution, the satellite occupations yields:

$$\langle N_s(M_h)\rangle = \int_{M_{\star,1}}^{M_{\star,2}} d\log M_\star \, \frac{p(\log M_\star, z)}{n_\ell(z_\ell)} \left.\left(\frac{d\langle N_s(M_h|\tilde{M}_\star)\rangle}{d\log\tilde{M}_\star}\right)\right|_{\tilde{M}_\star=M_\star} . \tag{4.34}$$

Therefore, the total HOD prediction, accounting for the full stellar-mass distribution of the galaxies is provided via:

$$\langle N(M_h)\rangle = \langle N_c(M_h)\rangle + \langle N_s(M_h)\rangle \, . \tag{4.35}$$

In figure 4.13 we demonstrate how the stellar-mass HOD looks in two cases: using the tophat model, from equation (4.22), and when we take into accout the full 1D stellar-mass distribution of lens galaxies in a single bin, calculated from equation (4.35) and where a $n_\star$ distribution from figure 4.6 was used, as mentioned in the caption of that figure. The tophat model with $M_\star \in [M_\star^{t1}, M_\star^{t2}]$ is the difference of the two thresholded HOD's from that range. The model prediction that uses the full $n_\star$ distribution is similar to the tophat case with two noticeable differences: 1) the central HOD is more spread out in the case where the full distribution is used, as the galaxy stellar masses are distributed over a wider range than the thresholds in the tophat model's case; 2) a small offset in the amplitude of the satellite HOD at large halo masses, since more massive halos lie within the 1D full distribution with $\log M_\star > \log M_\star^{t2}$ compared to the tophat HOD. Given these differences, in Section 4.6 where we utilize the HOD to model our observables, we test and compare both cases in the process

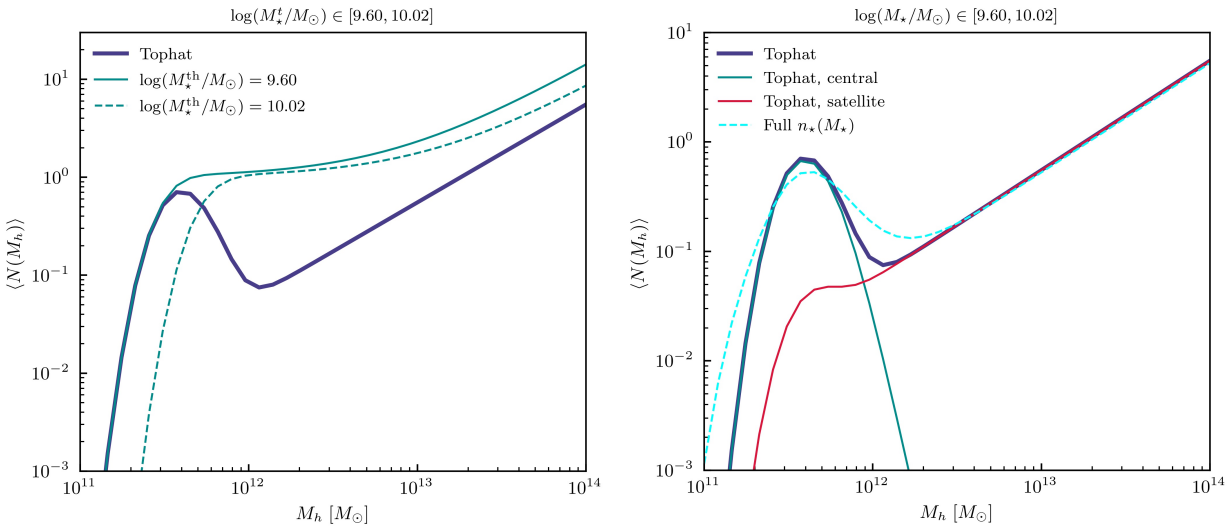


Figure 4.13: *(Left)* The stellar-mass HOD model prediction using the tophat model (dark blue), and the two thresholded HOD's (solid and dashed green) that produce it with the two thresholds shown in the legend. The HOD parameters we used to produce this plot are: $\sigma_{\log M_\star} = 0.3$, $\alpha_{\rm sat} = 1$, $B_{\rm sat} = 25$, $\beta_{\rm sat} = 1$, $f_{\rm cen} = 1 = f_{\rm sat}$; the SHMR parameters are: $A_{1,0} = 0.0351$, $A_{1,1} = -0.0247$, $M_{1,0} = 11.59$, $M_{1,1} = 1.195$, $\beta_{1,0} = 1.376$, $\beta_{1,1} = -0.826$, $\gamma_{1,0} = 0.608$, $\gamma_{1,1} = -0.329$. *(Right)* The equivalent HOD prediction using the full 1D stellar-mass distribution (dashed blue) for lens bin $(\ell, m) = (1, 1)$ from figure 4.6, compared with the tophat HOD from the left plot (solid dark blue). The central (solid green) and satellite (solid red) components of the tophat HOD are also shown for demostration of how they make up the total HOD prediction.

of determining the fiducial model we will be using for the main analysis in Section 4.7.

4.5.5 Variations of the HOD model

The number of ways one can modify a HOD model is only limited by imagination. In this above discussion, we have defined our HOD model so that the satellite occupation is dependent on the central one. The main purpose of this is to capture the cutoff on small halo mass values; if we had simply assumed a power-law form for $\langle N_s(M_h|M_\star)\rangle$ we would lose this cut-off shape and the the model would produce wrong descriptions of the data. However, we have allowed for some further freedom for the satellite term by introducing the $f_{\rm sat}$ parameter that lets the amplitude of $\langle N_s\rangle$ to vary independently from $\langle N_c\rangle$. If we want

to further allow the cutoff in $\langle N_s \rangle$ on low masses to be different from that in $\langle N_c \rangle$ we could follow various approaches. For example, we could enforce a cut-off below a mass-scale using a step function, e.g. multiplying by the *Heavyside function*, or, for a more smooth cut-off behavior, we could multiply by the *sigmoid function*. This would completely decouple the HOD's, but also introduce $4-5$ new parameters. Since there is already a lot of freedom in our HOD parametrization we would like to avoid that. We can, however, achieve the same behavior by adding just one additional parameter that would shift the cut-off mass in $\langle N_s \rangle$. Specifically, if we instead of $\langle N_c(M_h|M_\star) \rangle$ we use $\langle N_c(M_h|M_\star - \Delta M_\star^t) \rangle$ in the thresholded satellite HOD in equation (4.20), where we have introduce the mass parameter $\Delta M_\star^t$ that shifts the cutoff we have further decoupled the two HOD's in a much simpler way.

Whether or not the above modification is necessary is unclear and it depends on what the data prefers. One might assume that it is reasonable to not couple the central and satellite HOD's in order to allow the more freedom for the satellites, but in practice it is not easy to clearly determine which model is better able to fit the data.

Moreover, when we fit the data to constrain the galaxy-halo connection, we need all the constraining power we can have. Thus, we usually jointly fit multiple bins in stellar-mass and/or redshift. In such cases, we might assign a single set of HOD parameters to the joint fit. However, this might not be optimal since the HOD parameters could have a (either stellar or halo) mass and/or reshift dependence which would make each lens bin in our sample to constrain them differently. The type of dependence on, say halo mass, is not well understood, but usually a sigmoid-like functional may be assumed. A detailed study of this is left for future study, and in this work we assume that the HOD parameters remain constant within a redshift bin.

4.6 Modeling the observables

4.6.1 Galaxy-galaxy lensing

Equipped with the new HOD model, the main goal in this section is to compute the projected 2D lensing power spectrum, averaged over the $z - M_\star$ PDF, in order to transform into real space for the shear estimate. For a single lens-source redshift pair (z_ℓ, z_s) and for a stellar mass value $M_\star$ this power spectrum is computed, under the Limber approximation, by:

$$C_{\rm gm}(\ell|z_\ell, z_s, M_\star) = \frac{\rho_m \Sigma_c^{-1}(z_\ell, z_s)}{\chi^2(z_\ell)} P_{\rm gm}\left(\frac{\ell + 1/2}{\chi(z_\ell)}, z_\ell \middle| M_\star\right) , \tag{4.36}$$

where ρ_m is the average mass density of the Universe, and we define the critical surface density as

$$\Sigma_c(z_\ell, z_s) = \frac{c^2}{4\pi G} \frac{a(z_\ell)\chi(z_s)}{\chi(z_\ell)\chi(z_\ell, z_s)} , \tag{4.37}$$

with c and G being the speed of light and Newton's gravitational constant, and $\chi(z_1, z_2)$ is the comoving distance between redshifts z_1 and z_2. We then calculate the average spectrum in the following way:

$$C_{\rm gm}(\ell) = \int_{z_{s,1}}^{z_{s,2}} dz_s \; n_s(z_s) \int_{z_{\ell,1}}^{z_{\ell,2}} dz_\ell \int_{M_{\star,1}}^{M_{\star,2}} d\log M_\star \; p(\log M_\star, z_\ell) C_{\rm gm}(\ell|z_\ell, z_s, M_\star) \; . \tag{4.38}$$

In the above, the 3D galaxy-matter cross-power spectrum, $P_{\rm gm}(k, z)$, is the only function that depends on stellar mass, and this dependence enters via the HOD. Therefore, in practice we simply need to average $P_{\rm gm}$ over the stellar-mass distribution of the lenses and plug it into equation (4.38). We express these terms separately for each HOD component in what follows.

The central 1-halo term, using the equations from Section 4.5, is written as

$$\begin{aligned} P_{\rm gm}^{\rm c1h}(k,z) &= \frac{1}{\rho_m \bar{n}_g} \int dM_h \, \frac{dn}{dM_h} M_h u_{\rm dm}(k|M_h) \\ &\quad \times \int d\log M_\star \, \langle N_c(M_\star|M_h)\rangle \frac{p(\log M_\star, z)}{n_\ell(z)} \\ &= \frac{1}{\rho_m \bar{n}_g} \int dM_h \, \frac{dn}{dM_h} M_h \langle N_c(M_h)\rangle u_{\rm dm}(k|M_h) \, , \end{aligned} \tag{4.39}$$

where the average galaxy number density

$$\bar{n}_g = \int dM_h \, \frac{dn}{dM_h} \left[\langle N_c(M_h)\rangle + \langle N_s(M_h)\rangle\right] \tag{4.40}$$

is predicted from the HOD. Following the same procedure, the satellite 1-halo term turns into

$$\begin{aligned} P_{\rm gm}^{\rm s1h}(k,z) =& \frac{1}{\rho_m \bar{n}_g} \int dM_h \frac{dn}{dM_h} M_h u_{\rm dm}(k|M_h) u_{\rm sat}(k|M_h) \\ &\times \int d\log M_\star \langle N_s(M_h)\rangle \frac{p(\log M_\star, z)}{n_\ell(z)} \\ &= \frac{1}{\rho_m \bar{n}_g} \int dM_h \frac{dn}{dM_h} M_h \langle N_s(M_h)\rangle u_{\rm dm}(k|M_h) u_{\rm sat}(k|M_h) \, , \end{aligned} \tag{4.41}$$

where $u_{\rm dm/sat}(k|M_h)$ is the Fourier transform of the dark matter/satellite spatial distribution, for which we assume the NFW profile (Navarro et al., 1996). Thus, it turns out that, compared to Zacharegkas et al. (2022), we can simply replace the HOD prediction with the new one from Section 4.5 and use the same basic equations in order to calculate the projected lensing power spectrum. The 2-halo components are also the same as the ones in the aforementioned paper after we replace the HOD prediction. Namely, the central and satellite

2-halo terms are written as:

$$P_{\rm gm}^{\rm c2h}(k,z) = P_{\rm m}^{\rm nl}(k,z) \int dM_h \, \frac{dn}{dM_h} \frac{M_h}{\rho_m} b_h(M_h) u_{\rm dm}(k|M_h) \times \int dM_h' \, \frac{dn}{dM_h'} \frac{\langle N_c(M_h') \rangle}{\bar{n}_g} b_h(M_h') \tag{4.42}$$

and

$$P_{\rm gm}^{\rm s2h}(k,z) = P_{\rm m}^{\rm nl}(k,z) \int dM_h \, \frac{dn}{dM_h} \frac{M_h}{\rho_m} b_h(M_h) u_{\rm dm}(k|M_h) \times \int dM_h' \, \frac{dn}{dM_h'} \frac{\langle N_s(M_h') \rangle}{\bar{n}_g} b_h(M_h') u_{\rm sat}(k|M_h') \, . \tag{4.43}$$

In addition, the same $k \to 0$ limits also holds for the 2-halo terms, where the large-scale equations reduce to the linear model,

$$P_{\rm gm}^{\rm 2h}(k \to 0, z) \approx \bar{b}_g P_{\rm m}^{\rm lin}(k,z) \, , \tag{4.44}$$

where the average galaxy bias is

$$\begin{aligned} \bar{b}_g &= \frac{1}{\bar{n}_g} \int dM_h \, \frac{dn}{dM_h} \left[\langle N_c(M_h) \rangle + \langle N_s(M_h) \rangle u_{\rm sat}(k|M_h) \right] b_h(M_h) \\ &\approx \frac{1}{\bar{n}_g} \int dM_h \, \frac{dn}{dM_h} \left[\langle N_c(M_h) \rangle + \langle N_s(M_h) \rangle \right] b_h(M_h) \, , \end{aligned} \tag{4.45}$$

and we need to specify a model for the halo bias $b_h(M_h)$. In this work we adopt the functions from Tinker et al. (2010) for the latter.

To calculate the total power spectrum, $P_{\rm gm}(k,z)$, we combine the 1-halo and 2-halo components by summing up the two contributions at each k,

$$P_{\rm gm}(k,z) = P_{\rm gm}^{\rm 1h}(k,z) + P_{\rm gm}^{\rm 1h}(k,z). \tag{4.46}$$

As stressed by Mead et al. (2021), a linear halo bias is not necessarily a good description of the clustering relation between the halos and matter, especially on the transition scales. It could thus be important to incorporate a non-linear halo bias model into the halo model. Implementing such a "beyond-linear" halo bias model, as described in that paper, into our framework would change the shape of the 2-halo component as a function of k, especially around the scales corresponding to the size of individual dark matter halos. In this work we do so by utilizing the fitting functions from Tinker et al. (2005). Specifically, we multiply the 2-halo correlation function by $\zeta(r, z)$, which is defined as

$$\zeta(r, z) \equiv \sqrt{\frac{[1 + 1.17\xi_{\rm m}^{\rm nl}(r, z)]^{1.49}}{[1 + 0.69\xi_{\rm m}^{\rm nl}(r, z)]^{2.09}}} \,, \tag{4.47}$$

where $\xi_{\rm m}^{\rm nl}(r, z)$ is the non-linear real-space correlation function of $P_{\rm m}^{\rm nl}(k, z)$. This is the same approach also used in Zu & Mandelbaum (2015) and results in the 2-halo galaxy-cross-matter correlation function

$$\xi_{\rm gm}^{\rm 2h}(r, z) \approx \zeta(r, z)\bar{b}_g\xi_{\rm m}^{\rm nl}(r, z) \tag{4.48}$$

which we transform into Fourier space to obtain our final estimate of $P_{\rm gm}^{\rm 2h}(k, z)$. We find that using this model we are able to obtain a better fit to our data over all scales and improves the fits significantly in the 1-halo to 2-halo regime compared to what we utilized previously in Zacharegkas et al. (2022), which was also followed by Hayashi & White (2008); Zu et al. (2014); Clampitt et al. (2017). As we discuss in Section 4.7, this model is especially useful in fitting the low-redshift, low-mass bins.

We would like to more explicitly, however, write down the 1-halo part of the above power

spectrum since this is the part that is being modified on the small scales. It becomes:

$$
\begin{aligned}
C_{\rm gm}^{\rm 1h}(\ell) = \frac{1}{\bar{n}_g}\int_{z_{\ell,1}}^{z_{\ell,2}} dz_\ell \ \left[\int_{z_{s,1}}^{z_{s,2}} dz_s \ n_s(z_s-\Delta z_s)\frac{\Sigma_c^{-1}(z_\ell,z_s)}{a^2(z_\ell)\chi^2(z_\ell)}\right] \\
\times \int dM_h \ M_h \int_{M_{\star,1}}^{M_{\star,2}} d\log M_\star \ p(\log M_\star, z_\ell - \Delta z_\ell) \\
\times \left[\langle N_c(M_\star|M_h)\rangle + \langle N_s(M_\star|M_h)\rangle u_{\rm sat}(k|M_h)\right] u_{\rm dm}(k|M_h) \\
= \frac{1}{\bar{n}_g}\int_{z_{\ell,1}}^{z_{\ell,2}} dz_\ell \ n_\ell(z_\ell-\Delta z_\ell)\left[\int_{z_{s,1}}^{z_{s,2}} dz_s \ n_s(z_s-\Delta z_s)\frac{\Sigma_c^{-1}(z_\ell,z_s)}{a^2(z_\ell)\chi^2(z_\ell)}\right] \\
\times \int dM_h \ M_h \left[\langle N_c(M_h)\rangle + \langle N_s(M_h)\rangle u_{\rm sat}(k|M_h)\right] u_{\rm dm}(k|M_h) \ , \qquad (4.49)
\end{aligned}
$$

using equation (4.29) and after introducing the photo-z uncertainty parameters Δz_ℓ and Δz_s in the lens and source redshift distributions, respectively.

After all the above, to conclude this discussion, we express the final estimator for gravitational shear as:

$$\gamma_t(\theta) = (1+m_s)\int \frac{\ell d\ell}{2\pi} J_2(\ell\theta) C_{\rm gm}(\ell) \ , \qquad (4.50)$$

using the Hankel transform with the Bessel function J_2, and after introducing the *multiplicative bias* parameter m_s that quantifies any remaining errors in the γ_t estimation. In figure 4.14 (left) we show this model's prediction in the two cases of a tophat HOD and when the full 1D stellar-mass distribution for a single lens bin (as mentioned in the caption) is used, as discussed in detail in Section 4.5. The difference between the two is minuscule and mostly affects the small scales. For that reason, in this work we will be fitting the tophat model to the data in our fiducial study. This way the fitting is far more time efficient, while we do not obtain biased results, given that any differences in the two cases is captured by the uncertainty on the final constraints and the HOD's flexibility.

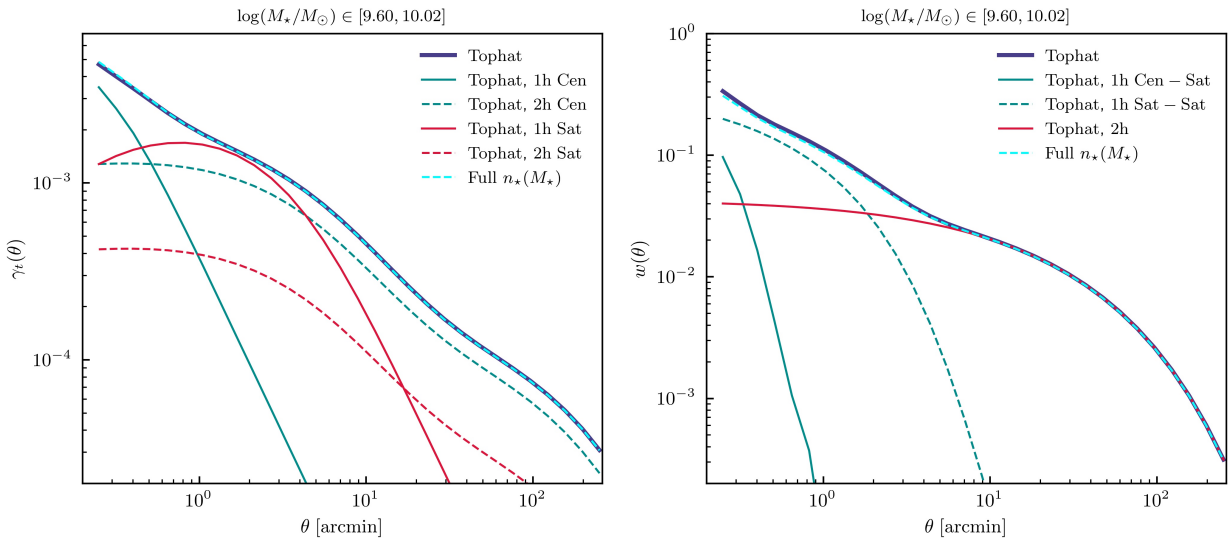


Figure 4.14: *(Left)* The γ_t prediction using the tophat model (solid dark blue) with its constituents (solid and dashed green and red lines in the legend), and the model incorporating the full 1D stellar-mass distribution (blue dashed) using the respective the HOD models from figure 4.13. *(Right)* Similar to the left plot but for w.

4.6.2 Galaxy clustering

In this section we derive the equations for modeling galaxy clustering using the stellar-mass HOD we use in this work and the equations are heavily based on van den Bosch et al. (2013). The total 3D power spectrum of the galaxy position auto-correlation is the sum of the 1-halo and 2-halo terms:

$$P_{\rm gg}(k) = P^{\rm 1h}_{\rm gg}(k) + P^{\rm 2h}_{\rm gg}(k)\ . \tag{4.51}$$

Further than that, the 1-halo term breaks down to the *1-halo central-central* $P^{\rm 1hcc}_{\rm gg}$, the *1-halo central-satellite* $P^{\rm 1hcs}_{\rm gg}$, and the *1-halo satellite-satellite* $P^{\rm 1hss}_{\rm gg}$ terms. The 2-halo term

does not break down to more components and in the $k \to 0$ limit it is written as:

$$P_{\rm gg}^{\rm 2h}(k,z) = \frac{1}{\bar{n}_g^2} P_{\rm m}^{\rm lin}(k,z) \int dM_h \, \frac{dn}{dM_h} \left[\langle N_c(M_h) \rangle + \langle N_s(M_h) \rangle u_{\rm sat}(k|M_h) \right] b_h(M_h)$$
$$\times \int dM_h' \, \frac{dn}{dM_h'} \left[\langle N_c(M_h') \rangle + \langle N_s(M_h') \rangle u_{\rm sat}(k|M_h') \right] b_h(M_h')$$
$$\approx \bar{b}_{\rm eff}^2 P_{\rm m}^{\rm lin}(k,z) \, , \tag{4.52}$$

where we have defined the "effective" large-scale average galaxy bias which in the limit where $u_{\rm sat}(k \to 0|M_h) = 1$ it is calculated by

$$\bar{b}_{\rm eff} = \frac{1}{\bar{n}_g} \int dM_h \, \frac{dn}{dM_h} \left[\langle N_c(M_h) \rangle + \langle N_s(M_h) \rangle u_{\rm sat}(k|M_h) \right] b_h(M_h) \approx \bar{b}_g \, , \tag{4.53}$$

i.e. it reduces to the average HOD galaxy bias from equation (4.45). Finally, for the 2-halo term, as we discuss in Section 4.6.2 for lensing, we find that our fits to the data behave significantly better if we modify the above by multiplying by the scale-dependent factor $\zeta(r,z)$, defined in equation (4.47). Thus, we write the final estimator of the 2-halo real-space correlation function of galaxy clustering as

$$\xi_{\rm gg}^{\rm 2h}(r,z) \approx \zeta^2(r,z) \bar{b}_g^2 \xi_{\rm m}^{\rm nl}(r,z) \tag{4.54}$$

which we then transform into Fourier space to obtain the final $P_{\rm gg}^{\rm 2h}(k,z)$.

In what follows we will model the 1-halo component of equation (4.51) starting from the following equation (Seljak, 2000):

$$P_{\rm gg}^{\rm 1h}(k,z;p) = \frac{1}{\bar{n}_g^2} \int dM_h \, \frac{dn}{dM_h} \langle N(N-1) \rangle |u_{\rm sat}(k|M_h)|^p \, , \tag{4.55}$$

where $N = N_c + N_s$ is the total number of galaxies, i.e. the sum of the number of central and satellite galaxies, N_c and N_s, respectively. We can make the approximation $p = 1$ when

$\langle N(N-1)\rangle < 1$ (since that term is be associated with the central-satellite spectrum) and $p = 2$ when $\langle N(N-1)\rangle > 1$ (since this corresponds to the satellite-satellite term). Therefore, we need to calculate $\langle N(N-1)\rangle$, which we do as follows:

$$\begin{aligned}\langle N(N-1)\rangle &= \langle N_c\rangle \left[\langle N(N-1)\rangle\right]_{N_c=1} + (1-\langle N_c\rangle)\left[\langle N(N-1)\rangle\right]_{N_c=0} \\ &= \langle N_c\rangle \left(\langle N_s^2\rangle + \langle N_s\rangle\right) + (1-\langle N_c\rangle)\left(\langle N_s^2\rangle - \langle N_s\rangle\right) \ . \end{aligned} \tag{4.56}$$

In this work we assume that the probability to find N_s satellite galaxies in a halo of mass M_h follows a *Poisson distribution* (see Kravtsov et al., 2004, for a theoretical justification), defined as

$$\mathrm{Prob}(N_s) = \lambda_s^{N_s} \times \frac{e^{-\lambda_s}}{N_s!} \ , \tag{4.57}$$

with the mean $\lambda_s = \lambda_s(M_h) = \langle N_s(M_h)\rangle$ and with

$$\langle N_s^2\rangle = \langle N_s\rangle^2 + \langle N_s\rangle \ . \tag{4.58}$$

Therefore, equation (4.56) leads to:

$$\langle N(N-1)\rangle = \langle N_c(M_h)\rangle\langle N_s(M_h)\rangle \left[2 + \langle N_s(M_h)\rangle\right] \ , \tag{4.59}$$

in which case we can write the 1-halo galaxy auto-correlation power spectrum from equation (4.55) as

$$\begin{aligned}P_{\rm gg}^{\rm 1h}(k,z) =& \frac{1}{\bar{n}_g^2}\int dM_h \ \frac{dn}{dM_h}\langle N_s(M_h)\rangle u_{\rm sat}(k|M_h) \\ & \times \left[2\langle N_c(M_h)\rangle + \langle N_s(M_h)\rangle u_{\rm sat}(k|M_h)\right] \ . \end{aligned} \tag{4.60}$$

We note here that the $P_{\rm gg}^{\rm 1hcs}$ and $P_{\rm gg}^{\rm 1hss}$ terms can be directly read from equation (4.60): they are the first and second terms in brackets, respectively. Moreover, in all the above

expressions we have already taken into account that the HOD terms are the only ones that depend on stellar mass, and thus we did not need to explicitly use the conditional probability notation, as in the galaxy-galaxy lensing case in Section 4.6.1, but we rather simply used the "usual" HOD notation from the start.

Finally, we note on the above discussion that, the central-central power spectrum is $P_{\rm gg}^{\rm 1hcc} \approx 1/\bar{n}_c$, where $\bar{n}_c = (1 - F_{\rm sat})\bar{n}_g$ is the HOD average number density of central galaxies alone and

$$F_{\rm sat} = \frac{1}{\bar{n}_g} \int dM_h \, \frac{dn}{dM_h} \langle N_s(M_h) \rangle \tag{4.61}$$

is the satellite fraction, which is always much lower than the rest of the terms. We thus ignored its contribution to the total signal. Furthermore, in $P_{\rm gg}^{\rm 1hss}$ we have ignored a contribution from the subdominant term $1/\bar{n}_s$, where $\bar{n}_s = F_{\rm sat}\bar{n}_g$ is the HOD satellite average number density. These terms would in theory dominate in the "very small-scale" limit where $u_{\rm dm/sat}(k \to \infty | M_h) = 0$. However, in practice we are never in that regime in our work.

We now conclude by modeling the projected 2D power spectrum, in the Limber approximation, as:

$$C_{\rm gg}(\ell) = \int dz_\ell \, \frac{n_\ell^2(z_\ell)}{\chi^2(z_\ell)} \left(\frac{d\chi}{dz}\Big|_{z=z_\ell} \right)^{-1} , \tag{4.62}$$

where $(d\chi/dz)|_{z_\ell}$ is the derivative of the comoving distance with respect to redshift at $z = z_\ell$, and we have used the galaxy selection function, which is essentially the normalized redshift distribution of our lenses, twice and hence the n_ℓ^2 factor. Via a Hankel transform, using J_0, we can then model the final galaxy clustering observable as

$$w(\theta) = \int \frac{\ell d\ell}{2\pi} J_0(\ell\theta) C_{\rm gg}(\ell) \tag{4.63}$$

as a function of angular scale. In figure 4.14 (right) the above model prediction is presented in the two cases of a tophat HOD and when the full 1D stellar-mass distribution for a single lens bin (as mentioned in the caption) is used, which is discussed in detail in Section 4.5. The

difference between the two is more evident on small scales and is small. Given the HOD's flexibility to fit the data, discussed in Section 4.7, as well as the error bars on the data and our constraints, this difference is small enough to allow us to use the tophat model which is more time efficient and can describe the data well. Thus, for our fiducial model we use that HOD model.

4.6.3 Lens magnification

We now consider the effects of weak lensing magnification on the estimation of our observables. In addition to the distortion (shear) of galaxy shapes, weak lensing also changes the observed flux and number density of galaxies – this effect is referred to as magnification. We only consider the magnification in flux for the lens galaxies, as that is the dominant effect in our work.

We first model lens magnification in γ_t. Similar to shear, magnification is expected to be an increasing function of redshift. In the weak lensing regime, the magnification power spectrum involves an integration of the intervening matter up to the lens redshift and is given by (Unruh et al., 2020)

$$\begin{aligned} C_{\rm gm}^{\rm lmag}(\ell) =& \frac{9H_0^3\Omega_m^2}{4c^3}\int dz_\ell n_\ell(z_\ell)\int_0^{z_\ell}\frac{dz}{E(z)}\frac{\chi(z,z_\ell)}{\chi(z)a^2(z)} \\ & \times\int dz_s n_s(z_s)\frac{\chi(z,z_s)}{\chi(z_s)}P_{\rm m}^{\rm nl}\left(\frac{\ell+1/2}{\chi(z)},z\right)\ , \end{aligned} \tag{4.64}$$

where the dimensionless Hubble function is

$$E(z)\equiv\frac{H(z)}{H_0}\ . \tag{4.65}$$

Then, the contribution of this effect to tangential shear can be written as

$$\gamma_t^{\rm lmag}(\theta) = 2(\alpha_{\rm lmag} - 1) \int \frac{\ell d\ell}{2\pi} J_2(\ell\theta) C_{\rm gm}^{\rm lmag}(\ell) \ , \tag{4.66}$$

where $\alpha_{\rm lmag}$ is a free parameter in this work.

We can similarly compute the contribution of lens magnification to galaxy clustering as

$$C_{\rm gg}^{\rm lmag}(\ell) = \frac{9H_0^3\Omega_m^2}{4c^3} \int dz_\ell n_\ell^2(z_\ell) \int_0^{z_\ell} \frac{dz}{E(z)} \left[\frac{\chi(z, z_\ell)}{\chi(z)a(z)}\right]^2 P_{\rm m}^{\rm nl}\left(\frac{\ell + 1/2}{\chi(z)}, z\right) \ , \tag{4.67}$$

where, in contrast to $C_{\rm gm}^{\rm lmag}$, we now have two factors of $n_\ell(z_\ell)$ since the galaxy selection function is used twice. Therefore, the contribution from lens magnification to $w(\theta)$ can be written as

$$w^{\rm lmag}(\theta) = 4(\alpha_{\rm lmag} - 1)^2 \int \frac{\ell d\ell}{2\pi} J_0(\ell\theta) C_{\rm gg}^{\rm lmag}(\ell) \ , \tag{4.68}$$

via a Hankel transform using J_0.

4.6.4 Tidal stripping of the satellites

As tidal disruptions in the outskirts of the host halo strips off the dark matter content of the satellite subhalo, the density profile of the subhalos drops off at large scales. Therefore, this term is usually modeled as a truncated NFW profile (see e.g. Mandelbaum et al., 2005; Velander et al., 2013) which is similar to that of the central 1-halo, $\gamma_t^{\rm c1h}$, out to the truncation radius R and falls off as $\propto r^{-2}$ at larger radii r. For this work, we fix the truncation radius to $R = 0.4R_{200c}$. Additionally, since this is a satellite term, it needs to be multiplied by the satellite fraction $F_{\rm sat}$, and we thus write:

$$\gamma_t^{\rm strip}(\theta) = F_{\rm sat} \times \begin{cases} \gamma_t^{\rm c1h}(\theta) & \text{if } r \leq R \\ \gamma_t^{\rm c1h}(R)\left(\dfrac{R}{r}\right)^2 & \text{if } r > R \end{cases} \ , \tag{4.69}$$

where $r = r(\theta; z_\ell)$ is the radius from the center of the sub-halo at redshift z_ℓ that corresponds to angular scale θ.

4.6.5 Point-mass contribution

This term quantifies the contribution to lensing by the baryonic content of the central galaxy (e.g. Velander et al., 2013). This term is simply modelled as a point-source term:

$$\gamma_t^{\rm PM}(\theta) = \int dz_\ell \; n_\ell(z_\ell) \frac{\langle M_\star \rangle}{\pi r^2(\theta; z_\ell)} \int dz_s n_s(z_s) \Sigma_c^{-1}(z_\ell, z_s) \; . \tag{4.70}$$

In this work, $\langle M_\star \rangle$ is the average stellar mass content in a given lens bin from Table 4.1. Then, the above is more conveninetly written as

$$\gamma_t^{\rm PM}(\theta) = \frac{\langle M_\star \rangle}{\rho_m} \int dz_\ell \; n_\ell(z_\ell) \frac{W_\kappa(z_\ell)}{\pi r^2} \; , \tag{4.71}$$

where the have used

$$W_\kappa(z) \equiv \rho_m \int dz' \; n_s(z') \Sigma_c^{-1}(z, z') \tag{4.72}$$

and corresponds to the lensing efficiency.

4.6.6 Intrinsic alignment

Galaxies are not randomly oriented even in the absence of lensing. On large scales, galaxies can be stretched in a preferable direction by the tidal field of the large scale structure. On small scales, other effects such as the radial orbit of a galaxy in a cluster can affect their orientation. This phenomenon, where the shape of the galaxies is correlated with the density field, is known as *intrinsic alignment* (IA); for a review see Troxel & Ishak (2015).

The contamination of shear by IA can become important when the source galaxies are

physically close to the lenses and gravitational interactions can modify the shape of the galaxies. This term in this work is modeled using the non-linear linear alignment (NLA) model (Hirata & Seljak, 2004; Bridle & King, 2007; Joachimi et al., 2013), modified so that the matter power spectrum is replaced by the galaxy-cross-matter power spectrum so that the small-scale contribution can accounted for. This is the same approach adopted in Zacharegkas et al. (2022); as also noted in that paper, this is not a tested methodology, but rather a trick to accommodate for small scales and a more detailed analysis of small-scale IA is left for future work. Thus, we model the 3D power spectrum of IA as:

$$P_{\rm NLA}(k,z|M_\star) = -A_{\rm IA}C_1\rho_c\Omega_m D_+^{-1}(z)P_{\rm gm}(k,z|M_\star)\left(\frac{1+z}{1+z_0}\right)^{\eta_{\rm IA}} \tag{4.73}$$

where $D_+(z)$ is the linear structure-growth factor at redshift z normalised to unity at $z=0$, $A_{\rm IA}$ determines the overall amplitude, $C_1 = 5\times10^{-14}h^{-2}{\rm M}_\odot^{-1}{\rm Mpc}^3$ is a constant, and the power-law index $\eta_{\rm IA}$ models the redshift evolution defined so that the pivot redshift is set to $z_0 = 0.62$. Then, we model the projected 2D power spectrum for NLA is given, in the Limber approximation, by:

$$C_{\rm NLA}(\ell) = \int dz_\ell \, \frac{n_\ell(z_\ell)n_s(z_\ell)}{\chi^2(z_\ell)(d\chi/dz)|_{z_\ell}} \int dM_\star \, n_\star(M_\star)P_{\rm NLA}\left(\frac{\ell+1/2}{\chi(z_\ell)}, z(\chi_\ell)\middle| M_\star\right) \ , \tag{4.74}$$

where we average over the lens stellar-mass distribution $n_\star(M_\star)$. To obtain the IA contribution to the tangential shear, we perform a Hankel transform on $C_{\rm NLA}(\ell)$ using $J_2(\ell\theta)$, as in Equation (4.50).

4.6.7 Dark matter and satellite galaxy density profiles

For the spatial distribution of dark matter and satellite galaxies within a dark matter halo we have assumed the NFW proflie (Navarro et al., 1996). Thus, their Fourier transform, $u_{\rm dm}$ and $u_{\rm sat}$, respectively, have analytic forms with concentration parameters $c_{\rm dm}$ and $c_{\rm sat}$,

respectively. However, the distribution of satellite galaxies is typically less concentrated than that of the dark matter (e.g. Carlberg et al., 1997; Lin et al., 2004; Nagai & Kravtsov, 2005; Hansen et al., 2005). To account for this, in this work we allow $c_{\rm sat}$ to be smaller than $c_{\rm dm}$ by introducing the free parameter $\kappa_{\rm sat} = c_{\rm sat}/c_{\rm dm}$, which is allowed to take values between 0 and 1, and is expected to not deviate significantly from the unity.

4.7 Parameter inference

In this section we discuss the fitting process of our HOD model, which is discussed in Sections 4.5 and 4.6 to our data, which is thoroughly described in Sections 4.2, 4.3 and 4.4. In our fiducial model, we vary the SHMR parameter vector $\boldsymbol{\theta}_{\rm SHMR} = \{M_{1,0}, \log A_0, \beta_0, \gamma_0\}$, whereas we ignore the redshift-dependent part of these parameters, i.e. $M_{1,1} = A_1 = \beta_1 = \gamma_1 = 0$. The two parameters $(B_{\rm sat}, \beta_{\rm sat})$ that control the satellite SHMR are also free parameters in our fiducial framework. Given that our samples are expected to be complete in our redshift-stellar mass bins, we decided to set the central and satellite parameters $f_{\rm cen}$ and $f_{\rm sat}$, respectively, to unity, and thus we assume $f_{\rm cen} = 1 = f_{\rm sat}$. We also note that from testing the effects from varying these two parameters with priors $\mathcal{U}[0, 1]$ on our data fits we found that they make no significant difference and they are unconstrained. In our HOD model, we furthermore vary the central parameter $\sigma_{\log M_\star}$, the scatter in stellar mass, and the satellite power-law index parameter $\alpha_{\rm sat}$. Our priors on the above parameters are shown in Table 4.2.

We furthermore have three parameters to account for systematic uncertainties (Δz_ℓ^i, Δz_s^j, m_s^j) per lens bin i and source bin j. The *photo-z uncertainty parameters* Δz shift the redshift distributions of lens and source galaxies, and quantify errors in the mean redshift of both samples, while the multiplicative bias parameter is introduced below equation (4.50) in Section 4.6.1. In addition to these, per lens bins we vary the parameter (Σ_ℓ^i) that corresponds to the stretching factors of the lens redshift distribution, similarly to what is discussed in

Porredon et al. (2021). There parameters modify the lens redshift distributions as

$$n_\ell^i(z_\ell) \to \frac{1}{\Sigma_\ell^i} n_\ell^i \left(\langle z_\ell^i \rangle + \frac{z_\ell - \langle z_\ell^i \rangle}{\Sigma_\ell^i} - \Delta z_\ell^i \right) , \tag{4.75}$$

where $\langle z_\ell^i \rangle$ is the mean redshift in lens bin i, and

$$n_s^j(z_s) \to n_s^j \left(z_s - \Delta z_s^j \right) . \tag{4.76}$$

Our priors on the uncertainty parameters are shown in Table 4.2.

Our full data vector consists of the γ_t and w measurements. The full covariance matrix is block-diagonal with the two blocks that correspond to the full covariance matrix for the two observables and the zero matrix as the off-diagonal blocks. We discuss this choice in Section 4.4.2. As our fiducial model in this work we choose to utilize the tophat HOD model of Section 4.5 since, as we discuss in Section 4.6, it is not significantly different from using the full 1D stellar-mass distribution when modeling the observables, but it is a lot more time-efficient. For our model fit we present in Section 4.8 we do not use any of the additional model complexity, beyond the basis HOD model predictions, of Section 4.6.

To sample the posterior of each data set we utilise the MULTINEST[4] sampler, which implements a nested sampling algorithm (see for example Feroz et al., 2009). In our analysis we assume that our data is generated by an underlying Gaussian process, thus making its covariance Gaussian in nature. Therefore, for data vector $\mathbf{d}$ of length N_d and model prediction vector $\mathbf{m}$ of the same length we express the log-likelihood as

$$\ln \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2} (\mathbf{d} - \mathbf{m})^T \mathcal{C}_H^{-1} (\mathbf{d} - \mathbf{m}) \equiv -\frac{1}{2} \chi^2 , \tag{4.77}$$

where $\boldsymbol{\theta}$ is the parameter vector of our model $\mathcal{M}$ and $\mathcal{C}_H^{-1}$ is the Hartlap-corrected data

4. `https://github.com/JohannesBuchner/MultiNest`

Parameter	Prior	Description
$\sigma_{\log M_\star}$	$\mathcal{U}[0.01, 3]$	Central galaxy's stellar-mass scatter; eq. (4.15)
$\alpha_{\rm sat}$	$\mathcal{U}[0.5, 1.5]$	High-mass satellite power-law index; eq. (4.18)
$f_{\rm cen}$	1	Central galaxy number incompleteness; eq. (4.15)
$f_{\rm sat}$	1	Satellite galaxy number incompleteness; eq. (4.18)
$B_{\rm sat}$	$\mathcal{U}[0, 30]$	Satellite SHMR amplitude; eq. (4.19)
$\beta_{\rm sat}$	$\mathcal{U}[0.1, 2]$	Satellite SHMR power-law index; eq. (4.19)
$M_{1,0}$	$\mathcal{U}[9, 14]$	Central SHMR characteristic halo mass; eq. (4.10)
$\log A_0$	$\mathcal{U}[-3, 0]$	Central SHMR overall normalization; eq. (4.10)
β_0	$\mathcal{U}[0.1, 2.5]$	Central SHMR low-mass power-law index; eq. (4.10)
γ_0	$\mathcal{U}[0.1, 1.5]$	Central SHMR high-mass power-law index; eq. (4.10)
Δz^i_ℓ	$\mathcal{N}(0, 0.01)$	Lens photo-z uncertainty; eq. (4.75)
Σ^i_ℓ	1	Stretch parameter of lens $n(z)$; eq. (4.75)
Δz^3_s	$\mathcal{N}(0, 0.01)$	Source photo-z uncertainty in 3rd bin; eq. (4.76)
Δz^4_s	$\mathcal{N}(0, 0.006)$	Source photo-z uncertainty in 4th bin; eq. (4.76)
m^3_s	$\mathcal{N}(-0.0255, 0.0013)$	Multiplicative bias in 3rd source bin; eq. (4.50)
m^4_s	$\mathcal{N}(-0.0322, 0.0118)$	Multiplicative bias in 4th source bin; eq. (4.50)
κ	1	Satellite concentration parameter; sec. 4.6.7
$\alpha_{\rm lmag}$	1	Lens magnification coefficient; eq. (4.6.3)
$A_{\rm IA}$	0	IA amplitude; eq. (4.73)
$\eta_{\rm IA}$	0	IA redshift-dependence; eq. (4.73)

Table 4.2: Priors and fixed value of model parameters. If the prior is flat we present its range in the brackets of the *Uniform distribution*, $\mathcal{U}$, while for Gaussian priors we list the mean and variance in the parenthesis of the *Normal distribution*, $\mathcal{N}$. For fixed parameters we provide their number.

covariance matrix. Notice that we have neglected the constant factors which are not useful while sampling the likelihood.

For our model fits, we report the total χ^2 of our best-fit model to the data, as a measure of the goodness of fit. Alongside this we report the number of degrees of freedom (dof), which we calculate as the effective number of parameters that are constrained by the data, $N_{\rm eff} = N_{\rm par} - {\rm tr}\left[\mathcal{C}_\Pi^{-1}\mathcal{C}_p\right]$, where $N_{\rm par}$ is the total number of parameters we vary, subtracted from the number of data points, N_d:

$$N_{\rm dof} = N_{\rm d} - N_{\rm eff} = N_{\rm d} - N_{\rm par} + {\rm tr}\left[\mathcal{C}_\Pi^{-1}\mathcal{C}_p\right] \ , \tag{4.78}$$

where $\mathcal{C}_\Pi$ and $\mathcal{C}_p$ are the prior and posterior covariance, respectively. We should note here that a goodness-of-fit estimation based on finding an effective number of parameters is not always straightforward when the parameters do not enter the model linearly, as discussed in Section 6.3 of Joachimi et al. (2021). Therefore, our approach of calculating a reduced χ^2 using Equation (4.77) based on the $N_{\rm dof}$ from (4.78) yields a conservative answer if model under-fitting is the main concern.

4.8 Results

In this section we present the first model fits to the new lens galaxy sample we introduced in Section 4.2 from the DES Y3 `GOLD` catalog. The theory predictions are discussed in Sections 4.5 and 4.6. The fiducial framework and fitting procedure we followed to obtain these results are described in Section 4.7. The bins for which we show data fits, as well as the scales we fit, are chosen based on logic presented in Section 4.4.5.

Figure 4.15 shows the best-fit model predictions to galaxy-galaxy lensing. Our model provides a good description of the data as we can see from the plot. While fitting the data, we noticed that how we model the 1-halo to 2-halo regime makes a significant difference in whether we can fit all scales. Specifically, using the prescription from Zacharegkas et al. (2022) the model was unable to fit bins $(\ell, m, s) = (1, 1, 3)$ and $(1, 1, 4)$. However, using the model described in Section 4.6 our HOD framework can describe all features in the data vector well. Moreover, we notice that in many bins the large scales seem to be under-predicted by our model. This might indicate that our transition regime modeling still needs to be modified, or that we need extra model complexity to make the 2-halo and 1-halo terms more compatible with each other. Especially lens magnification could be of significant help as it would add power to the large scales. We also note, however, that large scales are the most correlated which would make the act more as a "single data point" and thus not contribute significantly to the overall χ^2, even if we might expect them to by eye. We also note here

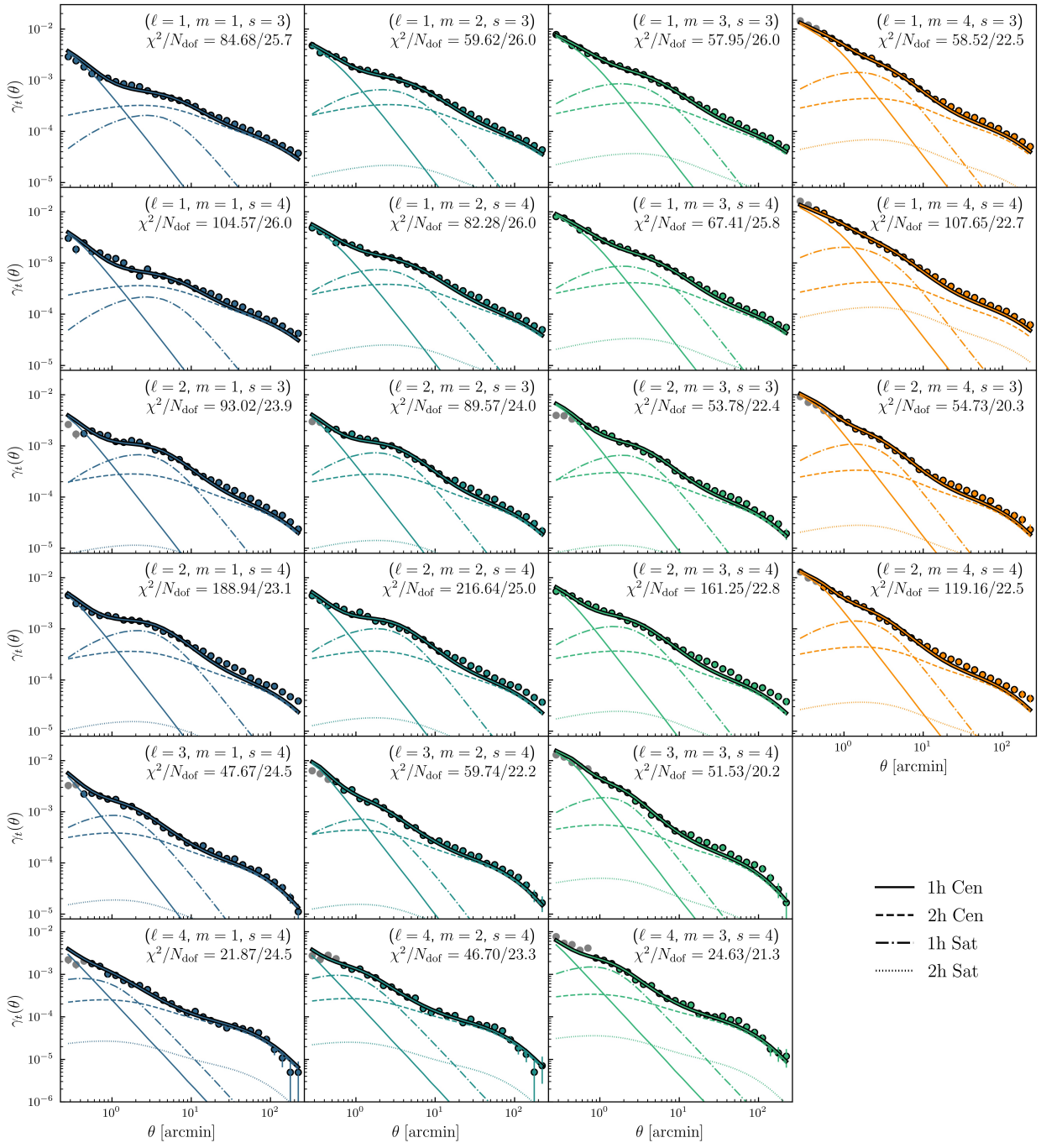


Figure 4.15: Measurements (points) and best-fit model (thick solid outlined) predictions for galaxy-galaxy lensing as a function of angular scale in arcmin. Each panel corresponds to a lens-source bin combination from the ones that we choose to model in this work, as indicated in parenthesis (ℓ, m, s). Each model component is also shown: 1-halo central (1h Cen, solid), 2-halo central (2h Cen, dashed), 1-halo satellite (1h Sat, dash-dotted) and 2-halo satellite (2h Sat, dotted). Gray points are removed from the fits by the scales cuts. The χ^2 goodness-of-fit is shown in each panel and the number of degrees of freedom is also reported.

that the χ^2 values reported in the figure should act as a general guide when determining how good the fits are and we do not expect perfect values, i.e. close to unity for the reduced chi-square $\chi^2_r \equiv \chi^2/N_{\rm dof}$. Interestingly, the highest redshift, and more massive, galaxies in redshift bins $\ell = 4$ are the best described by the model. In these bins, however, since they are so high-z, the 1-halo points are not "as visible", since the same angular scale corresponds to larger physical scales as redshift increases, and thus the compatibility between 1-halo and 2-halo is not as much of an issue there, which would explain why the fits look so much better. This is also evident by the fact that the 1-halo central term contributes less in the fits as we move to higher redshift bins, i.e. from top to bottom panels in figure 4.15.

Figure 4.16 presents the best-fit model predictions to galaxy clustering. The overall same comments about the goodness-of-fit to w as in the case of γ_t apply: the model is able to describe the data very well and the χ^2 should we used as a general guide in deciding how well the fits work. Moreover, we see again that, as we move from top to bottom in the figure, i.e. from lower to higher redshift, the 1-halo terms, and in particular the satellite-satellite term, contributes less for the same reason as in lensing. In contrast to γ_t, though, we do not see the fits to work better in the high-z bins, as compared to lower-z ones. The fits in this case seem to be more consistent throughout the various redshift bins. This might be a further indication that additional model complexity is needed in the case of galaxy-galaxy lensing for the model fits to improve in figure 4.15.

The data fits we have discussed so far have been separately done on the two observables. Our main goal, however, is to fit them jointly in order to constrain the SHMR optimally. In trying to do so, we found that small scales in galaxy clustering suffer from systematic effects, of unknown origin at this time, that lead to an overestimation on the measured signal. As a result, we cannot obtain good joint fits to both γ_t and w when small scales in w are included. We are therefore forced to remove scales of $\theta < 10$ arcmin from our w part of the data vector and thus only fit the 2-halo term in galaxy clustering. Still, given that w at those scales is

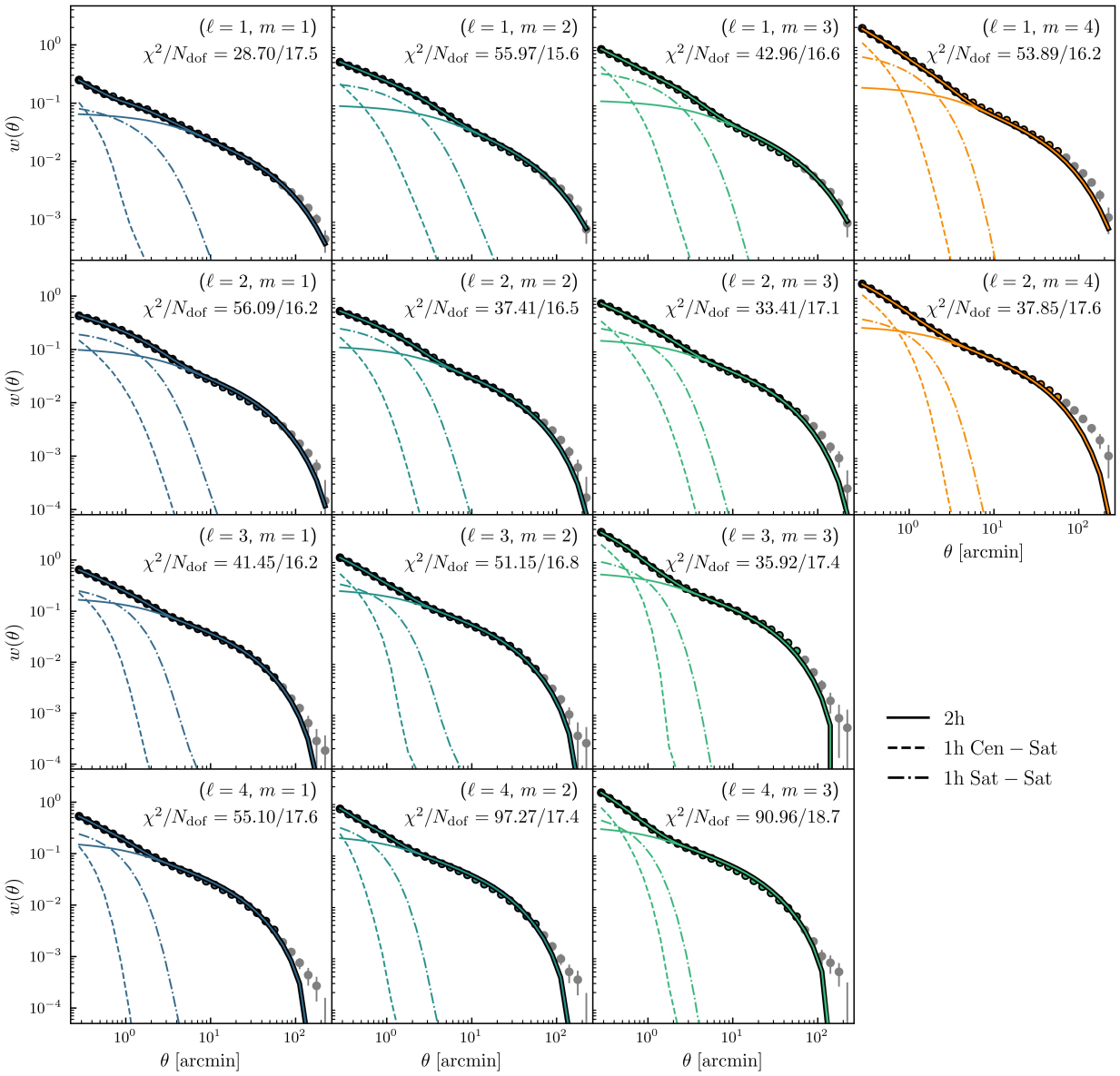


Figure 4.16: Measurements (points) and best-fit model (thick solid outlined) predictions for galaxy clustering as a function of angular scale in arcmin. Each panel corresponds to a lens bin from the ones that we choose to model in this work, as indicated in parenthesis (ℓ, m). Each model component is also shown: 2-halo (2h, solid), 1-halo central-satellite (1h Cen − Sat, dashed) and 1-halo satellite-satellite (1h Sat − Sat, dash-dotted). Gray points are removed from the fits by the scales cuts. The χ^2 goodness-of-fit is shown in each panel and the number of degrees of freedom is also reported.

$\propto \bar{b}_g^2$, which is a prediction from the HOD model, it provides additional constraining power when combined with the γ_t fits. A more detailed analysis on what is causing the small-scale discrepancy between the two observables is ongoing and more careful scale cuts on w on small scales as a function of redshift will be performed in the near future. After we removed the small-scale w data we are able to successfully fit γ_t and w together and examples in two lens bins are presented in figure 4.17. In that plot, only the 2-halo term in w is considered while all term are included during the fitting in γ_t.

It is important to note here that even though small-scale measurements in clustering seem to be overestimated, and thus it is in tension with lensing, our HOD framework is flexible enough to still be able to fit w on all scales when it is done separately from γ_t, as we see from figure 4.16. This is an indication that we need to be careful when we interpret the results from fitting our models to data and that we need to combine data that breaks degeneracies in the parameter space. To emphasize this even more, we found that when we allow the stretch parameter, Σ_ℓ, in the lens redshift distributions (see also table 4.2) with wide prior range we were still able to obtain good joint data fits over all angular scales. The inferred mass, however, was significantly miss-estimated. In that case, Σ_ℓ was constrained to values much less than unity, meaning that the lens $n(z)$ distributions were a lot narrower that what is shown in figure 4.5 and thus γ_t and especially w were boosted high as a result. The rest of the parameters were then forced to re-adjust so that the overall χ^2 was able to improve significantly. However, the resulting halo mass constraints were unphysically high. As a final note, while jointly fitting the observables in the 2-halo regime, we find no tension between the two, as also indicated by figure 4.17, and even if we allow Σ_ℓ to vary it does not prefer a value different from unity.

We present constraints on the stellar-to-halo mass relation as inferred from our data fits in figure 4.18. Our results when they are inferred from our γ_t-only fits to the data over all scales, which are shown in figure 4.15, are shown as outlined circles. In addition, we present

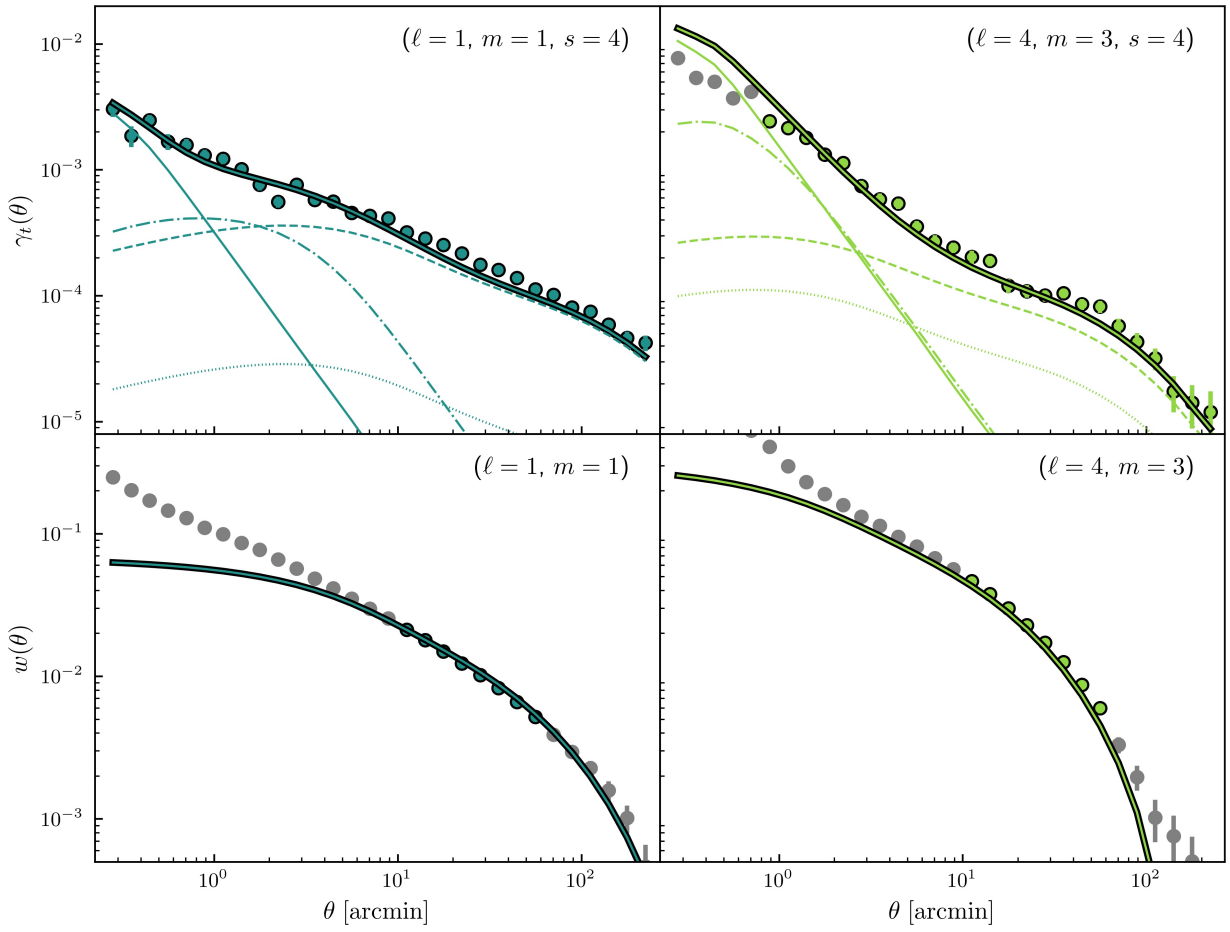


Figure 4.17: Measurements (points) and best-fit model (thick solid outlined) predictions for galaxy-galaxy lensing (top panels) and galaxy clustering (bottom panels) as a function of angular scale in arcmin. These fits were performed jointly for γ_t and w for single lens bins and each column shows the best-fit model prediction for the two observables, respectively. For γ_t we fit all scales after the initial scale cuts are applied, as discussed on Section 4.4.5, whereas for w we in addition remove the 1-halo regime as we discuss in Section 4.8. Thus, for γ_t panels show all components in addition to the total model prediction, exactly as in figure 4.15 while for w only the 2-halo component is plotted.

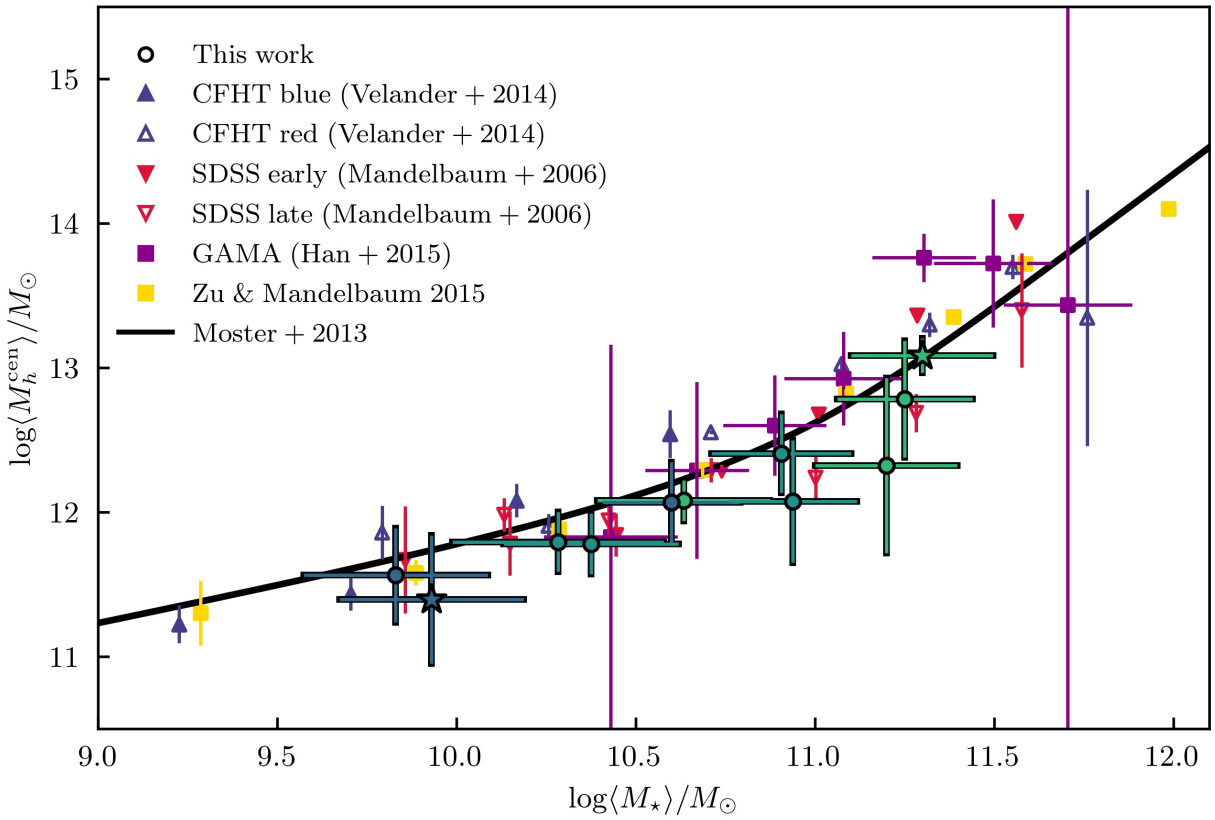


Figure 4.18: Constraints on the stellar-to-halo mass relation from our analysis (outlined points from γ_t-only fits and stars from jointly fitting γ_t and w) compared to constraints from the literature, as indicated in the legend. The bars on stellar mass (on the horizontal axis) correspond to the range in stellar mass per bin we that fit, whereas the error bar on halo mass (on the vertical axis) shows the uncertainty on our inference from the data fits. The black solid line is produced using the fitting functions from Moster et al. (2013) as a means to guide the eye.

the halo-mass constraints from the joint γ_t and w fits in figure 4.17 as outlined stars. We compare our results with independent constraints from the literature, as well as with the prediction from the Moster et al. (2013) model, and we find good overall agreement. The filled and open blue triangles present the results from CFHTLenS (Velander et al., 2013), were the samples they considered were flux-limited and were split into blue and red galaxies, respectively. The filled and open red triangles in the same plot present the constraints from SDSS (Mandelbaum et al., 2006a) and galaxies are, respectively, split into early and late types. The magenta squares show the average halo masses inferred from individual galaxy stellar mass samples using the weak lensing measurements in GAMA (Han et al., 2015) were the central galaxies of groups and clusters were almost volume-limited. We finally also compare our results with those from Zu & Mandelbaum (2015) were a HOD model similar to this work is used to jointly fit galaxy-galaxy lensing and galaxy clustering in SDSS. All constraints show a remarkable agreement with each other. We will present our final results in Zacharegkas et al. (2023), in prep. in the near future.

4.9 Summary and discussion

In this work we have produced a new lens galaxy sample from the DES `Y3 GOLD` that includes the redshift, stellar mass and color information of each individual galaxy. In addition, by using the software `GalPro` to do so, we were able provide the full 2D joint redshift-stellar mass distributions per galaxy and for all galaxies combined in a single bin. We have run a number of tests on this sample and we have validated our measurements on galaxy-galaxy lensing, γ_t, and galaxy clustering, w. This sample can be used in numerous studies in the future, within and outside the DES collaboration, when the stellar mass information is of interest. In this study, we have utilized our sample in order to bin galaxies into redshift and stellar mass in order to perform a small-scale HOD analysis on it to constrain the galaxy-halo connection of the galaxies in DES Y3.

In order to analyze the aforementioned stellar-mass sample, we developed an HOD model that is an update of the Zacharegkas et al. (2022) so that it can account for the stellar-mass information of the galaxies and to further incorporate $w(\theta)$ into the new framework. Importantly, this new HOD model depends on the stellar-to-halo mass relation and through fitting the data we were able to constrain that relation for our lenses. Our findings so far are the following:

- We can separately fit both observables, γ_t and w, over all scales we consider well.
- From fitting γ_t alone we can obtain good constraints on the the average halo mass in a $z - M_\star$ bins and thus on the SHMR. However, we have indications that additional model complexity in γ_t might be needed to capture all features of the data better.
- On the other hand, w alone cannot provide good SHMR constraints as the inferred mass when small scales are included is overestimated.
- In order to obtain good data fits to all scales, especially in low-stellar-mass bins, we needed to update how we model the 1-halo to 2-halo transition scales compared to what was utilized in Zacharegkas et al. (2022).
- We have concluded that there are systematic effects on the very small scales in w, at $\theta \lesssim 10$ arcmin and we thus remove those scales while jointly fitting both observables.
- The joint $\{\gamma_t, w\}$ data fits after small w scales are removed are successful in describing the data well and the SHMR from them are in agreement with the γ_t-only fits.
- Our final SHMR constraints are is good agreement with a large number of independent studies in the literature.

Moving forward, in the upcoming publication Zacharegkas et al. (2023), in prep. we will perform more careful small-scale cuts on w, and we will present SHMR constraints from fully

joint fits to the data after we have also incorporated the additional model complexity into our γ_t model. By tuning the scale cuts on galaxy clustering we might be able to probe the quasi-linear regime in w, or even the 1-halo scales, and thus further improve our final constraints. The additional complexity in γ_t will lead to an improved fit to the measurements were large and small scales show even better consistency. Furthermore, we will present results from fitting our model to galaxies after we split them into red and blue subsamples, as we have their color information as well.

This work is a major step towards the direction of performing cosmological analysis using data from galaxy surveys utilizing large scales, which have been commonly considered in the past in such studies, and the non-linear and quasi-linear regime simultaneously. This might further improve the constraints on the cosmological parameters, and will provide a more complete and unified picture of the large-scale structure of the Universe. In addition, our model and its implementation provides a galaxy-halo connection modeling framework that can be easily improved and expanded further in order to incorporate more of the lens galaxy properties, beyond stellar mass and redshift.

CHAPTER 5

APPLICATIONS OF THE GALAXY-HALO CONNECTION MODEL IN THE KEY COSMOLOGICAL ANALYSES IN DES

In this chapter we discuss briefly some applications of the galaxy-halo connection that was developed in Zacharegkas et al. (2022) to model the non-linear scales in galaxy-galaxy lensing, as described in Section 3. The two applications that follow make it clear that modeling those small scales is not only important for improving cosmological constraints and understanding the physics of the relation between galaxies and their host dark matter halos, but also for performing necessary tests on the data. These tests can be crucial in determining, for instance, the scale cuts that we need to apply to the data if small scales are removed so that the final constraints are unbiased.

5.1 Galaxy-galaxy lensing in DES Y3 and Y6

The principles and theory behind galaxy-galaxy lensing (the weak gravitational lensing of light emitted by a source galaxy by the dark matter halo of a lens galaxy) is discussed in Section 1.4.1. As mentioned in Chapter 2, this observable plays a key role in cosmological analyses, as it is very sensitive to cosmology, and combined with galaxy clustering and cosmic shear, it is used by the DES collaboration to derive their main results. Therefore, a major effort within the collaboration is to produce high-quality measurements of GGL, $\gamma_t(\theta)$, as a function of angular separation from the center of a lens, and to test (and correct) for systematics. This work in DES Y3 is fully described in Prat et al. (2022), and so, here we will only focus on some key applications from that study relevant to this thesis.

The details of how we measure GGL are discussed in Chapters 3 and 4, and, for the most part, they are the same as what was done in Prat et al. (2022). However, in these chapters we were not interested in scale cuts, as our focus was to include small scales into

our analyses. In contrast, Prat et al. (2022) is part of the fiducial cosmology work in the DES and is thus highly interested in finding the best possible scale cuts so that the maximum amount of data is utilized while the scales where the linear theory breaks down are removed. Below we briefly describe the idea of *contaminated simulated data vectors* and how they are used in scale-cut tests.

Figure 5.1, from Prat et al. (2022), shows a grid of panels of differences between GGL simulated data vector minus the data vectors contaminated by various terms, for all lens bins cross-correlated with the source bins (in the upper right corner the numbers indicate the $i-$th lens bins comma-separated from the $j-$th source bins). Therefore, it essentially shows how significant some effects, mostly relevant for small scales, are per lens-source bins combination. The terms tested in the figure are the following.

- The effects from using a *non-linear bias model* (orange dashed line): A non-linear model of galaxy bias can push the analysis to smaller scales compared to a linear model; on large scales the two are identical, but on smaller scales the non-linear model differs a lot from the linear one. This is evident from the large deviation of the contaminated vector compared to the fiducial one as we move to small angular separations. Note here, however, that although a non-linear galaxy bias model is a better description of the data on some of the scales in the quasi-linear regime, it still breaks down on sufficiently small scales. Finally, we note that, since the non-linear galaxy bias model accounts for the non-linear nature of the halo model, compared to the fiducial data vector, on small scales it predicts a higher GGL signal.

- Effects from *baryonic feedback* (gold dashed line): Baryons affect the contaminated data in the opposite direction from the non-linear galaxy bias modeling, as seen from the figure. As we can see (and as one would expect), this term mostly affects the very small scales and since it has to do with the feedback from baryonic matter, it predicts a lower GGL amplitude compared to the fiducial vector.

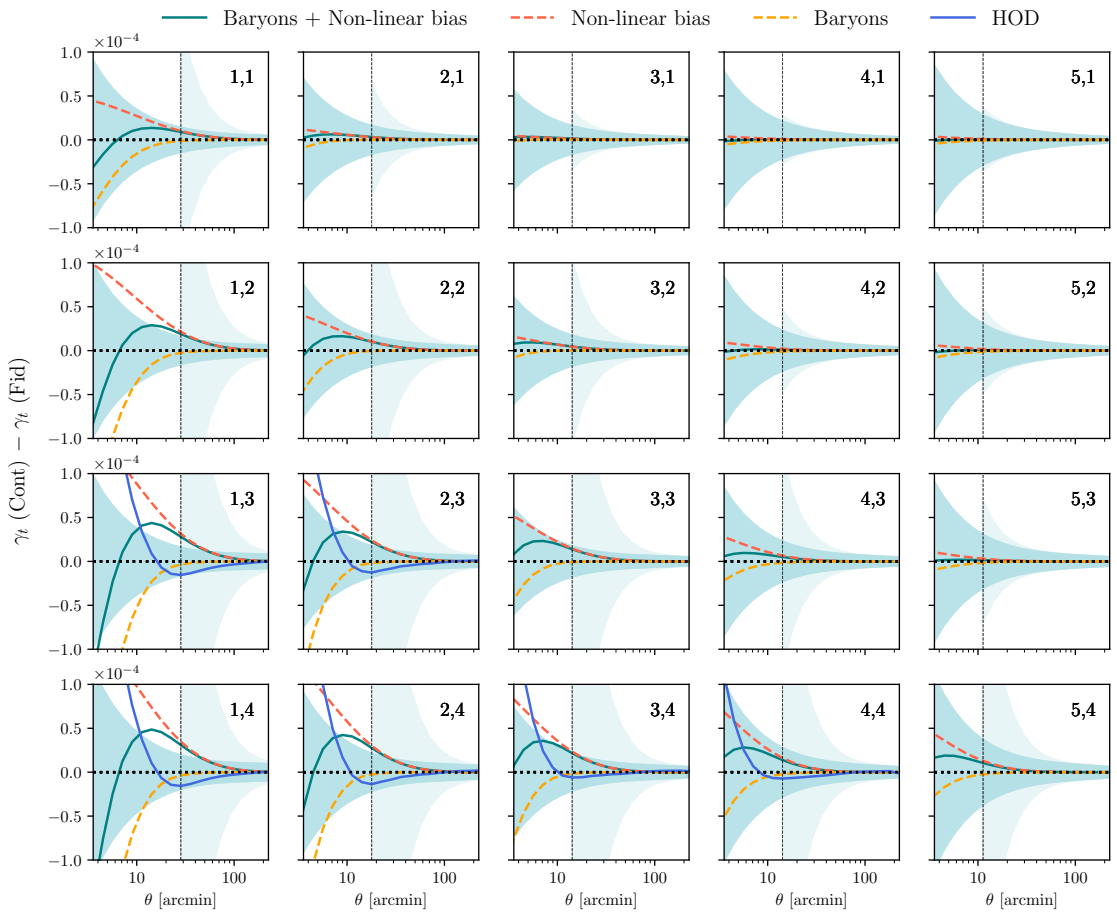


Figure 5.1: Difference between a simulated datavector contaminated with baryonic effects and non-linear galaxy bias with respect to the fiducial model (linear bias and `Halofit` non-linear matter power spectrum), and the equivalent difference for an HOD contaminated datavector using the model and results from Zacharegkas et al. (2022) – see Chapter 3. The darker blue shaded regions indicate the uncertainties coming from the theory covariance without point-mass marginalization and the lighter ones including the point-mass marginalization. The dashed vertical black lines indicate the 6 Mpc/h scale cuts. This figure is taken from Prat et al. (2022).

- Combining *non-linear galaxy bias and baryons* (cyan solid line): In order to make conclusions regarding scale cuts, these two effects are combined and the net effect on the fiducial data vector are seen in the figure. As a result of this test overall, the final verdict is that the modeling below the scale cuts, as shown the figure, cannot be trusted to fit the data.

- Testing a *Halo Occupation Distribution (HOD)* model (solid blue line): To further test how "wrong" the assumption of a linear bias model is on small scales, it makes sense that we would want to compare the fiducial data vector with a galaxy-halo model, which is the best possible way of modeling the non-linear regime. As we see from the figure, the HOD contamination is large on the small scales and identical to the linear theory on large scales, like we would expect. We note here, however that, to somewhat of a surprise, the HOD contamination is overall not as large as the net one from non-linear galaxy bias modeling and baryonic effects. This led to the additional conclusion that using the latter to derive the fiducial scale cuts is the conservative way to go. To conclude, we note here for completeness that the HOD effect was only tested on the 6 lens-source bin combination for which we had HOD fits (Zacharegkas et al., 2022), as presented in Chapter 3.

To conclude this section, we note that for DES Y6 a very similar approach to measuring and testing galaxy-galaxy lensing will be taken. As this work in under way, we will not show any results on this in this thesis. We will remark, however, that the galaxy-halo connection model will again serve as a powerful way to perform tests on the measurements, and it could be a key component of the methodology to determine the fiducial model in the main 3×2pt analyses. This further highlights the increasing importance of developing models for the non-linear regime.

5.2 Small-scale information from shear ratio in DES Y3

The idea of taking ratios of tangential shears between cross-correlations of different source bins, fixing the lens bins, is called *shear ratio*. As a brief note, taking ratios of quantities like shears is not unique to the applications we are discussing in this section, as they are a powerful way of constraining cosmology from geometrical quantities and the idea goes back to Jain & Taylor (2003). The point here is that, the shear ratios in this case can be used to extract information from scales much smaller than the fiducial scale cuts used in the DES Y3 3×2pt analyses that are discussed in Section 5.1. That is because, by taking ratios of two shears, in the approximation above, the galaxy bias and even the dependence on the matter power spectrum are almost exactly canceled out. Note, however, that this is true when we use the linear theory to model γ_t while taking the ratios, and thus we still cannot go down to the small non-linear scales without the usage of a galaxy-halo connection model.

The application we will talk about in this section is from Sánchez et al. (2022a), which is an application to the DES Y3 data. The theory of shear ratios is briefly described below. Lets consider lens galaxy ℓ at redshift z_ℓ and two source galaxies s_i and s_j at redshifts z_{s_i} and z_{s_j}, respectively; this ratio will be labeled henceforth as (ℓ, s_i, s_j). If we take the ratio of tangential shear of the galaxy pair (ℓ, s_i) divided by the pair (ℓ, s_j), the shear ratios can be modeled independently from angular scale. In particular, the ratio of shears is approximately written as:

$$\frac{\gamma_t^{(\ell,s_i)}}{\gamma_t^{(\ell,s_j)}} \approx \frac{\Sigma_c(z_\ell, z_{s_j})}{\Sigma_c(z_\ell, z_{s_i})} \,, \tag{5.1}$$

where the critical surface density is defined in Section 1.4.1. Then, we can extend this to taking ratios of average shears in cross-correlations of tomographic redshift bins, in the *thin-bin approximation*, integrating over the redshift distributions of the lenses, $n_\ell(z_\ell)$, and

sources, $n_s(z_s)$, which yields:

$$\frac{\left\langle \gamma_t^{(\ell,s_i)} \right\rangle}{\left\langle \gamma_t^{(\ell,s_j)} \right\rangle} \approx \frac{\Sigma^{-1}_{c,\mathrm{eff}(\ell,s_i)}}{\Sigma^{-1}_{c,\mathrm{eff}(\ell,s_i)}} \,, \tag{5.2}$$

where we have defined the *effective critical surface density* though its inverse as:

$$\Sigma^{-1}_{c,\mathrm{eff}(\ell,s)} \equiv \int dz_\ell \; n_\ell(z_\ell) \int dz_s \; n_s(z_s) \Sigma^{-1}_{c,\mathrm{eff}}(z_\ell, z_s) \; . \tag{5.3}$$

This approximation is only good enough when the redshift bins are sufficiently narrow. They are useful, however, in demonstrating that the shear ratios are only dependent on the redshift bins and geometrical factors. This model is what was used in DES Y1 (Prat et al., 2018); for the complete model that was instead used in DES Y3, which accounts for more terms, see Sánchez et al. (2022a).

In the full shear ratio modeling in DES Y3, the specifics of galaxy bias needed to be taken into account. In the fiducial model of shear ratios, the non-linear matter power spectrum is considered, which used the `Halofit` approach (see also discussion in Section 1.2.2). As we also mentioned before, shear ratios push the analysis to smaller scales (and thus the need of a non-linear matter power spectrum), and thus it is worth to compare the fiducial results with the predictions from a galaxy-halo connection model. For that, we used the results from the HOD analysis in Zacharegkas et al. (2022), and the results are shown in figure 5.2. The comparisons in this figure are limited to only the few bin combinations considered in that work, as presented also in Chapter 3. This test serves as a "robustness test" to assess the applicability of the `Halofit` model. The shear ratios in that figure are simulated using the `MICE` N-body simulations.

The shear ratio tests in figure 5.2 are split into large (orange points) and small (cyan points) scales. In the top panel, where the HOD is fixed within each bin, on large scales the

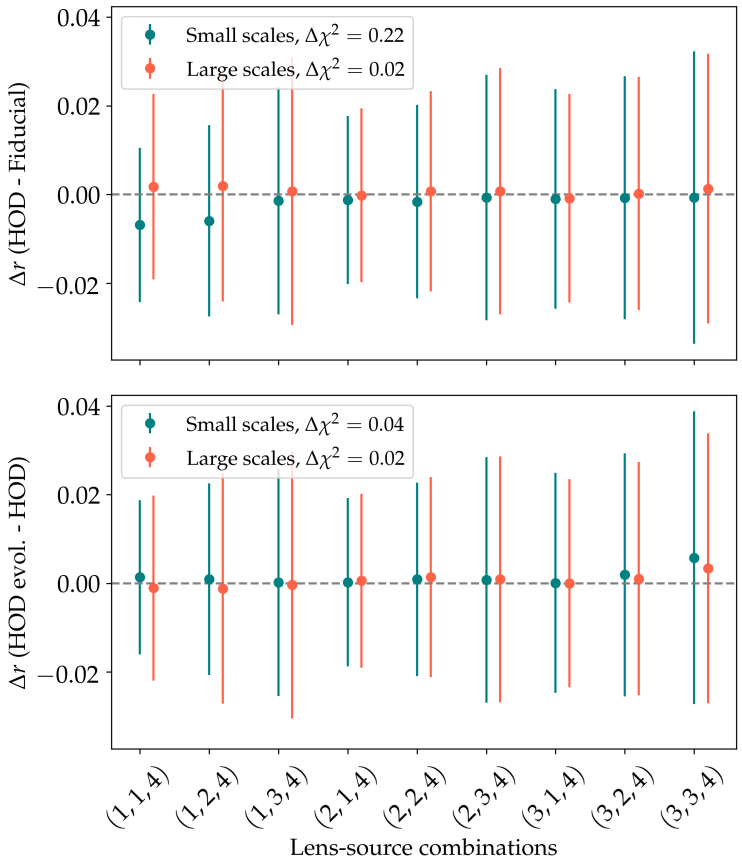


Figure 5.2: Effects of HOD modeling (Zacharegkas et al., 2022) and HOD evolution on the shear ratios, for both small and large angular scales. The error bars show the ratio uncertainties from the same covariance as used in the data. The comparison between the fiducial model and the HOD fixed inside a bin is shown in the top panel, while the comparison with the HOD that evolves with redshift within a bin is shows in the bottom panel.

`Halofit` approach is practically identical to the HOD one, as expected. On the small scales, there is a small deviation, however it is not significant compared to the error bars. The same is true for the second case, in the lower panel, where the differences are even smaller. Therefore, it is concluded that the `Halofit` approach is sufficiently good in describing the shear ratios on the scales considered in the analysis, at least when compared to the prediction from the HOD model.

CHAPTER 6

METHODS OF MITIGATION OF SMALL-SCALE CONTAMINATION IN GALAXY-GALAXY LENSING

In this section we present the work in Park et al. (2021). In that study, we compared and developed transformations to data from the DES Y3 that remove any contamination on the linear scales, which are utilized in the key cosmological analyses, from the non-linear scales. This is important in order to make sure that the final constraints on cosmology are not biased due to the contamination. By performing a careful comparison of various methods which do that, commonly employed in the literature, we provide recommendations as to which method is best to be utilized by future analyses on data sets even more constraining than DES Y3. To do so, we perform simulated data vector tests under a LSST Y1 setup.

6.1 Methods of mitigation of small-scale contamination in galaxy-galaxy lensing

Since in Cosmology an adequate small-scale model for galaxy-galaxy lensing, e.g. as discussed in Chapters 3 and 4, is not always available or when for various reasons such models are not utilized, we need to make sure that whenever we perform an analysis that includes this 2-point statistic we at least correctly account for the contribution from these scales in order to avoid biases in our results. This is the subject of this chapter: given the non-local nature of the galaxy-galaxy lensing estimators, how can one perform transformations on the data to remove the "contamination" from small scales in the larger scales that we consider in an analysis? We discuss this in detail in Prat et al. (2023) and in what follows below.

6.1.1 Introduction

When the light of background (*source*) galaxies passes close to foreground (*lens* or *tracer*) galaxies it gets perturbed, distorting the image of the source galaxies we observe. Galaxy-galaxy lensing refers to the cross-correlation between source galaxy shapes and lens galaxy positions. The amount of distortion is correlated with the properties of the lens sample and the underlying large-scale structure it traces. In the case of a spherical distribution of matter, the shear at any point will be oriented tangentially to the direction toward the center of symmetry. Then, the tangential component of the shear will capture all the gravitational lensing signal produced by a spherically symmetric distribution of mass. Because of this, the *tangential shear* averaged over many source-lens galaxy pairs is one of the typical measurements that is done to detect this correlation.

Galaxy-galaxy lensing has had a wide range of applications since it was first detected in Brainerd et al. (1996). It has been extensively used to probe the galaxy-matter connection at small scales, e.g. Choi et al. (2012), Yoo & Seljak (2012), Kuijken et al. (2015), Clampitt et al. (2017), Park, Krause et al. (2016), Zacharegkas et al. (2022) or Luo et al. (2022); to extract cosmological information using the well-understood large scales in combination with other probes such as galaxy clustering and/or CMB lensing as in Mandelbaum et al. (2013), Baxter et al. (2016), Joudaki et al. (2018), van Uitert et al. (2018), Prat, Sánchez et al. (2018), Singh et al. (2020), Lee et al. (2022), Prat et al. (2022); to obtain lensing shear geometric constraints e.g. Jain & Taylor (2003), Prat, Baxter et al. (2019), Hildebrandt et al. (2020), Giblin et al. (2021), Sánchez, Prat et al. (2022a), and also recently in Leauthaud et al. (2022) to assess the consistency of lensing across different data-sets and to carry out end-to-end tests of systematic errors.

Moreover, galaxy-galaxy lensing is a standard part of the so-called 3×2pt analyses that combine large-scale structure and weak lensing measurements to extract cosmological information. The 3×2pt stands for the combination of three two-point correlation functions: the

autocorrelation of the positions of galaxies (galaxy clustering), the cross-correlation of galaxy shapes and galaxy positions (galaxy-galaxy lensing) and the autocorrelation of galaxy shapes (cosmic shear). This combination was originally proposed in Hu & Jain (2004), followed by Bernstein (2009) and Joachimi & Bridle (2010) and has since then been applied to different galaxy survey data sets, such as to KiDS data, as in Heymans et al. (2021), and to DES data, e.g. in Abbott et al. (2022).

These kind of analyses commonly use the well-understood large scales, placing stringent scale cuts to remove the parts of the data vector that currently add too much uncertainty in the model due to non-linearities in the matter power spectrum, galaxy bias and baryonic effects, amongst others (e.g. Krause et al. 2021a). However, the galaxy-galaxy lensing signal is non-local in nature, that is, the predicted signal at a given separation between a source and a lens galaxy (at the redshift of the lens galaxy) depends on the modeling of all scales within that separation, including the non-linear small scales. This can be appreciated expressing the tangential shear of a single lens-source galaxy pair separated by an angular distance θ as a function of the excess surface mass density $\Delta\Sigma$:

$$\gamma_t\left(\theta, z_l, z_s\right) = \frac{\Delta\Sigma\left(\theta\right)}{\Sigma_{\mathrm{crit}}(z_l, z_s)}, \tag{6.1}$$

where $\Delta\Sigma$ can be expanded as the difference between the mean surface mass density *below* a certain angular scale ($< \theta$) and the surface mass density *at* this given scale θ:

$$\Delta\Sigma\left(\theta\right) = \overline{\Sigma}\left(< \theta\right) - \Sigma\left(\theta\right), \tag{6.2}$$

where the non-locality of the tangential shear quantity becomes apparent, since the tangential shear defined at some θ value will always carry information of all the scales below this value. Σ_{crit} is just a geometrical factor that depends on the angular diameter distances to the lens galaxy D_l, the one between the lens and the source D_{ls} and the one to the source galaxy

D_s, and is defined as:

$$\Sigma_{\rm crit}^{-1}(z_{\rm l}, z_{\rm s}) = \frac{4\pi G}{c^2}\frac{D_{ls}\, D_l}{D_s} \qquad \text{if} z_s > z_l, \tag{6.3}$$

and zero otherwise. In the equation above G is the gravitational constant and c is the speed of light.

Therefore, our inability to model the small scales accurately enough given the measurement uncertainties impacts the lensing signal at all scales. Expanding on this, in order to predict the lensing signal, an accurate prediction for the galaxy-matter correlation function $\xi_{\rm gm}(r)$ is required for some range of physical (3D) scales r, see e.g. MacCrann et al. (2020a). At large scales, we expect linear theory to hold and thus we can relate the galaxy-matter correlation function with the matter correlation through a linear galaxy bias factor: $\xi_{\rm gm}(r) = b\xi_{\rm mm}$ (see Desjacques et al., 2018, for a galaxy bias review). At smaller (nonlinear) scales we do not currently have a precise theory to model $\xi_{\rm gm}(r)$. Therefore, it is crucial to ensure that the tangential shear measurement is only sensitive to scales in $\xi_{\rm gm}(r)$ where we know that the modelling is sufficiently accurate. Since the galaxy-galaxy signal receives a non-local contribution that depends on scales in $\xi_{\rm gm}(r)$ that are much smaller than the separation at which the measurement is made (i.e. the impact parameter in the lens redshift), this non-locality can then force the scale cuts applied in real data to be significantly larger than the scale at which theoretical uncertainties become problematic. For example, due to this reason, the scale cuts in the DES Y1 3×2pt cosmological analysis (Abbott et al., 2018) were higher for the galaxy-galaxy lensing part (12 h^{-1}Mpc) than for the galaxy clustering part (8 h^{-1}Mpc).

To help with this issue, there have been a few independent efforts to mitigate the non-locality of the galaxy-galaxy lensing signal. The first that was proposed was the annular differential surface density estimator by Baldauf et al. (2010), which consists of modifying the data vector on all scales in a way that removes information about the lens mass dis-

tribution below a chosen scale R_0, based on the measured data vector around R_0. Later on MacCrann et al. (2020a) proposed a methodology to take into account the non-locality by analytically marginalizing over an enclosed point-mass directly when performing the cosmological parameters inference. Finally another estimator-based methodology was proposed by Park et al. (2021), which achieves the localization of the galaxy-galaxy lensing signal by performing a linear transformation of the tangential shear quantity. In the recent DES Y3 3×2pt work (Abbott et al., 2022), the point-mass marginalization methodology was applied to remove the information from small scales above a certain scale, which resulted in being able to model the galaxy-galaxy lensing until 6 Mpc/h, a much smaller scale cut than the one used in DES Y1 of 12 Mpc/h, even with smaller measurement uncertainties, while the galaxy clustering scale remained the same as in the Y1 analysis (8 h^{-1}Mpc).

In this paper we perform a thorough comparison of these different proposals with the main goal of understanding which of them is advantageous to use in future cosmological analyses. First, we use simulated data vectors with uncertainties mimicking the Rubin Observatory Legacy Survey of Space and Time (LSST) Y1 settings to perform such a comparison and then apply it to DES Y3 data.

6.2 Theory

Eq. (6.2) can be rewritten as a function of physical scale $R = \theta D_l$ in the small angle approximation as:

$$\Delta\Sigma(R) = \overline{\Sigma}\,(0, R) - \Sigma\,(R). \tag{6.4}$$

Then, expanding each of the terms, the surface density at a given transverse R scale between the lens galaxy and the source light can be expressed as the integral of the three-dimensional galaxy-matter correlation function $\xi_{gm}(r)$ over the line-of-sight distance Π, with

$r = \sqrt{\Pi^2 + R^2}$:

$$\Sigma(R) = \overline{\rho}_m \int_{-\infty}^{\infty} \mathrm{d}\Pi \left[1 + \xi_{gm}\left(\sqrt{\Pi^2 + R^2}\right)\right] , \tag{6.5}$$

where $\overline{\rho}_m$ is the mean matter density. The other term in Eq. (6.4) is the mean surface density between two transverse positions, which can be generally expressed as:

$$\overline{\Sigma}(R_1, R_2) = \frac{2}{R_2^2 - R_1^2} \int_{R_1}^{R_2} \Sigma(R') R' \mathrm{d}R'. \tag{6.6}$$

For $R_1 = 0$ and $R_2 = R$ it simplifies to:

$$\overline{\Sigma}(0, R) = \frac{2}{R^2} \int_{0}^{R} \Sigma(R') R' \mathrm{d}R'. \tag{6.7}$$

This term is the one containing the information from all scales down to $R = 0$, including the one halo regime for which we do not have an accurate model for ξ_{gm}. Assuming we can model ξ_{gm} accurately only down to some minimum scale $r_{\rm min}$, we do not want $\Delta\Sigma(R)$ to depend on ξ_{gm} below $r_{\rm min}$. From the equation above it becomes clear that $\overline{\Sigma}(0, R)$, and thus $\Delta\Sigma(R)$, are only sensitive to the *total* mass enclosed inside R (i.e. to some integral of ξ_{gm}), but not to how the mass is distributed (i.e. to the shape of ξ_{gm}). For instance, $\Delta\Sigma$ at scales larger than R will be the same for a point-mass distribution as for an NFW profile. Also, as shown by Eq. (6.7), the contribution from the total enclosed mass will scale as $1/R^2$. This is the key fact that all the estimators described below rely on to remove the dependency of the enclosed mass (or non-locality) of the galaxy-galaxy lensing measurements. Below we summarize each of these currently existing methodologies. We also visualize the modified data vectors for each of the methodologies in Fig. 6.1.

6.2.1 Annular Differential Surface Density Υ

The Annular Differential Surface Density statistic, $\Upsilon(R)$, is defined in the following way (Baldauf et al., 2010):

$$\Upsilon(R;R_0) \equiv \Delta\Sigma(R) - \frac{R_0^2}{R^2}\Delta\Sigma(R_0) = \tag{6.8}$$

$$= \frac{2}{R^2}\int_{R_0}^{R} \mathrm{d}R' R'\Sigma(R') - \frac{1}{R^2}\left[R^2\Sigma(R) - R_0^2\Sigma(R_0)\right] . \tag{6.9}$$

From the integration limits it is clear this estimator does not include information from scales below R_0, because $\Delta\Sigma(R_0)$ contains the same small-scale contribution as $\Delta\Sigma(R)$, just rescaled by R^2/R_0^2. The second line follows from the first one by substituting $\Delta\Sigma(R) = \overline{\Sigma}\,(0,R) - \Sigma\,(R)$ and using the definition

$$\overline{\Sigma}(R_1,R_2) = \frac{2}{R_2^2 - R_1^2}\int_{R_1}^{R_2} \mathrm{d}R' R'\Sigma(R'). \tag{6.10}$$

In this estimator and in the ones below that also involve a transformation of the data vector, the model is transformed in the same way as the measurements. Moreover, the Annular Differential Surface Density statistic can be equivalently defined for the tangential shear quantity using angular scales:

$$\Upsilon_{\gamma_t}(\theta;\theta_0) \equiv \gamma_t(\theta) - \frac{\theta_0^2}{\theta^2}\gamma_t(\theta_0)\,, \tag{6.11}$$

using the small-angle approximation to go from R to θ. In realistic scenarios, where each lens tomographic bin has a non-negligible width, a given value θ_0 will mix a range of physical scales R_0. In this work we use the Annular Differential Surface Density statistic based on the tangential shear quantity throughout the paper (instead of the $\Delta\Sigma$ one), and use θ_0 cuts computed with the mean z_l for each redshift bin, given a specified R_0 cut. In this paper we use values for R_0 of 6 Mpc/h and of 8 Mpc/h, depending on the data set and its constraining

power.

The covariance of Υ will also generally need to be modified. Given that we can write the transformation as $\Upsilon = \gamma_t - X$, with X being a constant, then $\mathrm{Var}[\Upsilon] = \mathrm{Var}[\gamma_t] + \mathrm{Var}[X] - \mathrm{cov}[\gamma_t, X]$. In the case that X is noiseless, the covariance of Υ will be identical to the γ_t one. In our implementation we always modify the covariance of the Υ statistic to include the noise of $\gamma_t(\theta_0)$.

6.2.2 Y-transform

The Y-transformation derived in Park et al. (2021) is a localizing linear transformation that utilizes the local quantity $\Sigma(R)$ underlying the galaxy-galaxy lensing observable $\Delta\Sigma(R)$. By inverting the $\Delta\Sigma(R)$–$\Sigma(R)$ relationship, the Y quantity defined as

$$Y(R) \equiv \Sigma(R) - \Sigma(R_{\rm max}) = \int_R^{R_{\rm max}} \mathrm{d}\ln R' \left[2\Delta\Sigma(R') + \frac{d\Delta\Sigma(R')}{d\ln R'}\right] \tag{6.12}$$

recovers the local $\Sigma(R)$ up to an overall additive constant $\Sigma(R_{\rm max})$. To treat the discretized observables most frequently used in real analyses, this relation is also discretized to a linear transform given by

$$\mathbf{Y} = (2\mathbf{S} + \mathbf{SD})\,\boldsymbol{\Delta\Sigma} = \mathbf{T}\boldsymbol{\Delta\Sigma}, \tag{6.13}$$

where the trapezoidal summation matrix $\mathbf{S}$ representing the log integral and the finite differences matrix $\mathbf{D}$ representing the log differentiation are used to define the final transformation matrix $\mathbf{T}$. The linear format of this transformation allows a further direct application to a γ_t vector, as γ_t is proportional to $\Delta\Sigma$. Thus, by transforming an observed galaxy-galaxy lensing vector $\boldsymbol{\gamma}_t$ and its covariance $\mathbf{C}_\gamma$ as

$$\mathbf{Y}_\gamma = \mathbf{T}\boldsymbol{\gamma}_t, \tag{6.14}$$

$$\mathbf{C}_{Y_\gamma} = \mathbf{T}\mathbf{C}_\gamma\mathbf{T}^{\mathbf{T}}, \tag{6.15}$$

we achieve a likelihood analysis with a localized galaxy-galaxy lensing observable. It is notable that the $\mathbf{T}$ matrix nulls out components in $\boldsymbol{\Delta\Sigma}$ ($\boldsymbol{\gamma}_t$) proportional to $1/R^2$ ($1/\theta^2$), which can also be seen from Eq. (6.12) when a term proportional to $1/R^2$ is added to $\Delta\Sigma(R)$. Another way to see this is that if $\Delta\Sigma(R)$ has a $1/R^2$ shape, the integral from Eq. (6.12) vanishes. Note that this is also true for the Υ statistic of section 6.2.1, i.e. adding a term proportional to $1/\theta^2$ makes no difference to the estimator (see Eq. (6.11)).

6.2.3 Point-mass marginalization

MacCrann et al. (2020a) proposed to analytically marginalize over the contribution from within the small-scale cut by treating it as a point-mass (PM) contribution scaling as $1/R^2$. This point-mass term can be expressed as an addition to the tangential shear model for a given lens redshift bin i and source redshift bin j:

$$\gamma_t^{ij}(\theta) = \gamma_{t,\mathrm{model}}^{ij}(\theta) + \frac{A^{ij}}{\theta^2}\,, \tag{6.16}$$

where we use the small-angle approximation to go from R to θ. Here the $\gamma_{t,\mathrm{model}}^{ij}$ is a prediction based on a model for the 3D galaxy-matter correlation function $\xi_{\mathrm{gm}}(r)$ that is correct for scales $r > r_{\mathrm{min}}$, but can be arbitrarily wrong for $r < r_{\mathrm{min}}$, and A^{ij} is some unknown constant that we can marginalize over. The simple form of this contamination model makes it suitable for an analytic marginalization approach given that the scale dependence is not dependent on cosmology or the lens galaxy properties. We want to marginalize $P(\gamma_{t,\mathrm{obs}}(\theta)|\gamma_{t,\mathrm{model}}(\theta), A)$ over the unknown constant A in order to obtain the likelihood we are interested in, namely:

$$P(\gamma_{t,\mathrm{obs}}(\theta)|\gamma_{t,\mathrm{model}}(\theta)) = \int \mathrm{d}A\, P(A) P(\gamma_{t,\mathrm{obs}}(\theta)|\gamma_{t,\mathrm{model}}(\theta), A). \tag{6.17}$$

In the case that the $\gamma_{t,\mathrm{model}}$ is Gaussian distributed with covariance matrix $\mathbf{C}_\gamma$, and we have a Gaussian prior on A with mean zero and width σ_A, one can show that (Bridle et al., 2002)

$P(\gamma_{t,\mathrm{obs}}(\theta)|\gamma_{t,\mathrm{model}}(\theta))$ is also Gaussian distributed with a covariance matrix

$$\mathbf{N} = \mathbf{C}_\gamma + \sigma_A^2 \vec{x}\vec{x}^{\mathrm{T}}, \tag{6.18}$$

where $\vec{x}$ has elements $x_n = (\theta_{\mathrm{min}}/\theta_n)^2$, and θ_{min} represents the scale cut. This means that in order to marginalize over the free parameter A, we only need to add this term to the original covariance rather than explicitly sampling over possible values of A in e.g. an MCMC chain. In this work we use an infinite prior for σ_A. In this case, the extra term is added to the inverse covariance directly, following the procedure described in MacCrann et al. (2020a) and in Prat et al. (2022).

Point-mass marginalization using geometric information

The amplitudes A^{ij} can be written as

$$A^{ij} = \int \mathrm{d}z_l \int \mathrm{d}z_s n_l^i(z) n_s^j(z) B^i(z_l) \Sigma_{\mathrm{crit}}^{-1}(z_l, z_s) D^{-2}(z_l)$$

where B^i represents the total mass enclosed within θ_{min} for the i-th lens bin, $n_l^i(z)$ is the redshift distribution of each lens bin, $n_s^j(z)$ for each source bin, and $D(z_l)$ is the angular diameter distance to the lens redshift z_l. If we assume that B^i evolves slowly across the width of the lens bins we can drop the z_l dependence and define the parameters β_{ij} in the following way:

$$A^{ij} \approx B^i \int \mathrm{d}z_l \int \mathrm{d}z_s n_l^i(z) n_s^j(z) \Sigma_{\mathrm{crit}}^{-1}(z_l, z_s) D^{-2}(z_l) \equiv B^i \beta_{ij}.$$

The parameters β_{ij} are purely geometrical (sometimes also called shear-ratio information), and thus can be exactly computed analytically given the input redshift distributions[1]. Then,

1. Note that since the second term of the RHS of Eq. (6.16) only accounts for the unmodeled enclosed mass caused by the mismodeling of the halo-matter correlation function, other effects such as IA or magnification

the predicted β_{ij} factors can be used to reduce freedom in the model by fixing the relative scales between the source bins sharing the same lens bin and only marginalizing over B^i instead of over a free-form A^{ij}.

We label this variant of the model as "Point-mass (free per z_l^i)", since in this case there is only one free parameter per lens bin. When this approximation is not used (i.e. only using Eq. 6.16) we label the model as "Point-mass (free per $z_l^i - z_s^j$)". In the DES Y3 analysis, where the point-marginalization was used, the variant of the model using geometrical information was employed. In this paper, we also explore the differences, advantages and caveats of these two variants of the point-mass marginalization method.

6.2.4 Mode projection: "Project-out" estimator

To further illustrate the similarities and/or differences among the above estimators, we also construct a new estimator that we refer to as the "project-out" estimator. This estimator is designed such that it follows the philosophy of the MacCrann et al. (2020a) approach, namely by focusing on the point-mass $1/R^2$ mode within the observed galaxy-galaxy lensing data vector, while it follows the implementation used in Park et al. (2021), namely by finding a suitable linear transformation and using it to transform the data vectors and covariances. In Appendix F we show that the "Proj-out" method is actually equivalent to the point-mass marginalization method when the prior of the point-mass is infinitely wide. Also, this basic equivalence between marginalizing over a parameter and projecting out a given mode had already been pointed out in Appendix A from Seljak (1998) for a general scenario.

The core idea behind this estimator, thus, is to identify the projection of the observed galaxy-galaxy lensing data vector onto the $1/R^2$ mode, and then to remove it from the

would be a part of the first term of the RHS of Eq. (6.16), and thus do not need to be included in the computation of the β_{ij} factors.

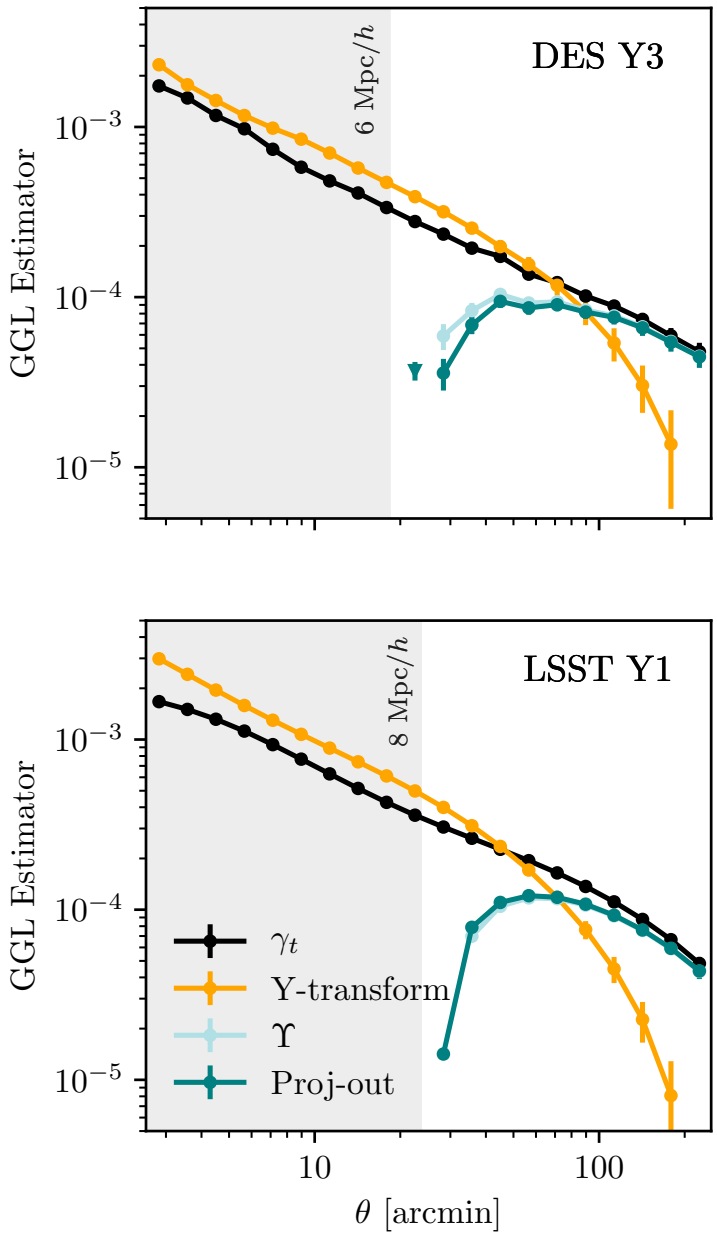


Figure 6.1: Visualization of the data vectors for each of the galaxy-galaxy lensing (GGL) estimators that we compare in this paper to localize the original tangential shear measurements (γ_t), corresponding to the second lens redshift bin and the highest source bin. We do not add the point-mass marginalization case since it does not involve a modification of the data vector, only of the (inverse) covariance. The triangle point for the proj-out case represents a negative point.

original vector. The projection operator $\mathbf{P}$ is given by (Aitken, 1936; Tegmark et al., 1998):

$$\mathbf{P} = \mathbf{A}\left(\mathbf{A^T C_\gamma^{-1} A}\right)^{-1}\mathbf{A^T C_\gamma^{-1}}. \tag{6.19}$$

where $\mathbf{A}$ has columns spanning the subspace onto which we wish to project. In our case, $\mathbf{A}$ has only one column given by $\{1/R_i^2\}$, or in practice $\{1/\theta_i^2\}$, with θ_i being the representative angular separation of the i-th bin in our data vector. Note that instead of the "vanilla" projection operator $\mathbf{A(A^T A)^{-1} A^T}$ we use the generalized least squares definition of the projection operator to properly account for the covariances in the data vector.

The "cleaned" data vector is then defined as

$$\begin{aligned}\gamma_{t,\text{proj-out}} &= \gamma_t - \mathbf{P}\gamma_t \\ &= \left[\mathbf{I} - \mathbf{A}\left(\mathbf{A^T C_\gamma^{-1} A}\right)^{-1}\mathbf{A^T C_\gamma^{-1}}\right]\gamma_t \\ &\equiv \mathbf{M}\gamma_t,\end{aligned} \tag{6.20}$$

and its covariance is given by

$$\mathbf{C}_{\text{proj-out}} = \mathbf{M C_\gamma M^T} \tag{6.21}$$

where $\mathbf{C}_\gamma$ is the original covariance matrix. The inversion of $\mathbf{C}_{\text{proj-out}}$ becomes problematic, as $\mathbf{M}$ is not a full-rank matrix. We thus follow Tegmark (1997) to define a pseudoinverse of the transformed covariance matrix as

$$\tilde{\mathbf{C}}^{-1}_{\text{proj-out}} = \mathbf{M}\left(\mathbf{C}_{\text{proj-out}} + \eta\mathbf{A A^T}\right)^{-1}\mathbf{M^T}, \tag{6.22}$$

where η is a constant whose value does not affect the performance of the pseudoinverse[2]. This can be understood intuitively as adding back in the lost mode to $\mathbf{C}_{\text{proj-out}}$, inverting,

2. We have used $\eta = 10^{-3}$ for the runs with DES Y3 data. We have checked that using a different value, e.g. $\eta = 10^3$, does not change the results.

and removing the added mode at the end. With $\gamma_{t,\text{proj-out}}$ and $\tilde{\mathbf{C}}_{\text{proj-out}}$ in hand, a full likelihood analysis can be defined using this new estimator.

6.3 Methodology

The question we are aiming to answer is whether the different methodologies to localize the galaxy-galaxy lensing measurements are consistent under the precision of current and future surveys. In order to do so, we perform a combined 2×2pt analysis using each of the methods to localize the tangential shear measurements and compare their performances at the cosmological posterior level. We choose to do the comparison in a 2×2pt analysis instead of a whole 3×2pt analysis to maximize the impact that the galaxy-galaxy lensing part of the data vector has on the cosmological parameter posteriors, thus maximizing potential differences between the localizing estimators. We perform the comparison using two different setups: 1) First we assume the characteristics of a future survey to test the differences under the smallest covariance. In particular, we choose the specifications of a LSST Y1-like survey, since that will become relevant in the near-future, and it will already be significantly more constraining than current generation surveys. 2) Secondly we apply the comparison of the methods to DES Y3 data, a noisy realistic scenario.

For the LSST Y1 simulated case, we generate the input data vectors using a *contaminated* model that includes baryonic effects and non-linear galaxy bias contributions that mostly affect small scales (see Sec. 6.3.2), and analyze it using the simpler *fiducial* model that does not take into account these contributions with a linear galaxy bias model and a dark matter only power spectrum (see Sec. 6.3.1). We find the appropriate set of scale cuts that allow us to recover unbiased cosmology in each case following the prescription described in Sec 6.3.3.

6.3.1 Fiducial model

Here we summarize the baseline or "fiducial" theory that we will use to model the observed tangential shear and galaxy clustering quantities. This is the same one used in the DES Y3 3×2pt analysis. In particular, for the LSST Y1 simulated analysis described below we assume the model presented in more detail in Krause et al. (2021a) and for the DES Y3 data analysis the one defined in Abbott et al. (2022)[3].

The tangential shear γ_t and angular clustering $w(\theta)$ observables can be expressed as various real space projections of angular power spectra. In particular we model γ_t as the following curved sky projection of the galaxy-matter angular power spectra and of other terms that encapsulate observational effects such as intrinsic alignments, lens magnification and their cross-talk, which add up to the total observed $C^{ij}_{\delta_{\rm obs}\,E}$:

$$\gamma_t^{ij}(\theta) = \sum_\ell \frac{2\ell+1}{4\pi\ell(\ell+1)} P_\ell^2(\cos\theta) C^{ij}_{\delta_{\rm obs}\,E}, \tag{6.23}$$

and we model $w(\theta)$ as:

$$w^{ij}(\theta) = \sum_\ell \frac{2\ell+1}{4\pi} P_\ell(\cos\theta) C^{ij}_{\delta_{\rm obs}\,\delta_{\rm obs}}. \tag{6.24}$$

We refer the reader to Krause et al. (2021a) for a detailed definition of $C^{ij}_{\delta_{\rm obs}\,E}$ and of $C^{ij}_{\delta_{\rm obs}\,\delta_{\rm obs}}$. P_ℓ and P_ℓ^2 are the Legendre polynomials and the associated Legendre polynomials respectively. For the matter power spectrum we use the dark matter only Halofit prescription from Takahashi et al. (2012) and assume a linear galaxy bias to relate the galaxy and matter density fluctuations.

We obtain the LSST Y1 covariances with the publicly available CosmoCov code (Krause

3. The two models are in essence identical, but differ in some of the input parameters such as the lens magnification parameters and the redshift distributions (which were measured later in the data analysis), and in the priors of the nuisance parameters.

	LSST Y1
Lens Sample	
Number density [arcmin^{-2}]	(2.29, 3.97, 6.06, 3.07, 2.62)
Galaxy bias	(1.7, 1.7, 1.7, 2.0, 2.0)
Source Sample	
Number density [arcmin^{-2}]	(2.50, 2.50, 2.52, 2.48)
Total shape noise	0.3677

Table 6.1: LSST Y1 lens and source sample specifications in the setup of this work. The listed shape noise is the total one including both ellipticity components. These values are taken from The LSST Dark Energy Science Collaboration (2018), which specifies a lens number density of 18 arcmin^{-2}, which we split in 5 redshift bins, and a source number density of 10 arcmin^{-2} which we split in 4 redshift bins.

et al., 2017; Fang et al., 2020), using the number densities and noise specified in Tab. 6.1. We include Gaussian and non-Gaussian terms computed using a halo model. We assume an area of 12300 deg^2 for LSST Y1, consistently with the specifications from the The LSST Dark Energy Science Collaboration (2018).

We use a ΛCDM model with 5 (6) free cosmological parameters for the simulated (data) case: Ω_m, A_s, n_s, Ω_b, h, (Ω_ν). We also free additional nuisance parameters to marginalize over uncertainties related to photometric redshifts – both for the lens and source samples, intrinsic alignments and shear calibration. The full list of free parameters and their respective priors can be found in table II[4] from Krause et al. (2021a) for the simulated analysis and in table I from Abbott et al. (2022) for the DES Y3 data analysis. For the simulated LSST Y1 analysis we assume the same redshift distributions that were used in the methodology paper of the DES Y3 analysis (Krause et al., 2021a), which are an early estimate of the DES Y3 redshift distributions[5]. For the DES Y3 data chains, we use the same settings and priors as

4. The only difference between the values that we use and the ones from table II from Krause et al. (2021a) is that we fix the neutrino density parameter $\Omega_\nu h^2$ to a null value. This is because CosmoCov is not able to generate a model for the covariance that takes into account neutrinos. However, as shown in figure 2 from Krause et al. (2021a), the impact of marginalizing over neutrino density is small for the DES Y3 3×2pt analysis, so we do not expect this choice to affect any of the conclusions of this work.

5. LSST is expected to use a different redshift binning with respect to the one we choose in this work. However, since we are matching the rest of the settings to Krause et al. (2021a) we decided to also match the input redshift distributions for simplicity, also given the fact that the redshift distributions that we assume

in Abbott et al. (2022), except that we do not include the shear-ratio likelihood and only combine the tangential shear and galaxy clustering measurements, since for the 2×2pt case it does not significantly change the results (Sánchez, Prat et al., 2022a).

We use the CosmoSIS (Zuntz et al., 2015) framework to generate the data vectors and perform the 2×2pt analysis. We use MultiNest (Feroz & Hobson, 2008; Feroz et al., 2009, 2019) to sample the parameter space and obtain the parameter posteriors, with the following accuracy settings:

- `live_points=500`
- `efficiency=0.3`
- `tolerance=0.01`
- `constant_efficiency=F`

For the DES Y3 data, we use Polychord with the same high-accuracy settings used in Abbott et al. (2022).

6.3.2 Contaminated input model

The baryonic contamination is obtained by rescaling the non-linear matter power spectrum with the baryonic effects from OWLS (OverWhelmingly Large Simulations project, Schaye et al. 2010; van Daalen et al. 2011) as a function of redshift and scale, comparing the power spectrum from the dark matter-only simulation with the power spectrum from the OWLS AGN simulation, following Krause et al. (2021a). For the non-linear galaxy bias contamination we utilize a model that has been calibrated using N-body simulations and is described in Pandey et al. (2020); Pandey et al. (2022). Overall we use the same procedure which is used in Krause et al. (2021a) to contaminate the fiducial data vector with these

in this work have a comparable binning and redshift range than the one predicted in The LSST Dark Energy Science Collaboration (2018) for the LSST Y1 sample.

effects. Note that while the scale cuts and constraining power for each setup will depend on these choices (both on the contaminated and fiducial model), the comparison of the estimators will be independent of it since we use the same input contamination for all the different localization methodologies.

We generate both the fiducial and contaminated data vectors at the same cosmological and nuisance parameters that were used to define the scale cuts in the DES Y3 3×2pt cosmological analysis. In Fig. 6.2 we display the differences between the contaminated and fiducial data vectors for the tangential shear in the top and angular galaxy clustering in the bottom.

6.3.3 Procedure to obtain the scale cuts

Here we describe how we obtain the scale cuts that we can use for LSST Y1 that yield unbiased cosmological results given our input contamination data vectors. We compute the differences of the posteriors in the 2D S_8–Ω_m plane between results using either the fiducial or the contaminated input data vectors. Specifically, we use the maximum a posteriori point (MAP) to compute the 2D offsets. We choose to use the "Point-mass (free per z_l^i)" model to perform this exercise, although we do not expect this choice to impact the results for the scale cuts. If the difference is above 0.3σ, it does not pass our criteria, following the same procedure as in Krause et al. (2021a). We have tested the following set of scale cuts: $w > 8$ Mpc/h, $\gamma_t > 8$ Mpc/h; $w > 12$ Mpc/h, $\gamma_t > 6$ Mpc/h; $w > 12$ Mpc/h, $\gamma_t > 8$ Mpc/h. Only this last set of scale cuts meets the criteria. See Appendix G for the plots showing these differences.

6.4 Results

Using the LSST Y1 setup described in the previous section, we find that all the estimators perform in a similar way. In Fig. 6.3 we show the results for the simulated 2×2pt analysis,

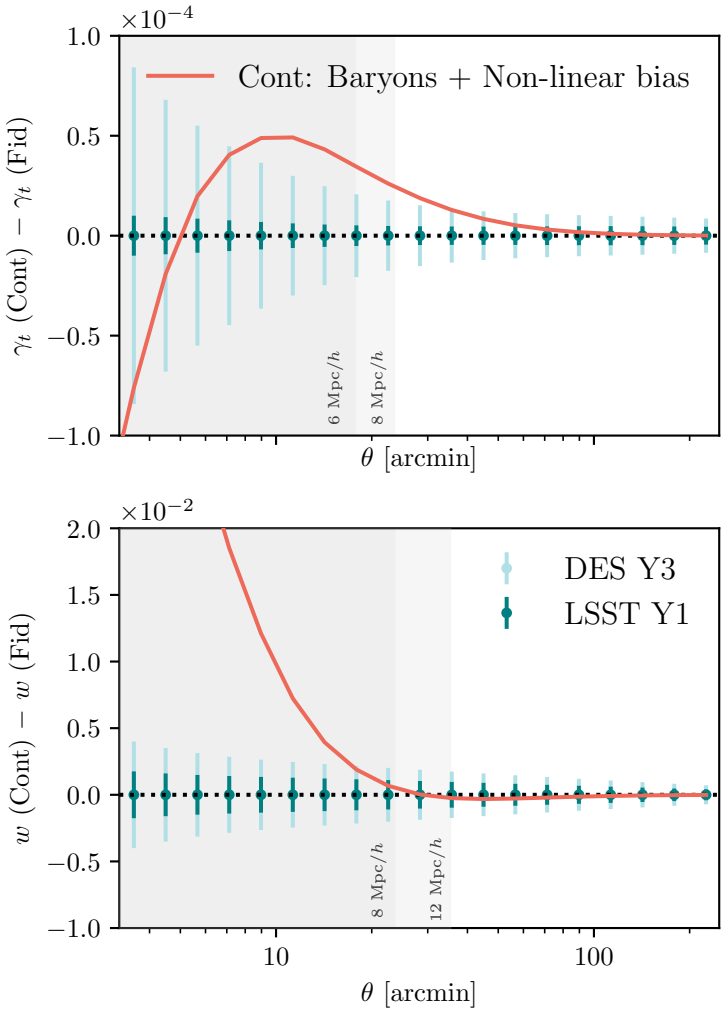


Figure 6.2: Difference between the contaminated data vector and the fiducial ones for the tangential shear (top) and angular clustering (bottom), for the redshift bin combination corresponding to the second lens bin and the highest source bin. This redshift bin combination corresponds to one of the higher S/N ones. Green error bars represent the uncertainties for DES Y3 and LSST Y1. The gray regions mark the scale cuts that are needed to obtain unbiased cosmological results from this contamination, which have been determined to be $w > 8$ Mpc/h and $\gamma_t > 6$ Mpc/h for DES Y3 and $w > 12$ Mpc/h and $\gamma_t > 8$ Mpc/h for LSST Y1 when using one of the methods to localize the tangential shear measurements (otherwise the scale cuts would need to be larger for the tangential shear quantity, as shown in Appendix H)

.

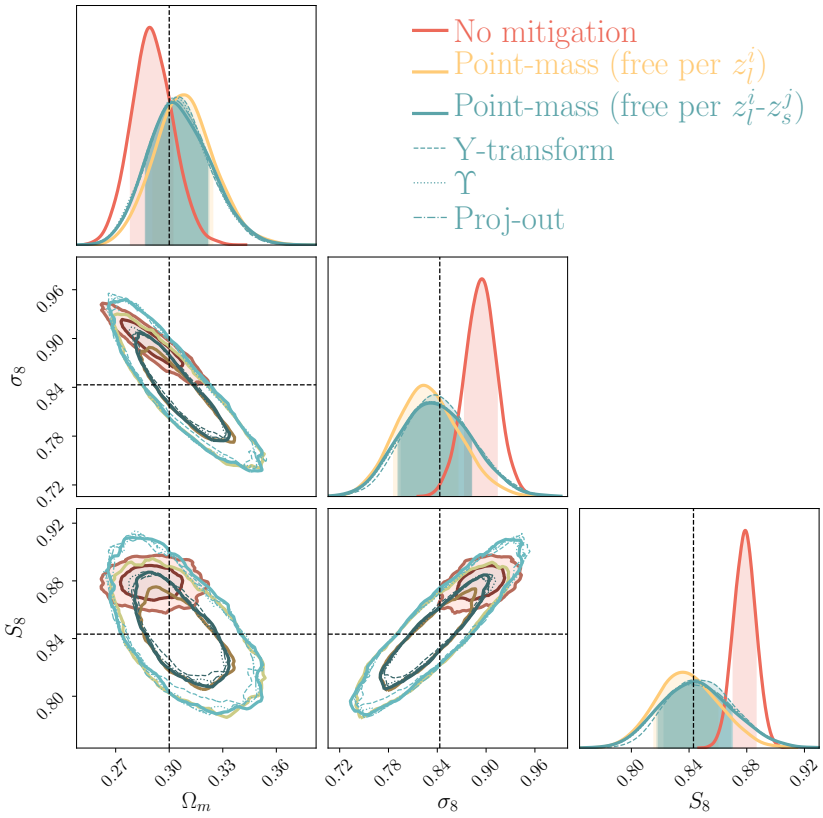


Figure 6.3: Cosmological parameter posteriors obtained from an input galaxy clustering and galaxy-galaxy lensing data vector (2×2pt) with non-linear bias and baryonic contamination, LSST Y1 covariance, 8Mpc/h scale cuts for galaxy-galaxy lensing and 12 Mpc/h for galaxy clustering. This figure demonstrates that all the methodologies to localize the galaxy-galaxy lensing measurements perform similarly at the cosmological posterior level with LSST Y1 uncertainties. The 2D contours represent 1σ and 2σ confidence regions. The shaded area under the 1D posteriors represents the 1σ confidence level in 1D.

combining galaxy-galaxy lensing and galaxy clustering for all the methodologies that we want to compare that localize the tangential shear measurements. We also add the result without applying any mitigation method, to illustrate the importance of using one of these methodologies to obtain unbiased cosmological constraints. All these results are applying the fiducial scale cuts that passed the criteria defined in Sec. 6.3.3: $w > 12$ Mpc$/h$ and $\gamma_t > 8$ Mpc$/h$. In Appendix H we also show that $\gamma_t > 40$ Mpc$/h$ cuts would be needed in order to recover unbiased cosmological constraints if we do not apply any mitigation scheme.

We find that all the methodologies are able to properly mitigate the impact of the input contamination and recover very similar uncertainties on the most constrained cosmological parameters of a 2×2pt analysis, that is, Ω_m and σ_8. In the comparison we also include the new estimator that we have developed that *only* projects out the $1/R^2$ mode, without doing any extra transformation as in some of the other methodologies, labelled as "Proj-out" in the plot. The fact that all the estimators agree with each other, and also with this new estimator, indicates that projecting out this mode is the only thing that has any effect in all the mitigation methodologies at the cosmological posterior level.

Instead we have found that differences between the methods arise from input assumptions. In particular, we observe the biggest difference is between the two different variants of the point-mass marginalization. The method labelled as "Point-mass (free per $z_l^i - z_s^j$)" does not use any extra information with respect to the other estimators and can be compared directly to them. On the other hand the one labelled as "Point-mass (free per z_l^i)" uses the approximation that the mass enclosed below the minimum scale used does not evolve within the redshift range of the lens bin, and moreover uses geometrical "shear-ratio" information to constrain the scaling between different sources sharing the same lens bin. We find that the posteriors for this case are slightly more constraining as expected since they use more information, but also slightly more biased with respect to the input true cosmology. Thus, we recommend that when applying the point-mass marginalization case using geometrical

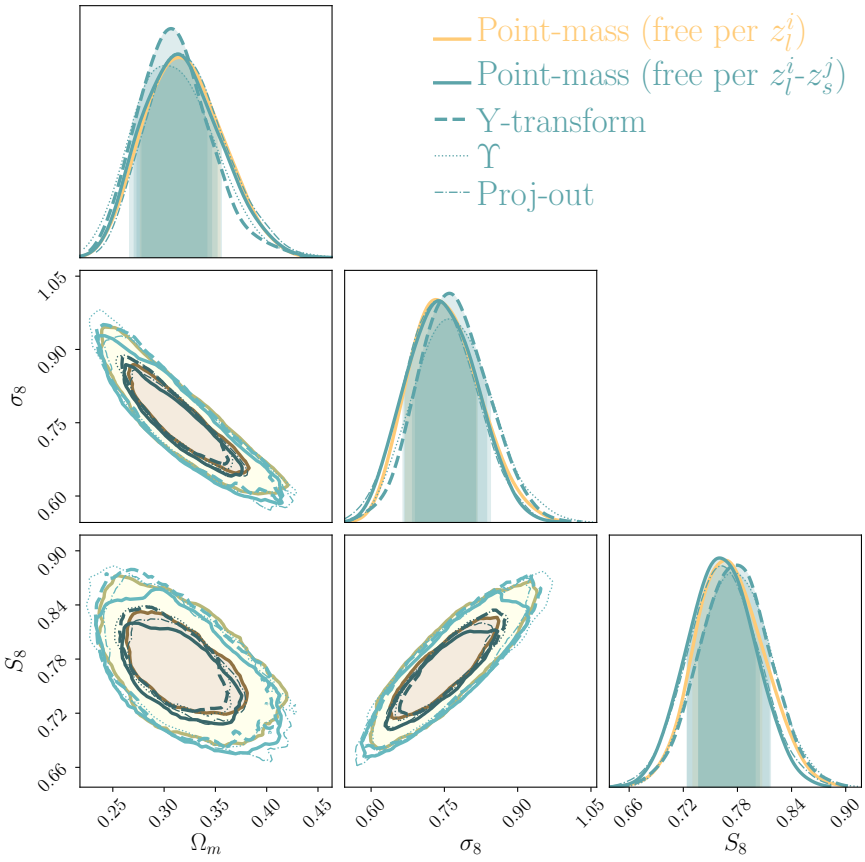


Figure 6.4: Application to DES Y3 data in a 2×2pt analysis for the MagLim sample of each of the different methodologies to localize the tangential shear measurements. This figure demonstrates that all the methodologies to localize the galaxy-galaxy lensing measurements perform similarly at the cosmological posterior level with DES Y3 uncertainties and the presence of noise. The 2D contours represent 1σ and 2σ confidence regions. The shaded area under the 1D posteriors represent the 1σ confidence level in 1D.

information to LSST Y1 or a more constraining data set, the assumption that the point mass evolves slowly within a lens redshift bin should be tested for a given lens sample.

In Appendix I we also compare the intrinsic alignment parameter posteriors. We find that all the estimators perform similarly, except for the point-mass using geometric information to constrain the scaling between the lens redshift bins that gives a much tighter (and still unbiased) posterior on the galaxy bias of the source sample b_{TA} parameter and a slightly tighter constraint for the A_2 and α_2 parameters, which are the parameters related to the

tidal torque contribution to the model.

6.4.1 Application to DES Y3 data

After testing on noiseless simulated data vector with an LSST Y1 setup we apply all the methodologies to localize the galaxy-galaxy lensing measurements to the DES Y3 data, in particular to the 2×2pt setup with the MagLim lens sample presented in Porredon et al. (2022) but without the shear-ratio likelihood. The scale cuts we use are the same as in Porredon et al. (2022), that is, 6 Mpc/h for galaxy-galaxy lensing and 8 Mpc/h for the angular galaxy clustering. In Fig. 6.4 we show the results of this comparison. Note that the "Point-mass (free per z_l^i)" case corresponds to the fiducial DES Y3 2×2pt result presented in Porredon et al. (2022). We find that all the methodologies give consistent results, even in this noisy and more realistic scenario, which presents non-linearities at the small scales and includes all the effects from the real Universe. In Appendix J we compare the constraining power of DES Y3 and LSST Y1 to give a sense of scale.

Moreover, we compare the posteriors on the TATT intrinsic alignment parameters for DES Y3 data and find similar conclusions as in the simulated case, as shown in Appendix I.

6.4.2 Performance differences

Regarding performance differences between the methods, we find the "Proj-out" estimator to be less numerically stable than the other approaches in its current CosmoSIS implementation. This is because the "Proj-out" method requires an input arbitrary η value to obtain the pseudoinverse of the transformed covariance matrix, as defined in Eq. (6.22). While we have checked that a considerable wide range of η values yield the same results under the conditions of this analysis, outside a certain range that is no longer the case. Thus, robustness against different η values might need to be revisited in other settings. Moreover, we find the point-mass marginalization method to be the simplest to use as currently implemented in

	Modify data vector?	Numerical Stability	Computational expense
Point-mass marginalization	No	Excellent	Fastest
Annular differential surface density (Υ)	Yes	Excellent	Fast
Y-transformation	Yes	Excellent	Slow
Proj-out	Yes	Poor	Slowest

Table 6.2: Comparison of the process and performance of each methodology to localize the galaxy-galaxy lensing measurements. The computational expense estimates are based on the current implementation of the CosmoSIS code. More details about the performance differences can be found in Sec. 6.4.2.

CosmoSIS, since it does not require modifying the input data vector or covariance matrix files. Finally, we also compare the running time of the different estimators. Using the LSST Y1 setup, the point-mass marginalization without geometrical information took 7h, the Υ statistic took 7h 45 min, the Y-transformation took 12h and the proj-out estimator took 16h 40 min, using the MultiNest sampler with the settings defined above and using the same number of cores. On DES Y3 data and the Polychord sampler with high-accuracy settings, the point-mass marginalization without geometrical information took 45h, the Υ statistic took 49h, the Y-transformation took 54h and the proj-out estimator took 70h. We summarize these findings in Table 6.2.

6.5 Summary and conclusions

In this paper we compare three existing methodologies to localize the galaxy-galaxy lensing measurements: the Annular Differential Surface Density estimator (Υ) presented in Baldauf et al. (2010), the Y-transformation derived in Park et al. (2021) and the point-mass marginalization described in MacCrann et al. (2020a). We compare them at the cosmological posterior level, performing a 2×2pt analysis which combines projected angular clustering and tangential shear measurements. We find that all these methods are able to mitigate the impact of small scale information when using a LSST Y1 setup with noiseless simulated data vectors, and that they are all performing in a very similar manner, yielding equivalent posteriors on the cosmological parameters.

To further illustrate the similarities and/or differences amongst the above listed estimators, we also construct a new estimator that we refer to as "project-out". The "project-out" method identifies the projection of the observed galaxy-galaxy lensing data vector onto the $1/R^2$ mode, and then it removes it from the original vector, following a similar procedure to the Y-transform methodology, but in this case *only* removing this mode. Then, we proceed to compare the posteriors obtained with the "project-out" method to the other ones, finding it yields equivalent results. Therefore, we conclude that the removal of the $1/R^2$ mode is the only relevant transformation that is needed to localize the tangential shear measurements and that the rest of the modifications in the other estimators are not producing any significant differences at the cosmological posterior level.

We also compare two different variations of the point-mass marginalization methodology, one that uses exactly the same information as the other estimators and one that uses extra geometrical information to constrain the scaling of the point-mass between different lens and source bin combinations, by assuming that the enclosed mass does not evolve with redshift within the width of the lens bin. We find that the point-mass marginalization using geometric information yields slightly more constraining but also slightly biased results on the cosmological parameters in the LSST Y1 simulated case, due to the approximation it makes. Thus, the assumption going into this point-mass variant should always be tested before applying it to more constraining data sets. Notably, we also find that the extra geometrical information significantly improves the precision (while keeping the accuracy) of the intrinsic alignment parameters of the tidal alignment and tidal torque (TATT) model. In particular, we find the biggest difference in the posterior for the galaxy bias of the source sample $b_{\rm TA}$ and in the parameters controlling the tidal torque part of the TATT model.

We also compare the results obtained using any of the mitigation schemes with the case of not applying any mitigation scheme but applying larger scale cuts. With the LSST Y1 setup, we find that the mitigation schemes yield $\sim$1.3 times more constraining S_8 results

than applying larger scale cuts without using any mitigation scheme.

Finally, we apply all the methods to DES Y3 data, reaching very similar conclusions as in the simulated case. However, even if the DES Y3 data has larger uncertainties than the simulated LSST Y1 case, this exercise is still meaningful since it provides an input data vector with the non-linearities and baryonic effects of the real Universe, together with any other other unforeseen contamination that is not present in our fiducial model. It also tests the methods in the presence of noise. In this case we still find that all the methodologies perform in a similar manner.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Conclusions

We live in a very exciting era for Cosmology. Our pursuit for answers to the most fundamental questions about the Universe is nowadays data driven and it is enabled by some of the most powerful experiments in Cosmology to date. The high precision of these observations, however, requires our models and numerical techniques to greatly improve in order to be able to keep up with the data. In many cases, due to our lack of understanding of how to accurately model the observables as we move deep into the non-linear regime, i.e. when we utilize the small angular scales, we opt to simply remove data points while we fit our models to the data vectors. This is not optimal as the non-linear scales contain a wealth of information which is not only constraining for Cosmology alone, but it also connects the large, linear scales to the non-linear regime. Developing models to describe the full range of scales is a big and difficult task; it is, however, the best path forward if we want to understand the full picture of how the Universe operates and how the various scales interact with each other to create the large-scale structure we observe today.

For the main part of this Thesis, we have focused on developing models for the galaxy-halo connection via the Halo Occupation Distribution approach. In Chapters 3 and 4 we presented applications of this to data from the Year 3 of the Dark Energy Survey. In Chapter 3, which is based on Zacharegkas et al. (2022), we developed a complex framework to model galaxy-galaxy lensing which we then applied to the DES Y3 data that was used in the cosmological analyses. We thus pushed these analyses to much smaller scales and we were able to constrain properties of the lens dark matter halos for the two cosmological samples. Moving forward, in Chapter 4 (Zacharegkas et al. in prep) we discussed an extension of the previous model that achieves two main purposes: 1) It accounts for more properties of the

lens galaxies, namely stellar mass and color; 2) We also modeled galaxy clustering on small scales, in order to perform a 2×2pt analysis that utilizes the full range of scales. In additions, in that chapter, since we were interested in utilizing stellar-mass and color, we created a new lens sample that we called the *stellar-mass sample* in DES Y3 which we described in detail. This work is in progress as this thesis is being written and the final paper is scheduled to be submitted shortly after this thesis is defended.

In Chapter 5, which is based on Prat et al. (2022) and Sánchez et al. (2022a), we presented some additional work that was done within the DES collaboration that utilizes aspects of the small-scale HOD framework on the previous two chapters in order to perform tests on the data. In the first chapter we showed how having a good grasp of the galaxy-halo connection can be important in determining the scale cuts in cosmological analyses, via test on contaminated simulated data vectors. In the next and final chapter, we talked about how small-scale information can be exploited in Cosmology via indirect ways by utilizing ratios of shears. Still, however, without a galaxy-halo connection model on cannot push the analyses to the smallest scales.

Again related to small scales, in Chapter 6 we discussed various methodologies one can use in order to mitigate the effects from contamination in galaxy-galaxy lensing from small scales; this is based on Park et al. (2016). This work is a nice complement to what we discussed in Chapters 3 and 4 in the sense that it is an approach to treating small scales that is the opposite of what was discussed in those chapters: Instead of modeling small scales and including them in the analysis, we remove both those non-linear scales and their effect on the linear scales that we consider. This also serves as a good overview of what is usually done in the cosmological literature regarding the non-linear regime and highlights even more how important it is to be able to model it.

Before we conclude this thesis, below we very briefly discuss interesting future directions and extensions of what is done this far.

7.2 Future work

In this section we discuss various future directions that we can take utilizing small scales in cosmology. As mentioned throughout this thesis, small scales in cosmology contain a wealth of information and they can be used to tests aspects of the ΛCDM model that have not been thoroughly tested before. Below we briefly introduce some of these ideas.

7.2.1 Cosmology with small scales

Having a framework to perform the equivalent of a 2×2pt or even a 3×2pt analysis using large and small scales has been one of the main reasons driving the development of the models in Chapters 3 and 4. Thus, in the future we would like to perform such an analysis in order to improve the constraints on cosmology and test the consistency between small and large scales. Of course, varying the cosmological and the galaxy-halo connection parameters simultaneously is computationally expensive and requires a lot of additional testing for systematics, and thus this would be a complex work.

Moreover, something that we realized during our DES Y3 small-scale analysis, is that pushing our modeling to small scales requires accounting for many additional effects. These include, but are not limited to: baryonic feedback, stripping of dark matter from the outskirts of subhalos, accounting for modeling uncertainties on the highly non-linear scales, halo exclusion, modeling dark matter density profiles inside a halo and also modeling how satellite galaxies are spatially distributed within the halo. If we want to perform cosmological analyses utilizing all scales we must carefully model and account for all these effects. It also needs to be emphasized that to properly add small-scale data to such an analysis, it is crucial to have a good model for the transition scales, from the 1-halo to the 2-halo regime.

7.2.2 Stress-testing the ΛCDM paradigm

Beyond "traditional" cosmological studies with the inclusion of small scales, it is very interesting to utilize the tools we have developed in this work to stress-test the ΛCDM model in the 1-halo regime. Some of ways we can do that are the following.

1) *Testing models of Dark Matter with small scales*

There are three key ingredients in our small-scale model that we can use to compare different models for dark matter: I) The density profiles of the dark matter distribution inside the central halos; II) The spatial distribution of satellite galaxies within a halo; III) The halo mass function and dark matter power spectrum models. By varying these model ingredients based on different dark matter models we can infer what are the dark-matter properties that the data prefers.

2) *The "lensing is low" problem*

When small-scale lensing model predictions are produced from galaxy clustering analyses of the galaxy-halo connection, the signal is found to be lower than the data. This is commonly referred to as the "lensing is low" problem and there is debate as to whether this is an indication of new physics or of an incomplete understanding of galaxy formation physics or cosmology. We can use the small-scale model from this work to test the effects on the small-scale modeling, in a combined lensing and clustering analysis, from additional model complexity or modifications to the theory. For example, baryonic effects or assembly bias could play a significant role in solving this problem, and we can also test modifications to General Relativity.

7.2.3 Applications of Machine Learning in Cosmology

Given how computationally demanding and complex analyses that simultaneously vary the cosmological parameters and the galaxy-halo connection are, Machine Learning (ML) is needed to efficiently perform the calculations. Some of the ways ML will assist in carrying

out such an analysis are the following.

1) *Inference of galaxy properties*

Inferring galaxy properties for millions of galaxies is computationally expensive. However, we can use a subset of galaxies with known properties to train a ML algorithm to efficiently predict these properties for the whole data set. For example, we can use this approach to infer the stellar mass, redshift and color for hundreds of millions of DES galaxies in just a few minutes to a few hours, as we have seen in Chapter 4.

2) *Production of large-scale-structure weights*

We usually correct for systematics in our measurements of the observables by weighting galaxies based on how likely they are to have been affected by observing conditions and other sources of contamination. In DES, we are using maps of these possible contaminants and by finding correlations between those and maps of galaxy counts, we can produce our weights. Training a Neural Network (NN) to do that is extremely efficient, and it becomes necessary since we often have to make changes to our samples which necessitates producing new weights every time we do so. An example of this we have also seen in Chapter 4.

3) *Construction of emulators*

The big volume of the parameter space to be explored and the large number of theory predictions we need to generate and compare with the data is so large when we vary cosmology and the galaxy-halo connection at the same time that running "traditional" Bayesian parameter inference codes is inefficient to solve this problem. The best way to approach that is to create emulators that interpolate in high dimensions and return fast theory predictions by utilizing ML algorithms. Therefore, using ML tools my goal is to create emulators that given a cosmology and a galaxy-halo connection would quickly provide predictions for galaxy clustering statistics.

4) *Forward modeling*

Forward modeling the galaxy halo-connection is a great way to construct a more realistic

framework that captures physics which is hard to model otherwise (up to limitations of the simulations). Such an alternative way of modeling the galaxy-halo connection can, for example, account for non-linearities in the model and additional complexity such as satellite striping, assebmly bias and halo exclusion in a precise way. More importantly, however, this approach can help model the transition scales from the 1-halo to the 2-halo regime more accurately, which is usually a hard thing to do with the more conventional methods. Moving forward, we can utilize simulations and Machine Learning techniques to forward model the galaxy-halo connection in order to perform cosmological studies using a wide range of scales.

7.2.4 Combining LSS data with other probes

One extension of the work in this thesis is a joint analysis of large-scale structure (LSS) data – including small scales – with other probes. The goal is to use the joint data set to constrain cosmology, the galaxy-halo connection, and any new quantities that the additional data can constrain. Some really interesting cross-correlations of data sets are the following.

1) *LSS and CMB joint analysis*

Weak lensing analyses, like from the DES collaboration, allow us to directly constrain the late-time Universe and play a key role in probing the dark matter and structure formation. We can also infer constraints on the latter using the Cosmic Microwave Background (CMB) data from Planck; however, the CMB probes the early Universe. By combining the two, we can probe a wider range of redshift, and furthermore test the compatibility of the two data sets while simultaneously constraining the cosmological parameters and the galaxy-halo connection. The increased constraining power coming from the inclusion of the small scales could also shed light on the well-known tension in the inference of the Hubble constant from the two probes.

2) *LSS and Compton-y joint analysis*

An especially interesting data combination is that between galaxies and the distribution

of hot gas in the Universe, probed via the thermal Sunyaev-Zeldovich effect and usually measured via the Compton-y parameter. This specific cross-correlation can constrain the thermal energy content of the Universe on large scales (Pandey et al., 2019) and the thermodynamics and properties of astrophysical feedback processes on small scales (Schaan et al., 2021). This data combination provides therefore a great way to constrain baryonic feedback and other small-scale processes which we need to account for in cosmological studies using all scales. Importantly, this cross-correlation is also robust to systematics.

3) *LSS and gravitational wave joint analysis*

Gravitational wave (GW) data from a binary system can be a cosmological tool to measure the luminosity distance to the source. Combining this information with redshift measurements from LSS data we can constrain the cosmic expansion and perform studies of dark energy and modified gravity.

7.2.5 Testing the compatibility of cosmological constraints

As we discuss in Appendix K, as the precision of the various experiments in Cosmology increases, it becomes more and more crucial that we have rigorous ways of assessing if there is tension between two or more data sets. The most complete way of doing so is to also account for any correlations between the data sets, as in many cases there is overlap between them. The inclusion of small scales makes the need for dedicated Concordance-Discordance Estimators (CDE's) even more evident, as the constrains from a full-range analysis are expected to be better. More than that, however, the addition of external data sets in our small-scale analyses (see e.g., Section 7.2.4) would require the assessment of internal and external tensions. Even within the same data set, we will have to quantify the agreement or disagreement between small and large scales by splitting the data in that way and for this the usage of "data splits" and "parameter splits", as discussed in Raveri et al. (2020), and extensions of that work is expected to be of great importance.

7.2.6 Small-scale constraints of Intrinsic Alignment

The orientation of galaxies, even in the absence of lensing, is not random, but it is rather correlated with the Large-Scale-Structure (LSS) density field. This phenomenon is know as Intrinsic Alignment (IA) and it is one of the effects that need especially careful consideration in future surveys, as the precision increases. While working on the small-scale analysis in DES, we encountered the difficulty of modeling this effect on small scales. Even on large scales, there is some debate on what the correct treatment of IA is, but on small scales its modeling is highly uncertain. For this reason, we chose to utilize redshift bins of our lens and source galaxy samples that produce galaxy-galaxy lensing signal with small IA contribution.

Moving forward, we can improve how we model IA on all scales in a consistent manner. To begin on this, we can utilize the HOD constraints in this work to produce theory predictions of galaxy-galaxy lensing curves for the lens-source redshift bin combinations with potentially high IA contamination. Taking the difference between theory and data, we can study the residuals and fit IA models to them. This will give us an initial intuition as to whether our current IA models can explain these residuals while constraining the model parameters. If this is not the case, we will develop novel ways of modeling IA on all scales. For example, in a halo model framework of modeling small scales, IA can naturally be incorporated, as discussed e.g. in Georgiou et al. (2019); Fortuna et al. (2021), where the properties of the galaxies and their host halos are linked to the IA signal.

APPENDIX A

MODEL VALIDATION

In this appendix, we present tests validating our modeling code using both external code and numerical simulations.

A.1 Comparison with DES cosmology pipeline

As part of validating our code we have done thorough comparisons with CosmoSIS (Zuntz et al., 2015). CosmoSIS is the official code basis for DES cosmological analyses. As a result, it is important to establish consistency with CosmoSIS on the regimes used for cosmology analysis, effectively the 2-halo regime. We compare the galaxy-cross-matter power spectrum $P_{\rm gm}$, the projected lensing power spectrum $C_{\rm gm}$ and the tangential shear γ_t. For this purpose we used constant galaxy bias values within a redshift bin to match the predictions from CosmoSIS at large scales, $\theta \gtrsim 30$ arcmin. In Figure A.1 we present the residuals between what our code produces and CosmoSIS. For our comparisons we have used the same $n(z)$ distributions and cosmological parameters in both CosmoSIS and our code. The parameter and bias values we used for this comparison are listed in the caption of Figure A.1. We note here that the cosmology, bias and redshift distributions we used are not the same as what is used or derived from the main analysis of this work.

The first panel of Figure A.1 validates that our implementation of the Eisenstein & Hu (1998) fitting functions for the linear matter power spectrum and our usage of Halofit to calculate the nonlinear spectrum is in good agreement with the results from CosmoSIS which uses CAMB for the linear spectrum prediction and Halofit to apply non-linear corrections to it. Going from $P_{\rm gm}$ to $C_{\rm gm}$ in the second panel we are also testing whether our treatment of the redshift distributions in our averaging procedure works as expected. Finally, to translate $C_{\rm gm}$ into γ_t and thus go from the second to the third panel we are

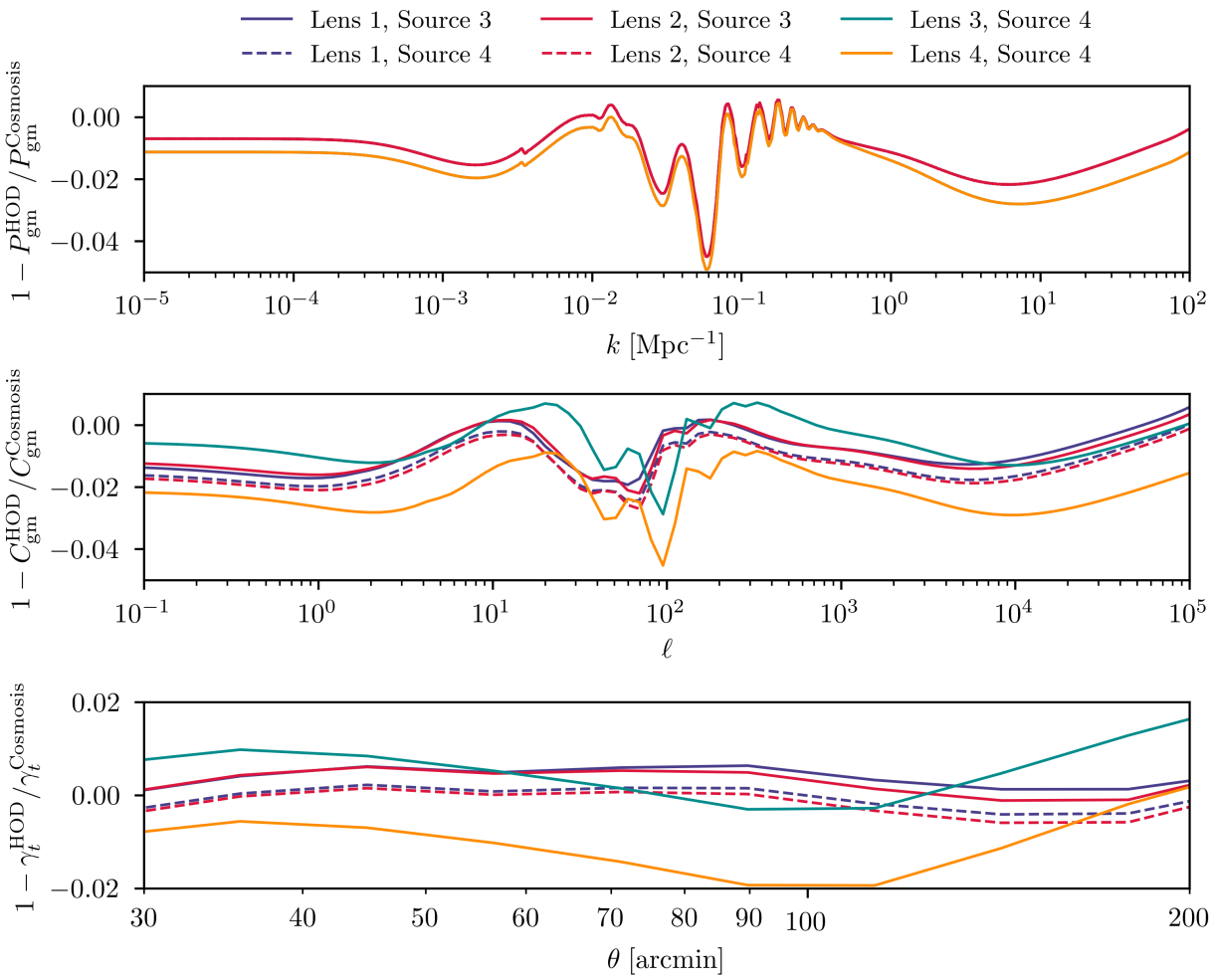


Figure A.1: CosmoSIS code comparison residuals for $C_{\rm gm}(\ell)$ and $\gamma_t(\theta)$ for 6 bins of interest. The bias values per each of the four lens bins $[1, 2, 3, 4]$ are $\bar{b}_g = [1.2, 1.6, 1.7, 1.7]$ respectively. For the first panel we have used the mean redshift of each lens redshift bin to calculate and compare the galaxy-matter cross power spectra. The other two panels show the projected power spectrum and tangential shear comparison for the average over the redshift distributions. These comparisons are done using the following parameters for a flat ΛCMD cosmology: $\Omega_m = 0.25$, $\Omega_b = 0.044$, $\sigma_8 = 0.8$, $n_5 = 0.95$, $H_0 = 70$ km/s/Mpc and $\Omega_\nu = 0$. Furthermore, note that the redshift distributions, $n(z)$, are not the same as what we used throughout this paper, but both our and the CosmoSIS results used the same $n(z)$ for lenses and sources.

confirming that our code is in agreement with CosmoSIS when transforming to real space. Note also that CosmoSIS is using the full-sky formalism to calculate the tangential shear, while we opt for the Hankel transform, i.e. flat-sky approximation, approach to gain in speed. However, for the angular scales we are interested in we have tested both approaches to confirm that the flat-sky approximation is sufficient, which is what the last panel of Figure A.1 essentially demonstrates.

The upper and middle panels of Figure A.1 show that our galaxy-dark matter cross power spectrum and, as a result, the projected lensing power spectrum, respectively, appear to be systematically lower than the CosmoSIS output. We trace that to a difference in the matter power spectrum from the two codes, as we are utilizing the Eisenstein-Hu fitting functions to calculate the dark matter transfer function whereas CosmoSIS is calling CAMB to evolve the primordial spectrum. Moreover, the presence of baryonic acoustic oscillations complicate the spectrum and the residuals appear worse around the scales that correspond to these wiggles. In addition to that, the calculation of $C_{\rm gm}$ involves the multiplication of $P_{\rm gm}$ by geometrical factors (Equation (3.14)). CosmoSIS is using a constant value in each redshift bin for $\Sigma_{\rm c}^{-1}$, while we are calculating that quantity as a function of redshift within a given bin, which leads to more differences in the resulting lensing power spectra when averaging over the $n(z)$ distributions. Overall, we find a non-significant $\sim 2\%$ deviation in $C_{\rm gm}$ and we also find a good overall agreement to within $\sim 2\%$ for the tangential shear outputs.In order to quantify the impact on our the derived halo properties from using the EH98 functions instead of CAMB we have produced a simulated data vector using CAMB which we then fitted with our fiducial model. From this test we found that the galaxy bias is recovered to $\sim 1\%$ accuracy, while the halo mass and satellite fraction is unchanged. To take this into account we have incorporated this uncertainty into our error bars on the galaxy bias from our main analysis.

A.2 Validation against simulations

Although a full end-to-end simulation test is not possible due to the limitations of existing simulations (resolution in mass, spatial resolution in ray-tracing, galaxy selection, etc.), we can validate different components of our analysis pipeline with simulations to ensure robustness of our results.

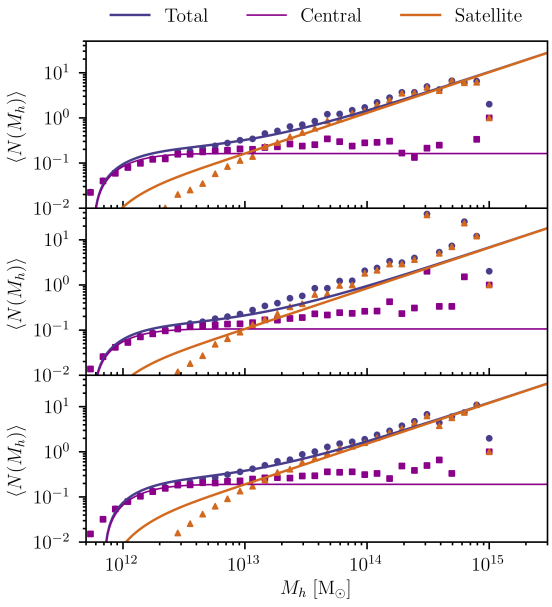


Figure A.2: Fits to the HOD measured from Buzzard high-resolution. Each panel corresponds to a different redshift bin. We fit the HOD directly using our model for central (magenta squares) and satellite (orange triangles) galaxies, as well as the total number of galaxies (blue points). The three panels, from top to bottom, correspond respectively to the following redshift bins: $z \in [0.0, 0.32]$, $[0.32, 0.84]$, $[0.84, 2.35]$.

First, we test whether our fiducial HOD model (Equations (3.1) and (3.2)) is sufficiently flexible to describe the underlying HOD of the lens galaxy sample. We note that this is not trivial especially for redMaGiC given the particular selection used in the algorithm (see Section 3.4.1). We check this by measuring the HOD from a set of high-resolution Buzzard mock galaxy catalog (DeRose et al., 2019), and fit the HOD with our fiducial model. A redMaGiC sample is constructed from the mocks using the same algorithm as applied to data, and should capture qualitatively the characteristics of the redMaGiC sample. Figure A.2 shows the measurement from the mocks together with our fit using Equations (3.1) and (3.2). We find that our model describes qualitatively the redMaGiC HOD well. The inferred satellite fraction from the fits to the Buzzard HOD is ~ 0.2.

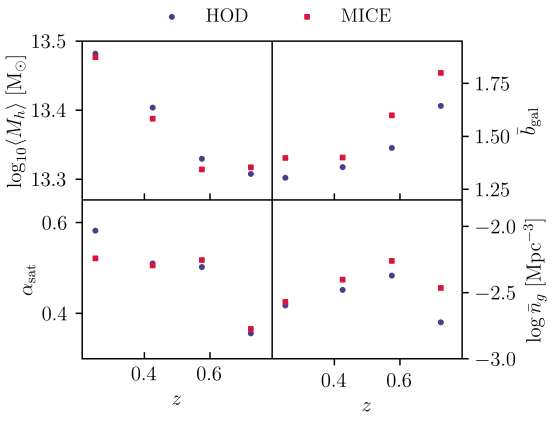


Figure A.3: Comparison between average halo mass, galaxy bias, satellite fraction and galaxy number density from our model prediction (blue points) and the corresponding measured quantities from MICE (orange squares) for the first four lens redshift bins. The HOD parameter vector $(\log_{10} M_{\rm min},\ \log_{10} M_1,\ \alpha,\ f_{\rm cen},\ \sigma_{\log M})$ used in the calculations are, for all 4 redshift bins respectively, $(12.38, 12.61, 0.73, 0.18, 0.5)$, $(12.15, 12.74, 0.84, 0.16, 0.22)$, $(12.16, 12, 72, 0.85, 0.17, 0.27)$, $(12.51, 13.3, 0.82, 0.2, 0.26)$.

Next, we perform a series of tests with the MICE simulations (Fosalba et al., 2015; Fosalba et al., 2015; Crocce et al., 2015; Carretero et al., 2015). The galaxies in the MICE simulations are populated according to a given HOD. This makes a similar *a priori* test as what was described above for BUZZARD slightly circular. We can, however, perform a number of other tests. First, for given HOD of galaxy samples, we check if our derived halo mass, galaxy bias, satellite fraction and galaxy number density agrees with what is measured directly from the simulations. Figure A.3 shows these comparisons. As we can see, our calculations are in good agreement with the MICE measurements, although they differ slightly. The trends followed by the points as a function of redshift, however, are always in very good agreement.

Second, for given HOD parameters and redshift distributions, we can compare our model prediction for γ_t with the measurements from the mock galaxy catalog. This is shown in Figure A.4 for six lens-source redshift bin combinations, as indicated in each panel. The large-scale measurements are generally in good agreement compared to the model prediction, especially for the higher lens redshifts. The small scales in each panel, however, are always in tension. Specifically, the measured γ_t is consistently lower than the model. Part of the explanation for this is the mass resolution in MICE which limits what we can measure, thus

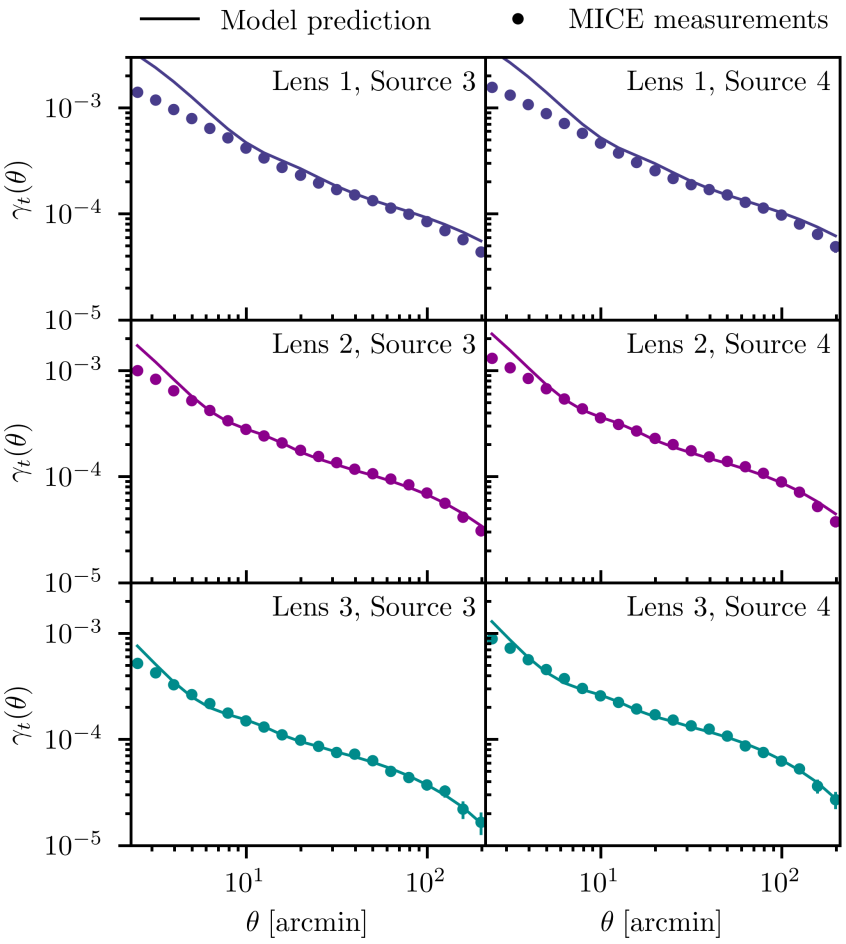


Figure A.4: Comparison of the measured γ_t as a function of θ in MICE simulations (points) and our model prediction (lines) for the lens-source redshift bins indicated in each panel. The HOD parameters used for each model line are the input to the simulations and are listed in the panels of Figure A.3.

leading to lower signal. This could also explain why the large-scale agreement is worse at the lowest redshift bin (Lens 1), since the same angular scale corresponds to smaller objects at low redshifts compared to higher redshifts. However, we do not expect this to be a big limitation in our case, given the big masses of REDMAGIC galaxies. More importantly, the dominant factor of the small-scale disagreement is that in MICE the galaxy positions do not correlate exactly with the underlying dark matter distribution. Instead, galaxies and dark matter trace each other on the mean, which could lead to small 1-halo power spectrum, and thus $\gamma_t^{\rm 1h}$, measurements. We have checked that the scales where we see the largest disagreement correspond to the 1-halo regime in each redshift bin.

APPENDIX B

RESULTS FROM SYSTEMATICS DIAGNOSTICS TESTS

In this appendix we present the results from the diagnostic tests we describe in Section 3.5.3, following the methodology from Prat et al. (2022). Figures B.1 and B.2 show a summary of all these tests for REDMAGIC and MAGLIM respectively, which include: the cross component, LSS weights and the responses. We also include the boost factor on this plot as discussed in Section 3.5.1. In the figures we also list the χ^2 between each curve, and the null hypothesis, using the covariance matrix of our γ_t measurements. We discuss below our findings for each test.

Cross component: The measurements of $\gamma_\times$ at large scales are consistent with zero. At smaller scales, below a few arcmin, $\gamma_\times$ fluctuates around zero, roughly within the error bars. The most noticeable exception is bin [Lens 1, Source 4] for which the smallest-scale measurements for the cross component get close approaches ~ 0.004. Considering that at small scales the level of noise increases, we do not find the behavior of $\gamma_\times$ worrisome. Furthermore, the reduced χ^2, even for bin [Lens 1, Source 4], is close to 1, which indicates the absence of significant problems.

Responses: Based on our results, when we compare our fiducial measurements which use a scale-averaged response per bin versus the same measurements when the exact scale-dependent responses are utilised, we find no strong evidence for disagreement between the two methods. In all the bins that we use in this work, this difference is subdominant to the statistical uncertainties, and the reduced χ^2 values always very small. We, therefore, conclude that our analysis based on the scale-averaged responses is good enough for our purposes.

LSS weights correcting for observing conditions: Comparing the measured shear with and without applying the LSS weights leads to no significant differences, as also indicated by the very small reduced χ^2 of each panel. This is shown by the fact that the difference between

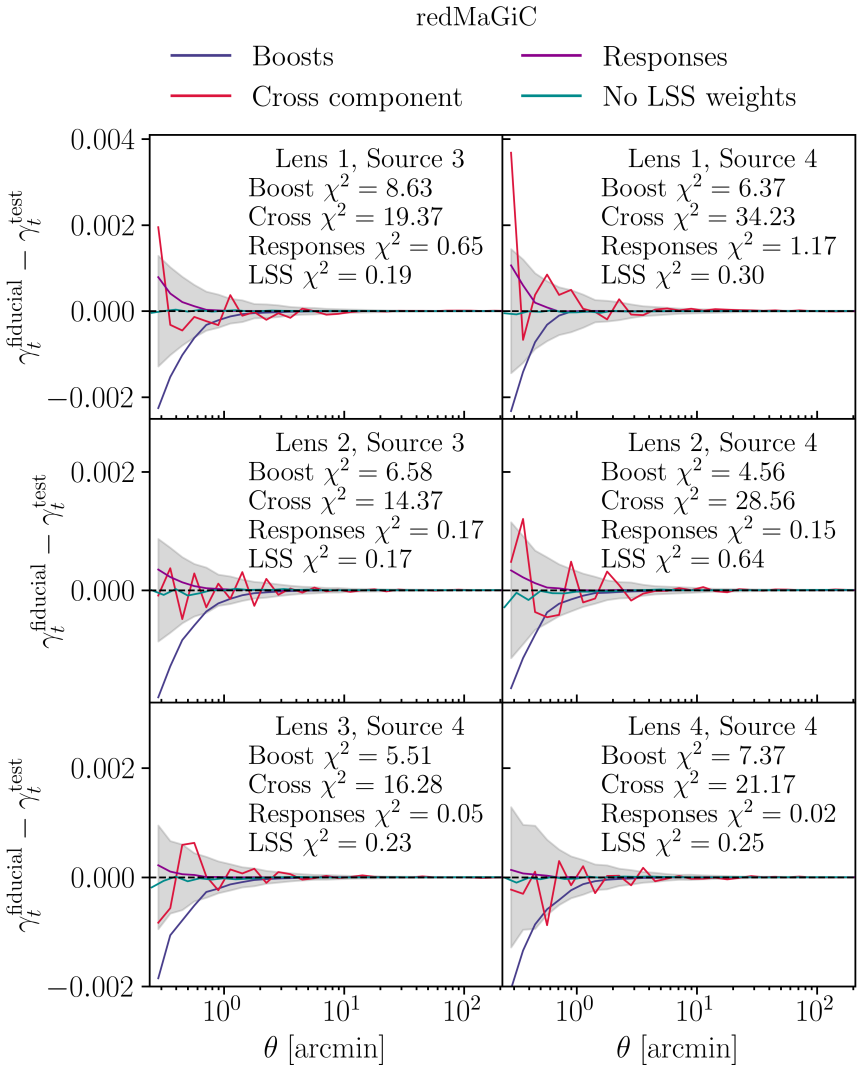


Figure B.1: Systematics tests, as discussed in Section 3.5.3, for the REDMAGIC sample. *Boosts:* Comparison of γ_t with and without applying the boost factor correction; *Cross component:* The cross-component of shear; *Responses*: Effect from using the scale-dependent responses compared to applying the average responses in each angular bin; *No LSS weights*: Effect from not applying the LSS weights to correct for observing conditions; *Gray area:* The error bars on the shear measurement. In each panel we also list the χ^2 between each test and the null, using the covariance of our γ_t measurements. The number of points for each of the lines is 30.

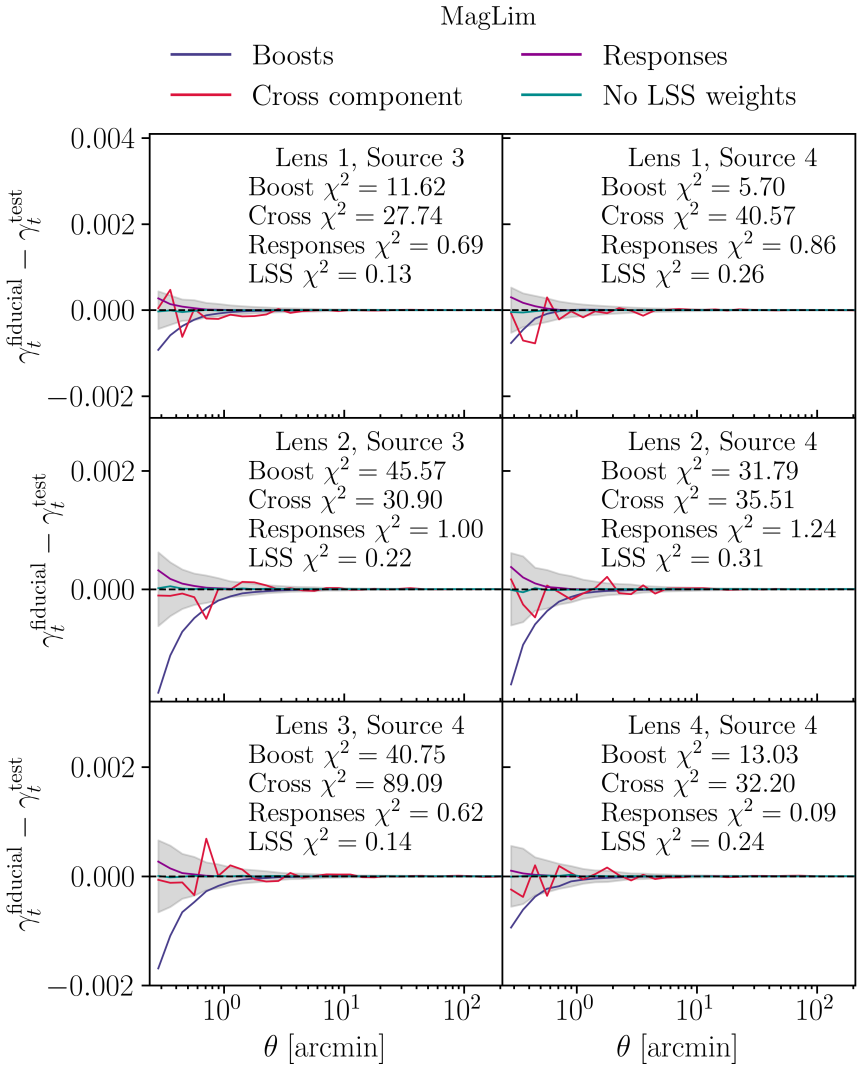


Figure B.2: Same as Figure B.1 but for the MagLim sample.

the two is always close to zero and smaller than our error bars. Thus, we find no problems with this test.

APPENDIX C

HALO EXCLUSION

In this appendix we discuss the effect of incorporating Halo Exclusion (HE) into our modeling. Based on HE, halos that overlap with each other are excluded from the 2-halo components of the galaxy-galaxy lensing model prediction, in order to avoid double counting. There are many different prescriptions for HE in the literature, some of which can be very computationally expensive. Some authors (e.g. Zheng, 2004; Tinker et al., 2005; Yoo et al., 2006) adopt the approach of choosing the appropriate upper limits to the halo masses when integrating over the mass function in Equations (3.10) and (3.11). The maximum masses, namely M_{h1} and M_{h2}, in these models, under the spherical-halo assumption, satisfy the requirement that the distance between the centers of the halos, r_{12}, is at least equal to the sum of their radii, $R_{200c}(M_{h1}) + R_{200c}(M_{h2}) \leq r_{12}$. Since this prescription is usually very computationally intensive, simplified versions of HE have been suggested (e.g. Magliocchetti & Porciani, 2003; Cacciato et al., 2009) which capture the effects of HE while making the computations more efficient.

We follow a simplified approach in this appendix based on the following prescription. For a given redshift bin of our lens sample and a set of HOD parameters, we estimate the average lens halo mass, $\langle M_h \rangle$, based on Equation (3.31) and the radius $\langle R_h \rangle \equiv R_{200c}(\langle M_h \rangle)$ it corresponds to. When then set the correlation function of the central 2-halo component, $\xi_{\rm gm}^{\rm c2h}(r)$, to -1 for $r < \langle R_h \rangle$. Since the HE effect is stronger in the central 2-halo term (Cacciato et al., 2009), compared to the satellite 2-halo component $\xi_{\rm gm}^{s2h}$, we did not apply HE on $\xi_{\rm gm}^{\rm s2h}$. Figure C.1 shows the fractional differences between the fiducial constraints on the average lens halo mass, galaxy bias and satellite fraction, and the constraints from fits that take halo exclusion, as described above, into account. We find that our results do not change significantly between the two cases. We also did not find a significant difference in the χ^2 of our fits. We therefore do not include halo exclusion in our fiducial model.

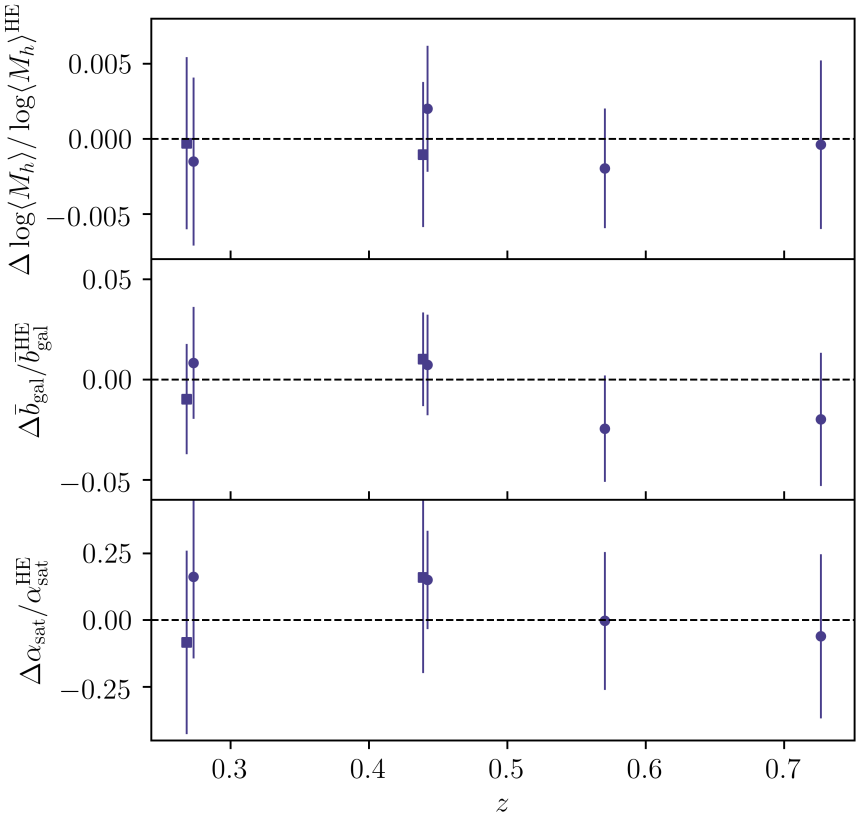


Figure C.1: Effect on our average lens halo mass, galaxy bias and satellite fraction constraints when halo exclusion is considered in our fits. This plot presents the fractional differences between the constraints from our fiducial fits and runs where we take into account halo exclusion, denoted by the "HE" superscript.

APPENDIX D

CONSTRAINTS FOR ALL MODEL PARAMETERS

In Tables D.1 and D.2 we summarise the best-fit parameters and derived quantities for the REDMAGIC and MAGLIM samples, respectively. We report the best-fit model parameters and the constraints on the average halo mass, linear galaxy bias and satellite fraction. The error bars show the 1σ posteriors.

REDMAGIC

Redshift bin	$\log(M_{\rm min}/M_\odot)$	$\log(M_1/M_\odot)$	$\sigma_{\log M}$	α	$f_{\rm cen}$	$\log(M_\star/M_\odot)$	$c_{\rm sat}/c_{\rm dm}$	$\log(M_h/M_\odot)$	$b_{\rm gal}$	$\alpha_{\rm sat}$
Lens 1 Source 3	$11.97^{+0.08}_{-0.07}$	$13.51^{+0.16}_{-0.17}$	$0.26^{+0.15}_{-0.15}$	$1.88^{+0.26}_{-0.27}$	$0.12^{+0.02}_{-0.02}$	$11.18^{+0.74}_{-0.75}$	$1.09^{+0.28}_{-0.29}$	$13.66^{+0.06}_{-0.06}$	$1.73^{+0.03}_{-0.03}$	$0.16^{+0.05}_{-0.05}$
Lens 1 Source 4	$12.13^{+0.09}_{-0.08}$	$13.64^{+0.16}_{-0.15}$	$0.50^{+0.15}_{-0.16}$	$2.06^{+0.26}_{-0.25}$	$0.13^{+0.02}_{-0.02}$	$11.09^{+0.70}_{-0.79}$	$0.99^{+0.27}_{-0.29}$	$13.67^{+0.06}_{-0.05}$	$1.71^{+0.03}_{-0.04}$	$0.13^{+0.04}_{-0.04}$
Lens 2 Source 3	$12.03^{+0.08}_{-0.08}$	$13.79^{+0.17}_{-0.16}$	$0.34^{+0.14}_{-0.16}$	$2.61^{+0.34}_{-0.33}$	$0.13^{+0.02}_{-0.02}$	$9.77^{+0.63}_{-0.61}$	$1.08^{+0.25}_{-0.26}$	$13.59^{+0.07}_{-0.07}$	$1.83^{+0.03}_{-0.03}$	$0.08^{+0.03}_{-0.04}$
Lens 2 Source 4	$12.08^{+0.08}_{-0.08}$	$13.73^{+0.12}_{-0.11}$	$0.49^{+0.14}_{-0.16}$	$2.48^{+0.25}_{-0.25}$	$0.13^{+0.01}_{-0.01}$	$9.48^{+0.64}_{-0.62}$	$1.08^{+0.22}_{-0.23}$	$13.59^{+0.05}_{-0.05}$	$1.81^{+0.03}_{-0.03}$	$0.09^{+0.02}_{-0.02}$
Lens 3 Source 4	$11.86^{+0.09}_{-0.08}$	$13.18^{+0.12}_{-0.11}$	$0.42^{+0.14}_{-0.15}$	$1.65^{+0.18}_{-0.17}$	$0.08^{+0.01}_{-0.01}$	$10.92^{+0.62}_{-0.62}$	$0.65^{+0.22}_{-0.21}$	$13.36^{+0.04}_{-0.04}$	$1.86^{+0.03}_{-0.03}$	$0.18^{+0.03}_{-0.03}$
Lens 4 Source 4	$12.16^{+0.12}_{-0.11}$	$13.26^{+0.19}_{-0.19}$	$0.46^{+0.12}_{-0.13}$	$1.59^{+0.24}_{-0.24}$	$0.06^{+0.03}_{-0.02}$	$11.01^{+0.58}_{-0.59}$	$0.71^{+0.24}_{-0.23}$	$13.27^{+0.09}_{-0.07}$	$2.12^{+0.06}_{-0.06}$	$0.19^{+0.06}_{-0.06}$

Table D.1: Statistical analysis summary of the chains for Y3 unblind REDMAGIC data (30 data points) using the fiducial cosmology; the average halo masses shown here use the $200\rho_m$-based definition. The error bars correspond to the 1σ posteriors.

MagLim

Redshift bin	$\log(M_{\rm min}/M_\odot)$	$\log(M_1/M_\odot)$	$\sigma_{\log M}$	α	$\log(M_\star/M_\odot)$	$c_{\rm sat}/c_{\rm dm}$	$\log(M_h/M_\odot)$	$b_{\rm gal}$	$\alpha_{\rm sat}$
Lens 1 Source 3	$11.74^{+0.05}_{-0.05}$	$13.32^{+0.19}_{-0.20}$	$0.27^{+0.12}_{-0.12}$	$1.66^{+0.31}_{-0.30}$	$11.26^{+1.09}_{-1.10}$	$0.41^{+0.23}_{-0.22}$	$13.44^{+0.07}_{-0.07}$	$1.57^{+0.03}_{-0.03}$	$0.14^{+0.04}_{-0.04}$
Lens 1 Source 4	$11.76^{+0.08}_{-0.07}$	$13.41^{+0.20}_{-0.21}$	$0.29^{+0.15}_{-0.15}$	$1.74^{+0.30}_{-0.31}$	$9.38^{+0.86}_{-0.89}$	$0.76^{+0.27}_{-0.27}$	$13.43^{+0.09}_{-0.10}$	$1.54^{+0.03}_{-0.03}$	$0.12^{+0.06}_{-0.05}$
Lens 2 Source 3	$11.96^{+0.07}_{-0.06}$	$13.44^{+0.12}_{-0.11}$	$0.26^{+0.14}_{-0.14}$	$1.82^{+0.22}_{-0.21}$	$10.83^{+1.08}_{-1.12}$	$0.63^{+0.30}_{-0.28}$	$13.46^{+0.04}_{-0.04}$	$1.84^{+0.04}_{-0.04}$	$0.14^{+0.03}_{-0.03}$
Lens 2 Source 4	$11.91^{+0.08}_{-0.07}$	$13.42^{+0.12}_{-0.13}$	$0.30^{+0.15}_{-0.15}$	$1.85^{+0.17}_{-0.18}$	$8.50^{+0.94}_{-0.94}$	$1.07^{+0.28}_{-0.26}$	$13.45^{+0.04}_{-0.04}$	$1.82^{+0.05}_{-0.04}$	$0.13^{+0.05}_{-0.04}$
Lens 3 Source 4	$11.88^{+0.09}_{-0.09}$	$12.84^{+0.31}_{-0.30}$	$0.21^{+0.14}_{-0.14}$	$1.24^{+0.24}_{-0.23}$	$8.59^{+0.96}_{-0.96}$	$0.21^{+0.25}_{-0.24}$	$13.27^{+0.06}_{-0.05}$	$1.99^{+0.04}_{-0.04}$	$0.37^{+0.13}_{-0.13}$
Lens 4 Source 4	$11.82^{+0.10}_{-0.10}$	$13.44^{+0.17}_{-0.15}$	$0.31^{+0.14}_{-0.15}$	$2.29^{+0.24}_{-0.24}$	$8.53^{+1.06}_{-1.04}$	$1.19^{+0.29}_{-0.31}$	$13.31^{+0.05}_{-0.05}$	$2.01^{+0.04}_{-0.05}$	$0.09^{+0.03}_{-0.04}$

Table D.2: Similar to Table D.1 but for the MagLim sample.

APPENDIX E

MODEL COMPLEXITY

In Section 3.7.3 we discuss how adding complexity to our model changes our results. In this appendix we provide details on our tests that led us to deciding what our fiducial framework is in this paper.

In Figure E.1 we show for all REDMAGIC redshift bins the fractional differences between the best-fit γ_t using the HOD-only model and the HOD-only model plus one additional contribution at a time. This plot shows how adding various terms to γ_t changes the best-fit model as a function of θ, providing more information than the difference in χ^2. Figure E.2 shows the constraints on the average halo mass, galaxy bias and satellite fraction corresponding to these fits, with the vertical bands representing the constraints from our fiducial runs and each point shows the constraints from adding an additional contribution to the model. In the same plot we also report in parenthesis the difference in goodness-of-fit as the difference in the reduced χ^2 between each tested model and the HOD-only fits.

Although adding complexity to the basic HOD-only model is informative, we point out that interactions between additional terms, when more than one of them are considered, can have a much different net effect. Due to the large number of combinations we could explore, it was not feasible to do this full analysis, but we also note that we did not have strong indications that specific combinations of model components lead to radically different results in our fits or halo property constraints. To test for that, as a complement to our tests in Figure E.2, we have performed a test where we start from the full model which includes all additional contributions from Section 3.3, removing one component at a time and re-fitting the data. Figure E.3 presents our findings from this test.

Below we discuss the effect of each contribution to the model fits separately when we simply add it to the basis of only HOD or remove it from the full model with all γ_t terms.

Point mass (PM): We find that the PM component mostly affects the small scales in the

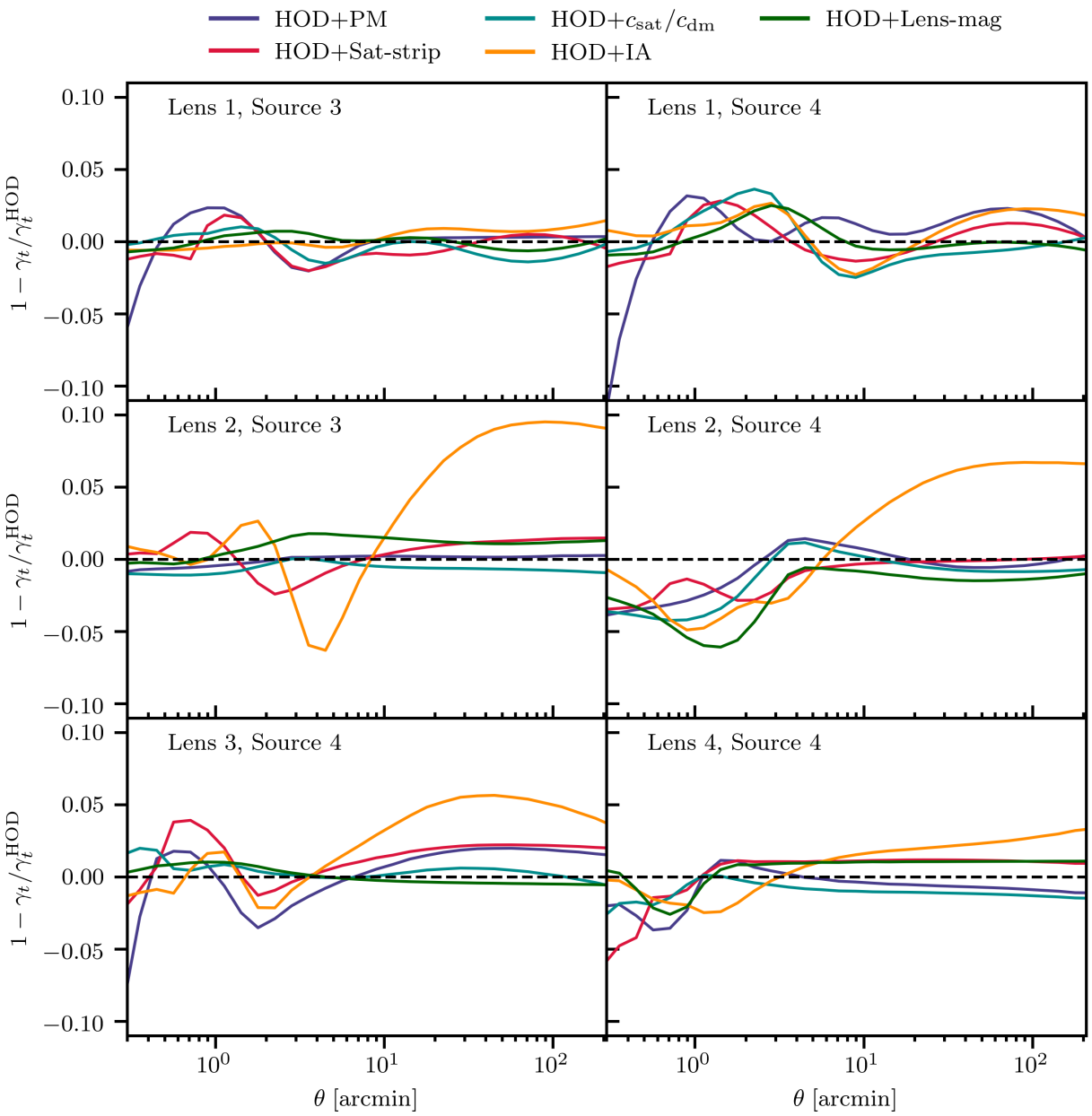


Figure E.1: Comparing the basic HOD-only best-fit γ_t model prediction for all REDMAGIC lens-source redshift bins to the best-fit γ_t after considering additional model complexity.

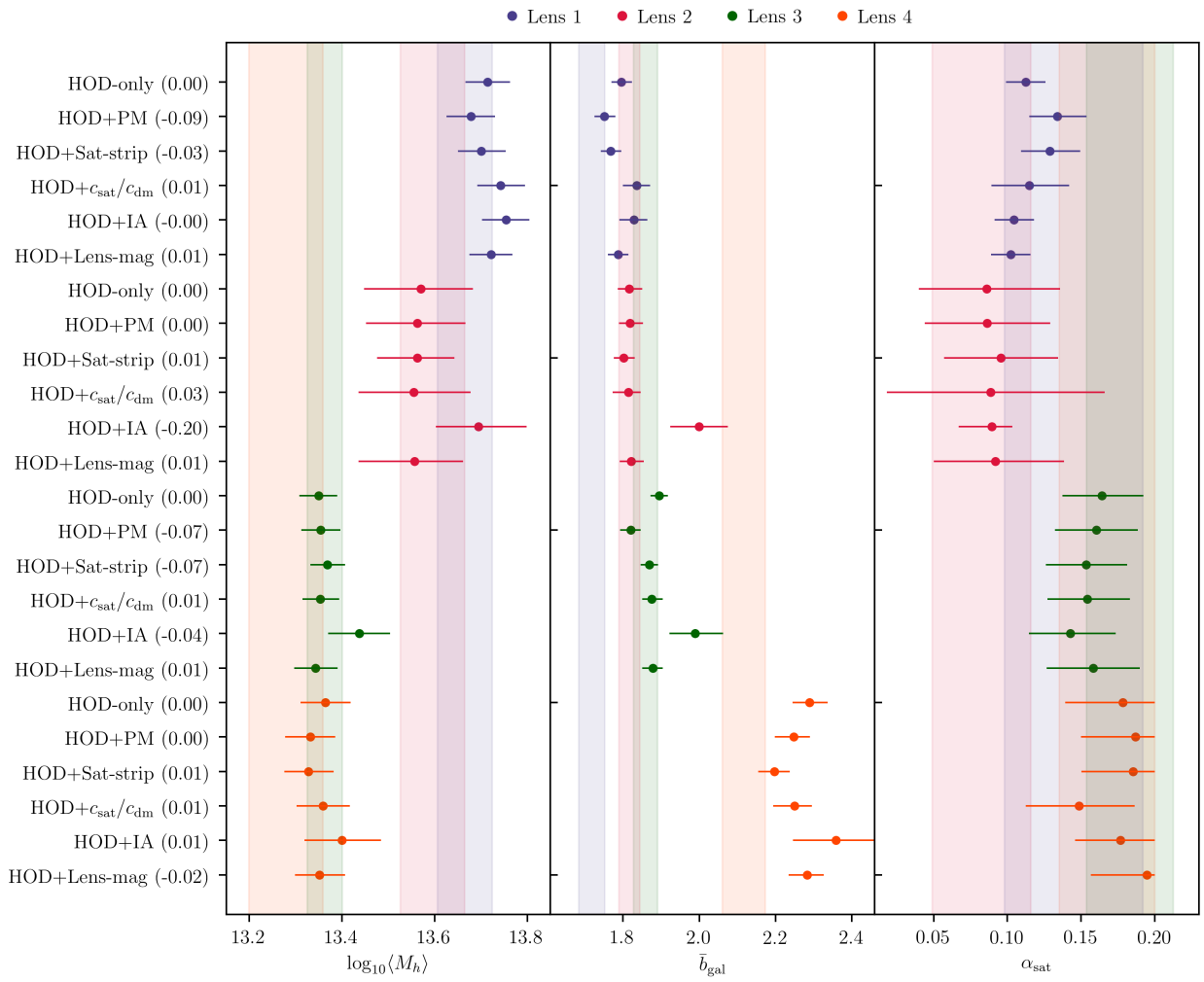


Figure E.2: Testing the robustness of the halo properties to adding complexity to our model. We begin from our basic HOD-only model and we add one additional component to it at a time. In parenthesis we report the difference in the reduced χ^2 between the best-fit HOD-only and the tested model fit. The vertical bands correspond to our constraints from the fiducial model and are added here for a direct comparison with our tests. Note that, to reduce the size of this figure we have combined bins with the same lenses and different sources by presenting the mean of the best-fit values and, to be conservative, the maximum of the error bars.

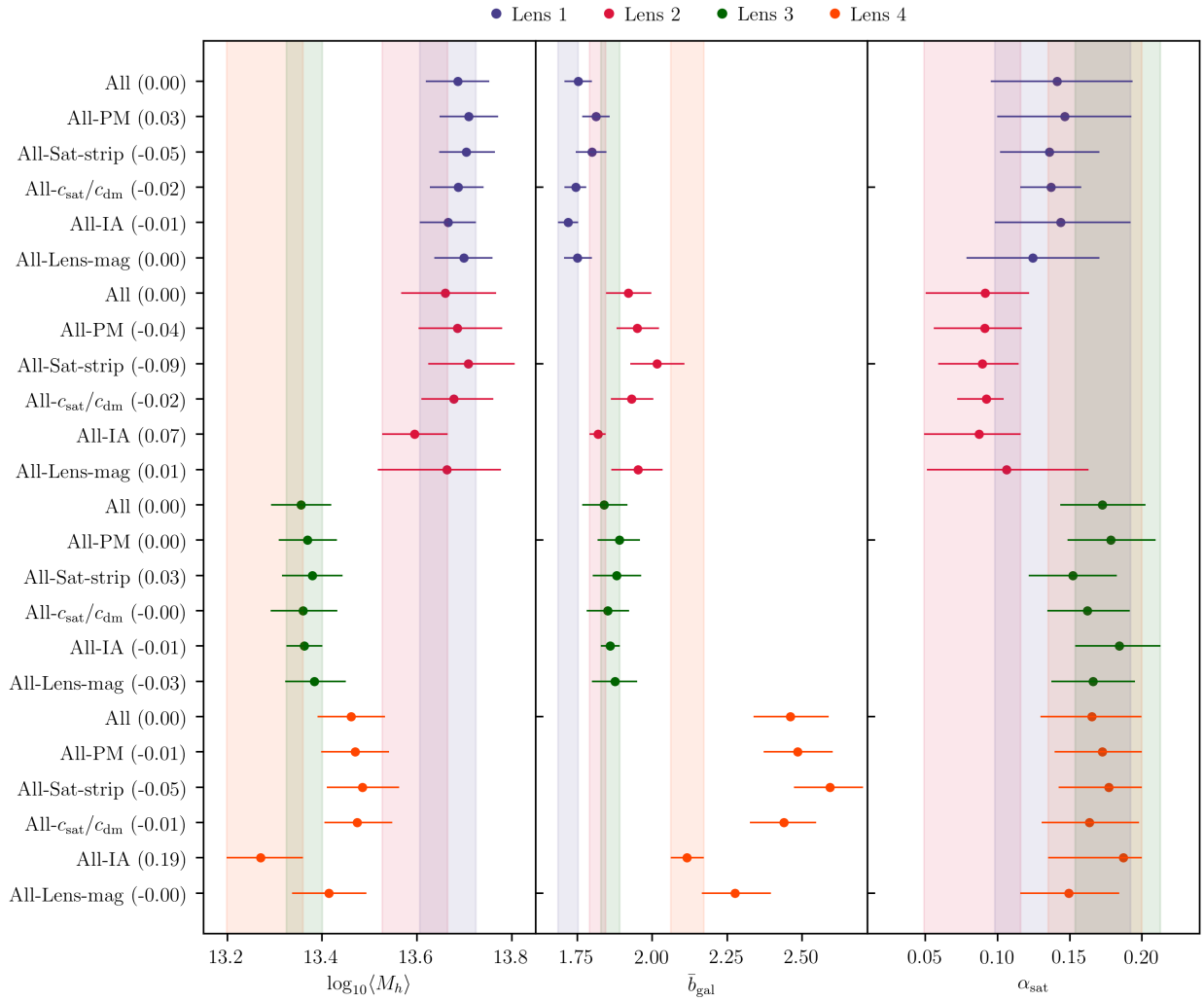


Figure E.3: Testing the robustness of the halo properties to adding complexity to our model. We begin from our basic HOD-only model and we add one additional component to it at a time. In parenthesis we report the difference in the reduced χ^2 between the best-fit from the runs with all components included and each tested model. The vertical bands correspond to our constraints from the fiducial model and are added here for a direct comparison with our tests. Note that, to reduce the size of this figure we have combined bins with the same lenses and different sources by presenting the mean of the best-fit values and, to be conservative, the maximum of the error bars.

first lens bin, with the largest effect being $\sim 10\%$ at the smallest angular scales. This is due to the fact that the smallest angular bins of that redshift bin correspond to the smallest physical scales we consider in this work. We have included PM in our fiducial model as a conservative approach to account for modeling uncertainties at scales below what we measure.

Satellite strip: The effect from striping of satellite galaxies to γ_t can make a quite significant change on the constraints in some redshift bins, especially in the last one. This component also introduces a nice physical picture to our modeling – it captures the tidal interactions between the central galaxy and the substructure in the lens halos. We have included this term in our fiducial model.

Satellite galaxy concentration parameter: Allowing for the concentration parameter for the spatial distribution of the satellite galaxies to vary mostly affects the bias constraints. This is because $a = c_{\rm sat}/c_{\rm dm}$ modifies the shape of the satellite terms in the 1-halo regime making the model more flexible and able to better fit small and large scales at the same time, which forces the large-scale bias to change and adjust accordingly. Furthermore, as discussed in Section 3.3.1, there is good motivation to allow the concentration of the satellite-galaxy distribution to be different from that of the dark matter's distribution. We have included this term in our fiducial model.

Lens magnification: The effect of lens magnification becomes stronger at higher redshift bins. Especially in the [Lens 4, Source 4] bin it can have a large impact on the final constraints, even on the halo mass, which is overall the most robust to changes in the model. Furthermore, magnification of lenses is well-motivated and its modeling is straightforward. Our magnification model only depends on fixed coefficients, as discussed in Section 3.3.5 and therefore does not introduce free parameters. We have included this term in our fiducial model.

Intrinsic alignment: Despite the uncertainty in the IA model in the 1-halo term (see discussion in Section 3.3.6), we test here this term's contribution to our fits. We find that

the change in the best-fit model can be heavily impacted as a function of angular scale by this component. The constraints can also be significantly affected by IA. In particular, lens bin 2 is mostly affected by the addition of IA to our basis HOD model, and the largest effect is noticed on large scales. This is caused by a combination how much overlap in the $n(z)$ distributions of the lenses and sources there is and how much of the 1-halo component we can observe in lens bin 2. Since a significant number of points in that bin's measurements belong to the 1-halo regime, if the HOD-only model cannot describe both small and large scales well at the same time, the added model flexibility from the inclusion of IA essentially accounts for that and improves the model fit. However, after adding other needed model complexity, besides IA, this effect is ameliorated and IA becomes negligible for the specific lens-source bin combinations we consider in this work. Therefore, and given that we do not trust that our modeling of IA is accurate at small scales, we decide to not take this term into account as part of our fiducial framework.

As a general note, we find that the constraints in the fourth bin are mostly affected by additional contributions to γ_t, while overall the bias constraints are the most sensitive to changes in our model. We note that our fiducial framework is effectively the "All-IA" model.

APPENDIX F

EQUIVALENCE BETWEEN "PROJECT-OUT" AND POINT-MASS MARGINALIZATION

In this appendix we show that the "Project-out" and the Point-mass marginalization methods are mathematically equivalent if the prior for the point-mass is infinitely wide. Let us define $\mathbf{m}$ to be the model prediction and $\mathbf{d}$ to be the data vector.

For the PM marginalization case we have:

$$\chi^2_{\rm PM} = (\mathbf{m}-\mathbf{d})^T \mathbf{N}^{-1} (\mathbf{m}-\mathbf{d}) \ .$$

However, in the infinite-prior case it is:

$$\mathbf{N}^{-1} = \mathbf{C}_\gamma^{-1} - \mathbf{C}_\gamma^{-1}\mathbf{A}(\mathbf{A}^T\mathbf{C}_\gamma^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}_\gamma^{-1} = \mathbf{C}_\gamma^{-1}(\mathbf{I}-\mathbf{P}) \ ,$$

which results in

$$\chi^2_{\rm PM} = (\mathbf{m}-\mathbf{d})^T \mathbf{C}_\gamma^{-1} \mathbf{M} (\mathbf{m}-\mathbf{d}) \ . \tag{F.1}$$

So, $\mathbf{N}^{-1}(\mathbf{m}-\mathbf{d})$ exactly simply removes the contribution from the $1/\theta^2$ mode from the model and data vectors, like the project-out approach. The project-out method is by construction only removing that mode from the data vector and does nothing else. Therefore, the two methods are equivalent in the above limit.

APPENDIX G

SCALE CUTS

In Fig. G.1 we show the result of the scale cuts tests for the successful scenario. We have found that scales cuts of $w > 12$ Mpc/h and $\gamma_t > 8$ Mpc/h are sufficient in the LSST Y1 setup to recover unbiased results when inputting a contaminated data vector with non-linear galaxy bias and baryonic effects. We obtain that in this case the difference in the $S_8 - \Omega_m$ plane between the fiducial and contaminated data vectors is 0.09σ which is below the threshold of 0.3σ, following the same procedure as in Krause et al. (2021a). We compare posteriors using the fiducial vs. the contaminated data vectors (instead of with the input cosmology) to remove any projection effects[1] impact on this test. We have also tried using $w > 8$ Mpc/h, $\gamma_t > 8$ Mpc/h scale cuts, which produced a difference of 0.46σ and $w > 12$ Mpc/h and $\gamma_t > 6$ Mpc/h cuts, which produced a difference of 0.93σ in the same $S_8 - \Omega_m$ plane. Note that these combinations of scale cuts are an arbitrary choice, and applications on actual data might want to optimize these choices.

1. Projection effects are residual differences between the input cosmology and the posteriors under ideal conditions (when the input data vector and model are the same) due to having broad parameter spaces.

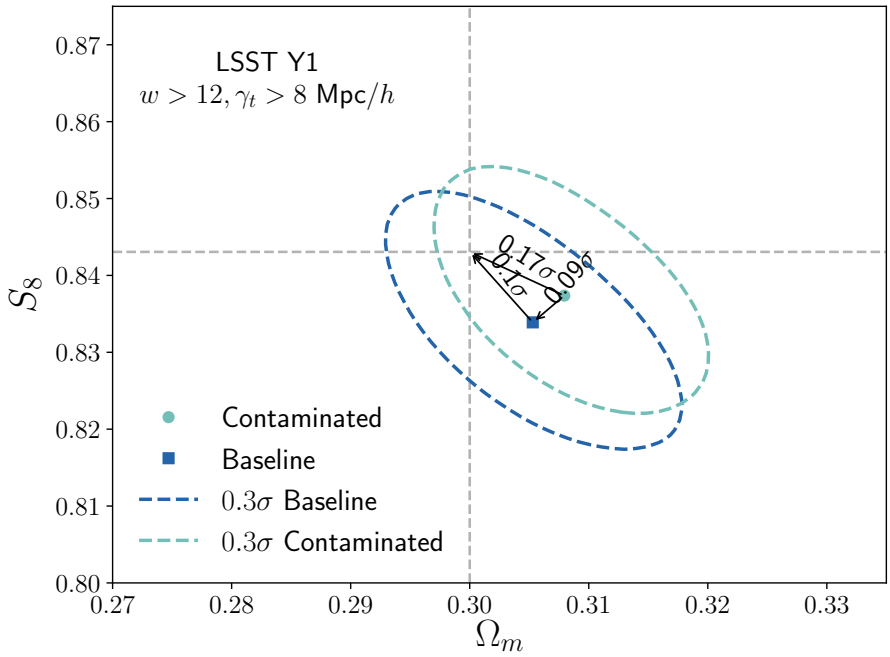


Figure G.1: $S_8 - \Omega_m$ plane showing the differences in the posteriors using either a fiducial input data vector or a contaminated one with baryonic and non-linear galaxy bias effects. The dashed gray lines mark the input fiducial cosmology. Comparing the contaminated and the baseline posteriors using different sets of scale cuts we have determined that $w > 12$ Mpc/h, $\gamma_t > 8$ Mpc/h cuts are sufficient for the LSST Y1 setup to recover unbiased results. Specifically, these cuts produce a difference of 0.09 σ in the $S_8 - \Omega_m$ plane, which is below the threshold of 0.3σ following the criteria from Krause et al. (2021a).

APPENDIX H

MITIGATION SCHEMES VS. NO MITIGATION

In this appendix we address the following question: How much constraining power do we gain by applying one of the mitigation schemes vs. not applying any of them and using fewer scales? To perform this comparison we choose the point-mass case that includes geometrical information since that is what we used to define the fiducial scale cuts, as described in Sec. 6.3.3. We perform this comparison for the LSST Y1 simulated analysis. We show the results in Fig. H.1. There we compare the posteriors between using the point-mass mitigation scheme and without applying any mitigation with the following two sets of scale cuts:

1. $w > 12$ Mpc/h, $\gamma_t > 8$ Mpc/h: the same scale cuts needed to recover unbiased constraints when applying the point-mass marginalization scheme including geometrical information.

2. $w > 12$ Mpc/h, $\gamma_t > 40$ Mpc/h: scale cuts needed to recover unbiased cosmological constraints without using any mitigation scheme, following the same criteria described in Sec. G. To obtain them, we have chosen to keep the galaxy clustering scale cut unchanged and increase the tangential shear cut until we recover unbiased results. Under this setup, we find a 0.17σ difference in the $S_8 - \Omega_m$ plane, while we find a 0.35σ difference if we use γ_t>32 Mpc/h instead, which does not meet the criteria.

We find that S_8 is $\sim$1.3 times more constraining when using the point-mass marginalization scheme vs. when not using any mitigation scheme and using larger scale cuts.

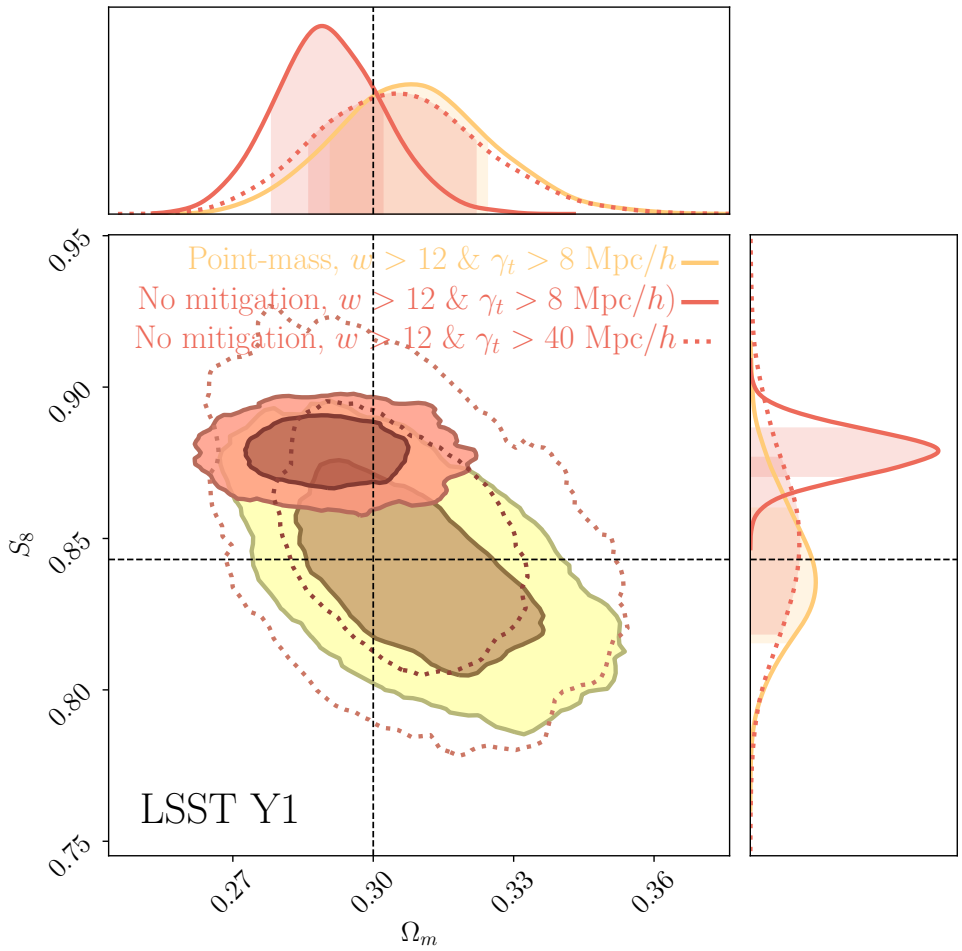


Figure H.1: We compare the constraining power when applying a mitigation scheme (in this case the point-mass marginalization) vs. when not applying any scheme and using less scales for the galaxy-galaxy lensing probe. We conclude that using a marginalization scheme yields $\sim$1.3 more constraining power on the S_8 parameter assuming a LSST Y1 simulated scenario.

APPENDIX I

EFFECT ON THE INTRINSIC ALIGNMENT PARAMETERS

In Fig. I.1 we show the posteriors for the tidal alignment and tidal torque (TATT) 5-parameter intrinsic alignment model, at the top for the LSST Y1 simulated case and at the bottom applied to DES Y3 data, in both cases for a 2×2pt analysis without including the shear-ratio likelihood. Using the simulated data in the LSST Y1 setup, we find that using the extra geometric information in the point-mass marginalization, i.e. the case labelled as "Point-mass (free per z_l^i)" and described in Sec. 6.2.3 is beneficial to constrain the intrinsic alignment parameters. In particular we find the parameter describing the galaxy bias of the source sample b_{TA} is more constrained, as well as the parameters affecting the tidal torque part of the model (a_2 which describes the amplitude of the IA effect and α_2 that modulates its redshift evolution). On the Y3 data, we find that the biggest gain in constraining power is in the a_2 parameter.

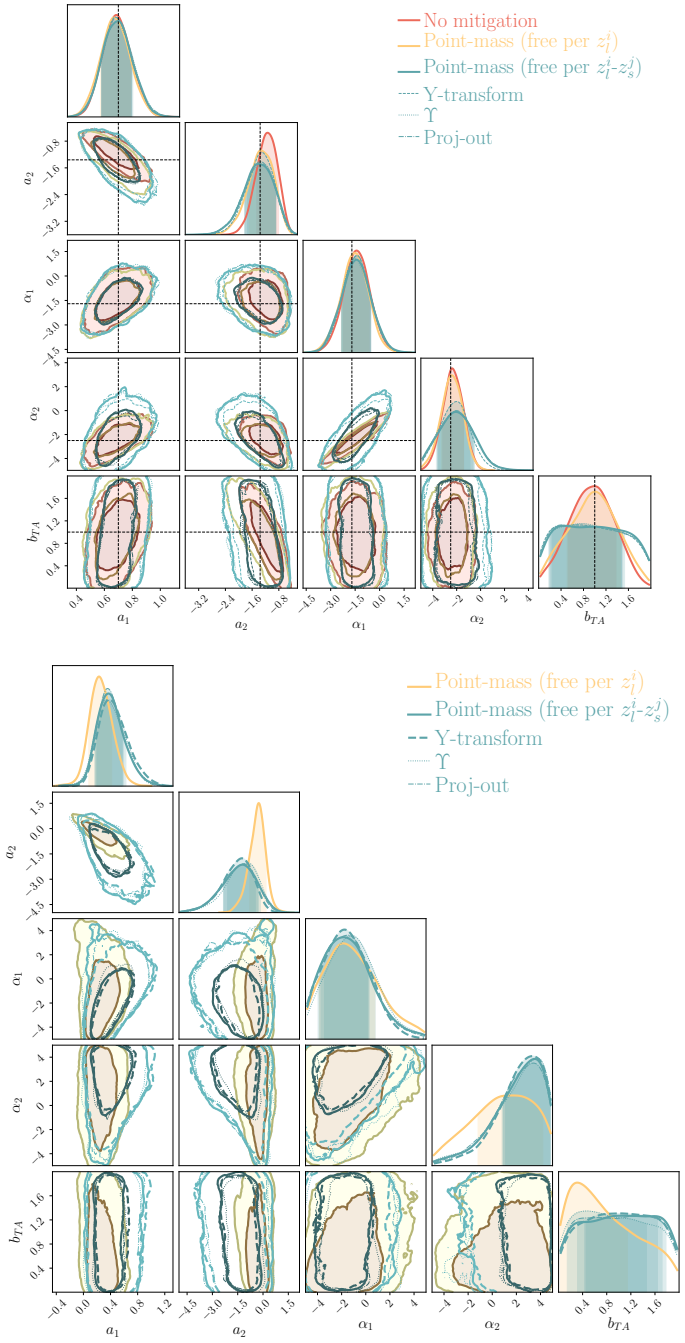


Figure I.1: **Top:** Intrinsic alignment TATT parameter posteriors obtained from an input galaxy clustering and galaxy-galaxy lensing data vector (2×pt) with non-linear bias and baryonic contamination with a LSST Y1 covariance, comparing the different methodologies to localize the tangential shear measurements. **Bottom**: Analogous comparison applied to DES Y3 data, for the MagLim 2×pt analysis without the shear-ratio likelihood.

APPENDIX J

COMPARISON BETWEEN DES Y3 AND LSST Y1

In Fig. J.1 we compare the constraining power between the DES Y3 and LSST Y1 setups, which besides being interesting on its own also provides some basic validation of the LSST Y1 covariance that we compute with CosmoCOV. First, we compare the fiducial LSST Y1 simulated analysis with the fiducial scales of $w > 12$ Mpc/h and $\gamma_t > 8$ Mpc/h, with the scales used in the DES Y3 data of $w > 8$ Mpc/h and $\gamma_t > 6$ Mpc/h. We observe that the degradation in the constraining power coming only from the differences in the scales is quite small. Then, we compare the size of the LSST Y1 contours with the simulated DES Y3 analysis, which was used in Krause et al. (2021a) to determine the scale cuts for that case. The only two differences between the contours labelled "LSST Y1 sim, with DES Y3 scales" and the ones labelled "DES Y3 sim" are the covariances and the input redshift distributions, as described in Sec. 6.3.1. Finally we also compare the DES Y3 simulated analysis with the actual final data DES Y3 2×2pt results, which use a different set of priors as the rest of the chains, had an updated covariance accounting for the best-fit parameters and are not centered at the same cosmology.

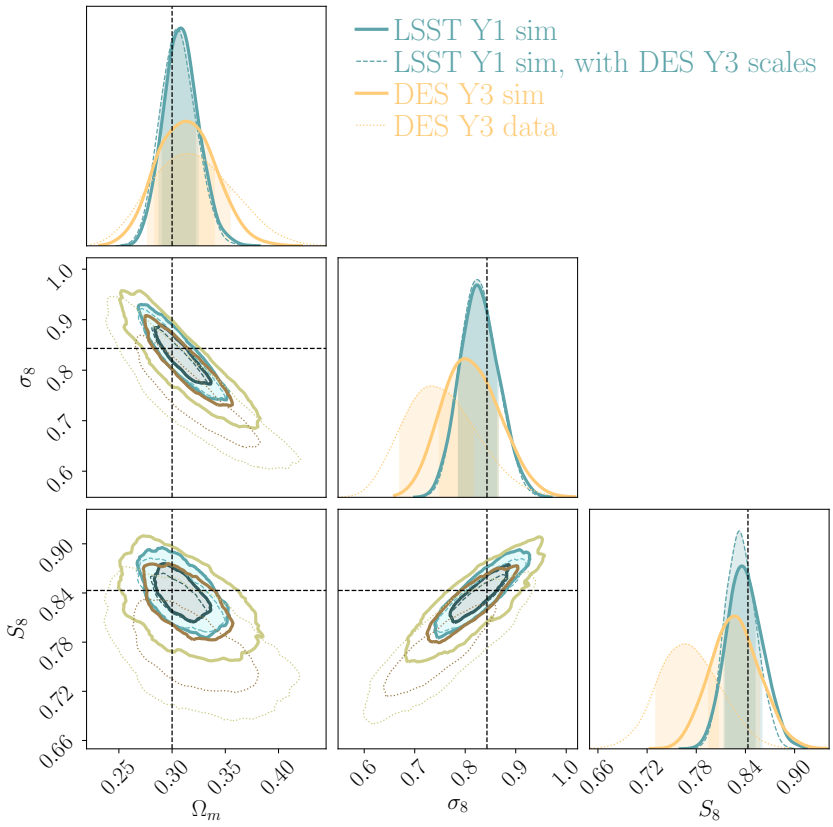


Figure J.1: Comparison of the constraining power between LSST Y1 and DES Y3 for a 2×2pt simulated analysis using the fiducial model data vectors. For LSST Y1 we use scales $w > 12$ Mpc/h, $\gamma_t > 8$ Mpc/h, and for DES Y3 $w > 8$ Mpc/h, $\gamma_t > 6$ Mpc/h. We also include the data results from DES Y3 2×2pt.

APPENDIX K

QUANTIFYING TENSIONS BETWEEN CORRELATED DATA SETS IN COSMOLOGY

Despite the remarkable successes of the ΛCDM model to explain a wide range of observed phenomena, as the precision of different experiments has increased, so too has the statistical significance of discrepancies between their inferences (for a recent review see Verde et al., 2019). Such discrepancies deserve close attention since they may hint at the existence of new physical phenomena or to the presence of residual systematic effects that are not yet understood. In addition, cosmological data sets have become increasingly complex to the point that understanding whether different probes agree or not requires the use of complex statistical tools (Marshall et al., 2006; Feroz et al., 2008; March et al., 2011; Amendola et al., 2013; Verde et al., 2013; Bennett et al., 2014; Martin et al., 2014; Karpenka et al., 2015; Larson et al., 2015; Addison et al., 2016; Raveri, 2016; Seehars et al., 2014, 2016; Grandis et al., 2016; Addison et al., 2018; Nicola et al., 2019; Weiland et al., 2018; Huang et al., 2018; Raveri & Hu, 2019; Motloch & Hu, 2018, 2019; Adhikari & Huterer, 2019; Kerscher & Weller, 2019; Handley & Lemos, 2019; Huang et al., 2019; Lin & Ishak, 2019).

In this appendix we briefly discuss the theory behind developing such tools to quantify the agreement or disagreement among cosmological data sets, when correlations between them are accounted for. This, as well as an illustrative example of applying these tools and more, is published in Raveri et al. (2020). Here we simply present the basic ideas of the statistical tools we considered.

The main focus is to extend the Concordance-Discordance Estimators (CDE's) introduced for uncorrelated data sets in Raveri & Hu (2019) to correlated data sets. In particular, we discuss estimators quantifying parameter shifts between two correlated data sets and goodness of fit loss when two data sets are joined together. We analyze these estimators under the Gaussian linear model (GLM), assuming Gaussianity of the data and model parameters.

We also comment on mitigation against non-Gaussianities.

When considering correlated data sets we follow two different strategies. In the first approach we consider the two disjoint data sets and build estimators working on their separate parameter inference though properly including their correlation. In the second one, we always consider the joint data set but fit the two parts of the data set with different cosmological parameters. This approach was also employed in Zhang et al. (2005); Chu & Knox (2005); Wang et al. (2007); Abate & Lahav (2008); Ruiz & Huterer (2015); Bernal et al. (2016); Köhlinger et al. (2019); Lemos et al. (2019). We call the first strategy the *data split* one while we refer to the second one as the *parameter split* strategy. The data and parameter split techniques are equivalent when data sets are uncorrelated but different when they are correlated.

K.1 The Gaussian linear model

In this section we gather some basic definitions that we will later use throughout the paper. For an in depth discussion of the Gaussian Linear Model (GLM) we refer the reader to Raveri & Hu (2019).

We denote the multivariate Gaussian distribution in N dimensions with mean $\bar{\theta}$ and covariance $\mathcal{C}$ as $\mathcal{N}_N(\theta;\bar{\theta},\mathcal{C})$. For a given data set, D, described by a model $\mathcal{M}$ that depends on a set of parameters θ, the posterior probability distribution of the parameters is given by:

$$P(\theta|D,\mathcal{M}) = \frac{\mathcal{L}(\theta)\Pi(\theta)}{\mathcal{E}} \,, \tag{K.1}$$

where the likelihood is the probability of the data at any given choice of parameters $\mathcal{L}(\theta) = P(D|\theta,\mathcal{M})$ and any prior knowledge is encoded in $\Pi(\theta)$. The normalization of the posterior, $\mathcal{E} \equiv P(D|\mathcal{M})$, is the evidence that provides the probability distribution of the data given the model $\mathcal{M}$.

In this section, we assume that the prior distribution is Gaussian in the model parameters, $\Pi(\theta) = \mathcal{N}_N(\theta; \theta_\Pi, \mathcal{C}_\Pi)$, with an unbiased mean θ_Π and covariance $\mathcal{C}_\Pi$. As discussed in Raveri & Hu (2019) this is a good choice to use in practice as it allows us to treat Gaussian priors on nuisance parameters exactly and models the most relevant features of informative flat priors: the scale of the prior and its central value.

We further assume that the likelihood is a Gaussian distribution in data space, $\mathcal{L}(\theta) = \mathcal{N}_d(x; m, \Sigma)$ and we denote by d the number of data points x and Σ their covariance matrix. The mean of the distribution is given by the model prediction, $m(\theta)$.

The GLM assumes that one can linearly expand the model prediction, $m(\theta)$, around a given parameter point. Since we are working with Gaussian priors, for simplicity, in the following we assume that the linear model expansion point is the prior center θ_Π and we can write

$$m(\theta) \approx m_\Pi + M(\theta - \theta_\Pi) \,, \tag{K.2}$$

where $m_\Pi \equiv m(\theta_\Pi)$ and $M \equiv (\partial m/\partial\theta)|_{\theta_\Pi}$ is the Jacobian of the transformation between parameter and data space.

Given the model prediction m_Π, the residual of a randomly chosen data point x, henceforth $X \equiv x - m_\Pi$, can be projected onto a component along the linear model, $\mathbb{P}X$, and another component orthogonal to it, $X - \mathbb{P}X = (\mathbb{I} - \mathbb{P})X$. The projector can be thought of as a two-step process. The first step is to construct the linear combinations of data, namely $\tilde{M}X$, that give the parameter estimates: $\theta - \theta_\Pi = \tilde{M}X$, where $\tilde{M} = \mathcal{C}M^T\Sigma^{-1}$, with

$$\mathcal{C} = \langle(\theta - \theta_\Pi)(\theta - \theta_\Pi)^T\rangle = (M^T\Sigma^{-1}M)^{-1} \,, \tag{K.3}$$

as the parameter covariance or inverse Fisher matrix. Then, as a second step, given the parameter estimates we transform back into data space using the Jacobian, $M(\theta - \theta_\Pi) =$

$M\tilde{M}X \equiv \mathbb{P}X$. Thus $\mathbb{P} \equiv M\tilde{M}$ is the full projector:

$$\mathbb{P} = M\mathcal{C}M^T\Sigma^{-1}\,, \tag{K.4}$$

and $(\mathbb{I} - \mathbb{P})$ is its complement.

In the GLM the maximum likelihood (ML) is given by:

$$\ln \mathcal{L}_{\max} = -\frac{1}{2}X^T(\mathbb{I}-\mathbb{P})^T\Sigma^{-1}(\mathbb{I}-\mathbb{P})X - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|)\,, \tag{K.5}$$

where we used $|\cdot|$ to denote the determinant of a matrix. Notice that the first line of Eq. (K.5) contains all the pieces that depend on the data while the second one contains normalization constants that are often neglected. The parameters corresponding to the maximum likelihood model are given by:

$$\theta_{\rm ML} = \tilde{M}(x - m_\Pi + M\theta_\Pi)\,. \tag{K.6}$$

Over realizations of data, the maximum likelihood parameters are distributed as $\mathcal{N}_N(\theta;\theta_\Pi,\mathcal{C})$. These expressions refer to the true maximum likelihood of a model and they should be obtained without reference to the prior; θ_Π appears here due to the assumption that the prior mean is unbiased.

The maximum posterior (MAP) parameters combine the ML parameters with the prior:

$$\theta_p = \mathcal{C}_p(\mathcal{C}_\Pi^{-1}\theta_\Pi + \mathcal{C}^{-1}\theta_{\rm ML})\,, \tag{K.7}$$

where $\mathcal{C}_p^{-1} = \mathcal{C}_\Pi^{-1} + \mathcal{C}^{-1}$. Under the GLM the maximum posterior parameters are distributed as $\mathcal{N}(\theta;\theta_\Pi,\mathcal{C}_p)$.

Within the GLM the probability of the data, i.e. the evidence, is Gaussian distributed for the Gaussian priors that we consider and is given by $\mathcal{E} = \mathcal{N}_d(x; m(\theta_\Pi), \Sigma + M\mathcal{C}_\Pi M^T)$.

We define all the statistics, Q, that we discuss in this paper to follow the convention that: if $P(Q > Q^{\rm obs})$ approaches zero then the observed value lies in the tail of the distribution that we would associate with a tension; if it approaches one the observed value would be in the tail associated with excess confirmation.

K.2 Impact of correlations

The stronger the correlation between two data sets the more crucial it is to use statistical tools which take these correlations into account. In multiple dimensions the correlation strength can be quantified by means of the Canonical Correlation Analysis (CCA) Hotelling (1936). CCA allows us to understand the change in parameter variances. The maximum error that we would make on the determination of the parameter variance, with respect to the full joint estimate, if we were to neglect correlations is given by:

$$\max\left(\frac{\sigma^2_{\rm no\,corr}}{\sigma^2_J}\right) = \frac{1}{1 \pm \rho_{12}}\,, \tag{K.8}$$

where $\sigma^2_{\rm no\,corr}$ is the wrong estimate of the variance of one parameter obtained neglecting correlations, σ^2_J is the correct estimate of the variance, accounting for all data correlations, and ρ_{12} is the maximum correlation coefficient.

As discussed in Raveri & Hu (2019) the value of the likelihood at MAP, $\hat{\theta}_p$, can be used as a goodness of fit measure to test the consistency of a data set with the model at hand. The MAP measure for goodness of fit is then given by:

$$Q_{\rm MAP} \equiv -2\ln\mathcal{L}(\hat{\theta}_p) + 2\left\langle \ln\mathcal{L}(\hat{\theta}_p)\right\rangle_D + \langle Q_{\rm MAP}\rangle_D \tag{K.9}$$

where the average is over data realizations. For Gaussian priors $Q_{\rm MAP}$ is distributed as a sum of Gamma distributed variables which can be (conservatively) approximated by a

chi-square distribution of $d - N_{\rm eff}$ degrees of freedom, $Q_{\rm MAP} \sim \chi^2(d - N_{\rm eff})$, where

$$N_{\rm eff} = N - {\rm tr}[\mathcal{C}_\Pi^{-1}\mathcal{C}_p] \tag{K.10}$$

is the number of effective parameters that are being constrained by the data over the prior, with N being the total number of model parameters. Notice that there might be cases where the data covariance matrix is singular. In these cases one needs to compute the number of data points as the rank of the covariance, $d = {\rm rank}(\Sigma)$.

K.3 Data splits and parameter splits

We can split a correlated data set by taking partitions of the joint dataset, that we denote as D_J, in two pieces that we indicate with D_1 and D_2. Hereafter we denote quantities that refer to the joint data set with J and quantities referring to the first and second data sets with the subscript 1 and 2 respectively. In Section K.6 at the end we discuss the natural generalization to an arbitrary number of splits. We indicate the joint $D_J = D_1 \cup D_2$ data covariance as

$$\Sigma_J \equiv \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix} , \tag{K.11}$$

which is in general not block diagonal. Since the full covariance has to be symmetric then $\Sigma_1 = \Sigma_1^T$, $\Sigma_2 = \Sigma_2^T$ and $\Sigma_{21} = \Sigma_{12}^T$. Notice that Σ_1, Σ_2 and Σ_J have all to be symmetric and positive definite.

Within the GLM this data separation projects onto parameter space through the single and joint Jacobian matrices that are related by $M_J^T = (\partial m_1^T/\partial\theta, \partial m_2^T/\partial\theta) \equiv (M_1^T, M_2^T)$.

The GLM estimate of the ML parameters in this case is given by:

$$\begin{aligned}\theta^{\mathrm{ML}}_{aS} &= \mathcal{C}_{aS} M_a^T \Sigma_a^{-1}(x_a - m_{\Pi a} + M_a \theta_\Pi)\,,\\ \theta^{\mathrm{ML}}_{J} &= \mathcal{C}_{J} M_J^T \Sigma_J^{-1}(x_J - m_{\Pi J} + M_J \theta_\Pi)\,,\end{aligned} \tag{K.12}$$

with $a \in \{1, 2\}$. Hereafter we denote quantities that are obtained within the single analysis of the split data sets with S. The maximum posterior parameters are obtained by adding on top of these estimates the prior, as in Eq. (K.7). As we can see the inference of the parameters for the separate data splits depends only on the given data set, while their joint inference is influenced by the complementary set, through the correlation between the two. For this reason it is not possible, in presence of data correlations, to reconstruct the joint ML parameters as a linear combination of parameter quantities that live in the single parameter space.

We can still, however, compute the covariance between different data split parameters both at the ML and MAP level as:

$$\begin{aligned}\mathrm{cov}(\theta^{\mathrm{ML}}_{1S}, \theta^{\mathrm{ML}}_{2S}) &= \mathcal{C}_{1S} M_1^T \Sigma_1^{-1} \Sigma_{12} \Sigma_2^{-1} M_2 \mathcal{C}_{2S}\,,\\ \mathrm{cov}(\theta^{p}_{1S}, \theta^{p}_{2S}) &= \mathcal{C}_{p1S} \mathcal{C}_\Pi^{-1} \mathcal{C}_{p2S} + \mathcal{C}_{p1S} M_1^T \Sigma_1^{-1} \Sigma_{12} \Sigma_2^{-1} M_2 \mathcal{C}_{p2S}\,.\end{aligned} \tag{K.13}$$

As we can see these depend on both parameter space and data space quantities to take into account that the single parameter covariances do not include correlation contributions.

As an alternative strategy we can think of the split as originating in parameter space, describing the two parts of the joint data set with different sets of parameters and always fitting the joint likelihood. We denote with the subscript C quantities that refer to this strategy of parameter duplication and, for example, we work with a $2N$ parameter vector that is defined by $\theta_C \equiv (\theta_{1C}, \theta_{2C})^T$. Similarly, we can define the duplicated prior parameter vector by $\theta_{\Pi C} = (\theta_\Pi, \theta_\Pi)^T$. One subtlety is that our null hypothesis is still that the data

is drawn from the prior distribution of a single parameter. We shall see that this difference between the split analysis and statistical properties of the data causes minor issues when counting these parameters if they are partially, but not fully constrained, by the prior.

In the GLM the block structure of the covariance in Eq. (K.11) then projects on the two parameter copies with the Jacobian given by:

$$M_C = \begin{pmatrix} \partial m_1/\partial\theta_1 & \mathbb{O} \\ \mathbb{O} & \partial m_2/\partial\theta_2 \end{pmatrix} = \begin{pmatrix} M_1 & \mathbb{O} \\ \mathbb{O} & M_2 \end{pmatrix} . \tag{K.14}$$

The maximum likelihood estimate of the copy parameters is given by Eq. (K.6) and explicitly reads:

$$\theta_C^{\mathrm{ML}} = \begin{pmatrix} \theta_{1C}^{\mathrm{ML}} \\ \theta_{2C}^{\mathrm{ML}} \end{pmatrix} = \mathcal{C}_C M_C^T \Sigma_J^{-1} \left(x_J - m_{\Pi C} + M_C \theta_{\Pi C} \right) , \tag{K.15}$$

where we have used the definition of the parameter copies covariance $\mathcal{C}_C^{-1} = M_C^T \Sigma_J^{-1} M_C$ and we have defined the duplicate prior center model prediction $m_{\Pi C} = (m_\Pi, m_\Pi)^T$. The parameter copy ML is then Gaussian distributed, over the space of data draws, with $\theta_C^{\mathrm{ML}} \sim \mathcal{N}(\theta_C; \theta_{\Pi C}, \mathcal{C}_C)$. The maximum posterior parameters are obtained by adding copies of the Gaussian priors on top of the ML result. We write explicitly the block structure of the parameter copies posterior covariance as:

$$\mathcal{C}_{pC} \equiv \begin{pmatrix} \mathcal{C}_{p1C} & \mathcal{C}_{p12C} \\ \mathcal{C}_{p21C} & \mathcal{C}_{p2C} \end{pmatrix} , \tag{K.16}$$

that allows us to write the posterior estimate for the first parameter copy as:

$$\theta_{1C}^p = \theta_{1C}^{\mathrm{ML}} - \mathcal{C}_{p1C}\mathcal{C}_\Pi^{-1}(\theta_{1C}^{\mathrm{ML}} - \theta_\Pi) - \mathcal{C}_{p12C}\mathcal{C}_\Pi^{-1}(\theta_{2C}^{\mathrm{ML}} - \theta_\Pi) , \tag{K.17}$$

and similarly for the second parameter copy.

As we can see the GLM posterior distribution for one of the parameter copies is related to the parameters of the other copy in two ways: first indirectly in the ML estimate and then directly at the MAP level. In particular, at the ML level, the parameters of one copy are related to the parameters of the other because of the shared data in Eq. (K.15). This is a natural consequence of the parameter duplication technique: since we always fit the joint data set, split parameters are influenced by data in both splits through their correlation.

In the posterior, there is a direct coupling between the ML and posterior estimators of the copies. In Eq. (K.17) this coupling is mediated by $\mathcal{C}_{p12C}$ in the last term. In the limit where the data is uncorrelated this term would vanish and, therefore, the sets would not be able to communicate with each other; we would, therefore, retrieve the expressions in Raveri & Hu (2019), which would also be the same as the corresponding expressions under the data split methodology. With correlated copy parameters, the maximization of the *joint* posterior depends on the ML values of each. Contrast this with the case of the separate parameters of the data split. Even though the ML values are still correlated according to Eq. (K.13), maximization of the posterior for each parameter is performed without reference or knowledge of its complement.

The joint parameter results can be viewed as a subspace of the parameter copies where all the copies have the same value. We define the projection matrix on this subspace as:

$$D_C^T = (\mathbb{I}_{N\times N}, \mathbb{I}_{N\times N}) \, . \tag{K.18}$$

When expressed as D_C, it is known as the design matrix, which takes a single set of N parameters and produces two separate parameters, i.e. the $2N$ copy parameters, to be estimated by the data. Notice that the relation between the joint and copy Jacobian is given by $M_J^T = D_C^T M_C^T$. The joint parameter covariance is a linear combination of the copy parameter covariance given by $\mathcal{C}_J^{-1} = D_C^T \mathcal{C}_C^{-1} D_C$. The linear combination of the copy parameter

estimates that forms the joint parameter estimate is:

$$\theta_J^{\rm ML} = \mathcal{C}_J D_C^T \mathcal{C}_C^{-1} \theta_C^{\rm ML} \,. \tag{K.19}$$

Likewise $D_C \mathcal{C}_J D_C^T \mathcal{C}_C^{-1}$ is the parameter projector that projects the copy parameters onto the space where they are the same $(\theta_{1C}^{\rm ML}, \theta_{2C}^{\rm ML})^T \to (\theta_J^{\rm ML}, \theta_J^{\rm ML})^T$.

K.4 Data Split CDEs

In this section we discuss CDEs in case of data splits. Specifically, in Sec. K.4.1 we present parameter shift statistics and in Sec. K.4.2 we discuss Goodness of Fit loss.

K.4.1 Parameter shifts

Given two data sets we can compute the difference between the parameters obtained by considering the two data sets alone: $\Delta\theta_S \equiv \theta_{1S}^p - \theta_{2S}^p$. Within the GLM this is Gaussian distributed and it can be shown that its expectation value over data realizations is zero. To form the optimal quadratic form to detect shifts in parameters,

$$Q_{\rm DM}^S \equiv \Delta\theta_S^T [\mathcal{C}(\Delta\theta_S)]^{-1} \Delta\theta_S, \tag{K.20}$$

we need to compute the parameter difference covariance $\mathcal{C}(\Delta\theta_S)$. For a discussion of optimal quadratic forms see App. D in Raveri & Hu (2019). Within the GLM the shift covariance can be obtained starting from the covariance in data space and results in:

$$\begin{aligned}\mathcal{C}(\Delta\theta_S) =& \mathcal{C}_{p1S} + \mathcal{C}_{p2S} - \mathcal{C}_{p1S}\mathcal{C}_\Pi^{-1}\mathcal{C}_{p2S} - \mathcal{C}_{p2S}\mathcal{C}_\Pi^{-1}\mathcal{C}_{p1S} \\ & - \mathcal{C}_{p1S} M_1^T \Sigma_1^{-1} \Sigma_{12} \Sigma_2^{-1} M_2 \mathcal{C}_{p2S} - \mathcal{C}_{p2S} M_2^T \Sigma_2^{-1} \Sigma_{21} \Sigma_1^{-1} M_1 \mathcal{C}_{p1S} \,.\end{aligned} \tag{K.21}$$

As we can see this expression agrees with Raveri & Hu (2019) in the limit of uncorrelated data sets. It cannot be, however, expressed in terms of parameter space quantities only when data correlations are present. In this case the parameter shift covariance depends on both the parameter and data covariance that are connected through the model Jacobian to account for the fact that data correlations are omitted from the single parameter estimates.

In addition to this, we can also write parameter shifts in update form, by comparing the parameters of one posterior (for simplicity 1 here) to the joint parameter determination: $\Delta\theta^U_S \equiv \theta^p_{1S} - \theta^p_J$. This is, again, Gaussian distributed with zero mean and covariance:

$$\begin{aligned}\mathcal{C}(\Delta\theta^U_S) =& \mathcal{C}_{p1S} + \mathcal{C}_{pJ} - \mathcal{C}_{p1S}\mathcal{C}_\Pi^{-1}\mathcal{C}_{pJ} - \mathcal{C}_{pJ}\mathcal{C}_\Pi^{-1}\mathcal{C}_{p1S} \\ &- \mathcal{C}_{p1S}M_1^T\Sigma_1^{-1}(\Sigma_1,\Sigma_{21})^T\Sigma_J^{-1}M_J\mathcal{C}_{pJ} \\ &- \mathcal{C}_{pJ}M_J^T\Sigma_J^{-1}(\Sigma_1,\Sigma_{12})\Sigma_1^{-1}M_1\mathcal{C}_{p1S}\,. \end{aligned} \tag{K.22}$$

This agrees with Raveri & Hu (2019) in the limit of uncorrelated data sets, where we recover $\mathcal{C}(\Delta\theta^U_S) = \mathcal{C}_{p1S} - \mathcal{C}_{pJ}$, but becomes significantly more complicated in general due to the presence of data correlations. We denote with

$$Q^S_{\rm UDM} \equiv (\Delta\theta^U_S)^T[\mathcal{C}(\Delta\theta^U_S)]^{-1}\Delta\theta^U_S\,, \tag{K.23}$$

the optimal data split parameter shift statistics in update form. Under the GLM, both $Q^S_{\rm DM}$ and $Q^S_{\rm UDM}$ are chi-squared distributed with number of degrees of freedom equivalent to the rank of their respective covariance matrix.

In case of uncorrelated data sets the statistical significance of parameter shifts in update form is the same as the statistical significance of the difference between $\theta^p_1 - \theta^{\rm ML}_2$ since these two quantities are related by a linear transformation. However, in the presence of data correlations this is not the case since the single parameters do not contain the information on the data correlation that is contained in the joint parameter determination. In other

words, it is not possible to write the update parameter shift as a linear combination of the shift in the two single parameters. Hence, we would expect to see some differences between the two estimates, related to the presence of correlated data and parameter modes.

From the previous discussion it appears clear that using the optimal, inverse covariance weighted, CDEs for data split parameter shifts is challenging in presence of data correlations. Their covariances cannot be written in parameter space and depend on both the posterior and data covariance. These can be related to each other by projection operations involving derivatives of the observables that are cumbersome to compute accurately. These considerations limit the applicability of these methods in practice.

The SN data Jacobian, M_J, is estimated numerically by linear finite differences computed around the best fit of the joint SN data set. The finite difference parameter step is computed such that it would correspond to a SN chi-square difference of one, ensuring that the derivatives are estimated on the scale at which they are relevant and are not contaminated by numerical noise. We assume that the model is fully linear so that the joint Jacobian determines the single data split Jacobian.

All the other quantities that are needed to compute $Q^S_{\rm DM}$ and $Q^S_{\rm UDM}$ are estimated from the GLM. A numerically challenging aspect of computing $Q^S_{\rm DM}$ and $Q^S_{\rm UDM}$ is identifying directions that can contribute to parameter shifts and those that do not. The latter parameter combinations can be either prior constrained or fully correlated, as can be seen from Eq. (K.21) and Eq. (K.22). In practice, due to numerical noise, the parameter shift covariances are never exactly zero along these directions.

In the uncorrelated case this problem is solved, at least for parameter shifts in update form, by computing the quadratic form using the Karhunen-Loeve (KL) decomposition of the covariances involved, as discussed in Raveri & Hu (2019). In this case we select the directions that are used to compute the two parameter shift estimators based on the KL decomposition of the shift covariance matrices and the parameter covariance of the most

constraining of the two data sets. Once the KL decomposition is performed the spectrum of the KL eigenvalues can be examined to understand if there is a clear separation of modes with KL eigenvalues very close to zero and directions that are significantly different from zero. This strategy also avoids problems with parameters having different units since the KL modes are invariant under changes of parameter basis. This also results in a wide separation between directions that can and cannot contribute to a shift making it easier to identify and remove the latter. Once the directions that cannot contribute any shift are isolated and removed the parameter shifts and their covariance are both projected on the other directions and $Q^S_{\rm DM}$ and $Q^S_{\rm UDM}$ are computed. The number of degrees of freedom of the two tests is given by the number of KL modes that are retained. In the SN example, this number is two, since the absolute magnitude constraint does not differ between the sets.

K.4.2 Goodness of fit loss

In addition to shifts in parameters we can use, as a CDE, the statistics of the ratio of the joint and single likelihoods at maximum posterior, $Q_{\rm DMAP}$ Raveri & Hu (2019). In the case where we consider data split we refer to this estimator as:

$$Q^S_{\rm DMAP} \equiv 2\ln\mathcal{L}_1(\theta^S_{p1}) + 2\ln\mathcal{L}_2(\theta^S_{p2}) - 2\ln\mathcal{L}_J(\theta^J_p)\ , \tag{K.24}$$

This quantifies goodness of fit loss as it corresponds to the degradation of the performances of the model when fitting two data sets jointly vs fitting the joint data. When two data sets are considered separately the model can invest all its parameters in improving the fit to data. On the other hand, when the two data sets are joined, the parameters have to compromise between the two and the joint fit will be worse. However note that in the correlated case the two data sets are not independent so that the joint likelihood is not the product of the two independent likelihoods regardless of the parameters. Consequently $Q^S_{\rm DMAP}$ is not

necessarily positive definite, complicating its interpretation as a goodness of fit loss.

Computing the statistics of $Q^S_{\rm DMAP}$ for correlated data sets, within the GLM, proves extremely hard in case of data set splits, as we see below in more detail.

We first consider the ratio of maximum likelihoods of the joint data set and the two subsets. By direct calculation it can be shown that, up to constant offsets that is irrelevant to the calculation of statistical significance this ratio can be written as the quadratic form in the data:

$$\begin{aligned} Q^S_{\rm DML} &= X_J^T \Bigg[(\mathbb{I}_J - \mathbb{P}_J)^T \Sigma_J^{-1} (\mathbb{I}_J - \mathbb{P}_J) \\ &\quad - \begin{pmatrix} (\mathbb{I}_1 - \mathbb{P}_1)^T \Sigma_1^{-1} (\mathbb{I}_1 - \mathbb{P}_1) & \mathbb{O} \\ \mathbb{O} & (\mathbb{I}_2 - \mathbb{P}_2)^T \Sigma_2^{-1} (\mathbb{I}_2 - \mathbb{P}_2) \end{pmatrix} \Bigg] X_J \\ &\equiv X_J^T A^S_{\rm DML} X_J \ , \end{aligned} \tag{K.25}$$

where X_J is the full data vector, distributed according to the evidence of the joint data set, and $A^S_{\rm DML}$ is the matrix that defines $Q^S_{\rm DML}$. The indices 1 and 2 denote the two subsets of the data after we split the joint set.

Through explicit computation, the form of the joint projector is given by:

$$\mathbb{P}_J \equiv \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

with

$$
\begin{aligned}
p_{11} &= M_1 \mathcal{C}_J (M_1^T - M_2^T \Sigma_2^{-1} \Sigma_{21}) K_1^{-1} \,, \\
p_{12} &= M_1 \mathcal{C}_J (M_2^T - M_1^T \Sigma_1^{-1} \Sigma_{12}) K_2^{-1} \,, \\
p_{21} &= M_2 \mathcal{C}_J (M_1^T - M_2^T \Sigma_2^{-1} \Sigma_{21}) K_1^{-1} \,, \\
p_{22} &= M_2 \mathcal{C}_J (M_2^T - M_1^T \Sigma_1^{-1} \Sigma_{12}) K_2^{-1} \,,
\end{aligned}
$$

where we have defined:

$$
\begin{aligned}
K_1 &\equiv \Sigma_1 - \Sigma_{12} \Sigma_2^{-1} \Sigma_{21} \,, \\
K_2 &\equiv \Sigma_2 - \Sigma_{21} \Sigma_1^{-1} \Sigma_{12} \,.
\end{aligned}
$$

In order to calculate the distribution of Eq. (K.25) we follow the procedure discussed in App. A of Raveri & Hu (2019) and compute the eigenvalues, λ, of $A^S_{\rm DML} \mathcal{S}_J$, where $\mathcal{S}_J$ is the covariance for the joint distribution of the data which, for Gaussian priors, is $\mathcal{S}_J = \Sigma_J + M_J \mathcal{C}_\Pi M_J^T$. This allows to decompose $Q^S_{\rm DML}$ in the following way:

$$
Q^S_{\rm DML} = \sum_i \lambda_i \mathcal{U}_i^2 \,, \tag{K.26}
$$

where each $\mathcal{U}_i \sim \mathcal{N}_J(x_J; \mathbb{O}, \mathbb{I})$ so that $Q^S_{\rm DML}$ is a weighted sum of chi squared variables.

By direct calculation we have:

$$
\mathcal{A}^S_{\rm DML} \equiv A^S_{\rm DML} \mathcal{S}_J \equiv \begin{pmatrix} A & B \\ C & D \end{pmatrix} \,, \tag{K.27}
$$

where:

$$
\begin{aligned}
A &= \left[\Sigma_1^{-1} M_1 \mathcal{C}_1 - K_1^{-1}(M_1 - \Sigma_{12}\Sigma_2^{-1} M_2)\mathcal{C}_J\right] M_1^T \ , \\
B &= -K_1^{-1}(M_1 - \Sigma_{12}\Sigma_2^{-1} M_2)\mathcal{C}_J M_2^T - (\mathbb{I}_1 - \Sigma_1^{-1} M_1 \mathcal{C}_1 M_1^T)\Sigma_1^{-1}\Sigma_{12} \ , \\
C &= -K_2^{-1}(M_2 - \Sigma_{21}\Sigma_1^{-1} M_1)\mathcal{C}_J M_1^T - (\mathbb{I}_2 - \Sigma_2^{-1} M_2 \mathcal{C}_2 M_2^T)\Sigma_2^{-1}\Sigma_{21} \ , \\
D &= \left[\Sigma_2^{-1} M_2 \mathcal{C}_2 - K_2^{-1}(M_2 - \Sigma_{21}\Sigma_1^{-1} M_1)\mathcal{C}_J\right] M_2^T \ .
\end{aligned}
$$

An analytic solution to the above eigenvalue problem is not easily obtained, but can be obtained numerically to evaluate the exact distribution of Q^S_{DML}. We highlight that, similarly to what happens for data split parameter shifts, the calculation of the statistics involves quantities that are defined both at the parameter space and data space level.

Note that the expressions we derived above reduce to the corresponding ones in Raveri & Hu (2019) in the limit of uncorrelated data sets.

Furthermore we can notice that the quadratic form defined by Q^S_{DML} is not necessarily positive definite. This is a consequence of the fact that the projector on the joint parameter space is not a sub-space of the span of the single data set projector. This severely limits the possibility of approximating Q^S_{DML} with a chi squared distribution, which is positive definite, especially for events in the confirmation tail that would be very close to $Q^S_{\mathrm{DML}} = 0$.

In addition, the fact that Q^S_{DML} is not chi squared distributed means that correlated data fluctuations are not optimally weighted.

We then consider the ratio of likelihoods at maximum posterior (DMAP) in the data split case. To do so we add the extra terms that transform ML into MAP so that the matrix

that controls $Q^S_{\rm DMAP}$ is given by:

$$\mathcal{A}^S_{\rm DMAP} = \mathcal{A}^S_{\rm DML} + \Sigma_J^{-1} M_J \mathcal{C}_J \mathcal{C}_\Pi^{-1} \mathcal{C}_{pJ} M_J^T - \begin{pmatrix} \tilde{M}_1^T \mathcal{C}_\Pi^{-1} \mathcal{C}_{p1} M_1^T & \tilde{B} \\ \tilde{B}(1 \leftrightarrow 2) & \tilde{M}_2^T \mathcal{C}_\Pi^{-1} \mathcal{C}_{p2} M_2^T \end{pmatrix}, \tag{K.28}$$

where, for compactness, we have defined $\tilde{B} \equiv \tilde{M}_1^T \mathcal{C}_\Pi^{-1} \mathcal{C}_{p1} \mathcal{C}_1^{-1} \mathcal{C}_{p1} (\mathcal{C}_\Pi^{-1} \tilde{M}_1 \Sigma_{12} + M_2^T)$.

These results, for both $Q^S_{\rm DML}$ and $Q^S_{\rm DMAP}$, can be used to compute the respective exact distributions. The trace of these distributions coincides with the results obtained in the uncorrelated case, but we notice that it is problematic to approximate them with simpler distributions because both of them are not positive definite.

The fact that both $Q^S_{\rm DML}$ and $Q^S_{\rm DMAP}$ are not positive definite means that there are aspects of the data where the joint likelihood is better than the product of the separate likelihoods. This can never happen for uncorrelated data sets and is a consequence of the presence of correlated data modes. In particular, the data modes that are fit by the model separately are the ones that would contribute to the positive definiteness of the above statistics, since they can zero out different data fluctuations, whereas the correlated data modes that are left out can contribute negatively. Fitting the data jointly, however, always takes correlated modes into account, so that the contribution to the chi-square from them is considered.

K.5 Parameter Split CDEs

In this section we follow the strategy of quantifying agreement and disagreement by duplicating model parameters. We first go through the analytic aspects of calculating the CDEs and then report the results of applying them to the SN example that we consider in this work. In Sec. K.5.1 we present parameter shift statistics, in Sec. K.5.2 we discuss exact

Monte Carlo parameter shift statistics, while in Sec. K.5.3 we consider goodness of fit loss.

K.5.1 Parameter shifts

We first consider the difference between the duplicate parameter posteriors, denoted by $\Delta\theta_C \equiv \theta^p_{1C} - \theta^p_{2C}$, in the case of parameter splits. To form the optimal estimator of the significance of the shifts, we construct the quadratic form:

$$Q^C_{\rm DM} \equiv (\Delta\theta_C)^T [\mathcal{C}(\Delta\theta_C)]^{-1} \Delta\theta_C \tag{K.29}$$

using their covariance to weight shifts in different parameter space directions. In this case the covariance reads:

$$\mathcal{C}(\Delta\theta_C) = \mathcal{C}_{p1C} + \mathcal{C}_{p2C} - \mathcal{C}_{p12C} - \mathcal{C}_{p21C} \,. \tag{K.30}$$

Notice that, in the uncorrelated limit $\mathcal{C}_{pijC} = \mathcal{C}_{piC}\mathcal{C}^{-1}_{\Pi}\mathcal{C}_{pjC}$ for $i, j \in [1, 2]$. Furthermore, Eq. (K.30) implies that in the case of parameter duplication we can express the covariance of the parameter shifts using just the results from the MCMC chains. This is not true in the case of data splits, however, where the expression of the covariance includes terms related to the data covariance which account for the correlations.

We then calculate the covariance of parameter shifts in update form using one of the two parameter copies, namely θ^p_{1C}, and the parameters inferred from the joint data set, θ^p_J. Therefore, defining $\Delta\theta^U_C = \theta^p_{1C} - \theta^p_J$, the covariance of parameter shift in update form is written as

$$\mathcal{C}(\Delta\theta^U_C) = \mathcal{C}_{p1C} - \mathcal{C}_{pJ} \,, \tag{K.31}$$

which is invariant in form to the one without correlations, which is discussed in Raveri &

Hu (2019).

We denote with

$$Q^C_{\rm UDM} \equiv (\Delta\theta^U_C)^T[\mathcal{C}(\Delta\theta^U_C)]^{-1}\Delta\theta^U_C \tag{K.32}$$

the optimal parameter-split parameter shift statistic in update form. Notice that, since Eq. (K.31) is invariant in form with respect to the uncorrelated case considered in Raveri & Hu (2019), we can compute $Q^C_{\rm UDM}$ by means of the KL decomposition to filter out modes that are not improved by the data over the prior and hence subject to sampling noise. Under the GLM, both $Q^C_{\rm DM}$ and $Q^C_{\rm UDM}$ are chi-squared distributed with number of degrees of freedom equal to the rank of their covariances.

The statistical significance of the two $Q^C_{\rm DM}$ and $Q^C_{\rm UDM}$ estimators is the same for the maximum likelihood parameters while it might differ at the maximum posterior level in case of partially informative priors. This difference stems from the fact that the update form of parameter shifts contains only one copy of the prior in the joint, whereas in the single parameter shift the prior is applied once to each data set. Therefore, θ^p_J cannot be formed from a linear combination of θ^p_{iC}. We can instead define a joint parameter estimate that is so constructed

$$\tilde{\theta}^p_J = \tilde{C}_{pJ}D^T_C\mathcal{C}^{-1}_{pC}(\theta^p_{1C},\theta^p_{2C})^T \tag{K.33}$$

with covariance $\tilde{\mathcal{C}}^{-1}_{pJ} = \mathcal{C}^{-1}_{pJ} + \mathcal{C}^{-1}_{\Pi} = D^T_C\mathcal{C}^{-1}_{pC}D_C$, so that

$$\theta^p_{1C} - \tilde{\theta}^p_J = \tilde{\mathcal{C}}_{pJ}D^T_C\mathcal{C}^{-1}_{pC}(\mathbb{O},\theta^p_{1C} - \theta^p_{2C})^T \ , \tag{K.34}$$

where the vector $\mathbb{O}$ has length N_p. This clearly shows that the statistical significance of $\theta^p_{1C} - \tilde{\theta}^p_J$ is the same as $\theta^p_{1C} - \theta^p_{2C}$ since the two are related by a linear, invertible transformation. We can then write the update parameter difference as:

$$\Delta\theta^U_C = (\theta^p_{1C} - \tilde{\theta}^p_J) + (\tilde{\theta}^p_J - \theta^p_J) \ , \tag{K.35}$$

which, in the uncorrelated limit reduces to $\theta^p_{1C} - \theta^{\rm ML}_2$. This agrees with the discussion in Raveri & Hu (2019) of their Eq. (47). More generally, the additional difference can be computed from C_{pJ} and C_Π and can cause $\Delta\theta^U_C$ to be larger than the difference implied by $\theta^p_{1C} - \theta^p_{2C}$ since the Gaussian priors in each copy tend to bring the posteriors closer together. Note that for flat, range bound, priors as in our SN example the two copies do not lead to a stronger joint prior so that $\tilde{\theta}^p_J = \theta^p_J$.

Furthermore the difference between $Q^C_{\rm DM}$ and $Q^C_{\rm UDM}$ becomes relevant only if there is a non-negligible shift along partially prior constrained directions since the two estimators agree in the fully data and prior constrained limits.

It is clear at this point that making use of the parameter split methodology provides some advantages compared to the data splitting method. Equations (K.30) and (K.31) for the covariances for the parameter split statistics should be compared with Eqs. (K.21) and (K.22) for data split statistics. Crucially the former can be simply calculated from parameter covariances whereas the latter require manipulations of the data covariance. We can also therefore check the GLM results using parameter covariances taken from the MCMC chain when evaluating Eqs. (K.29) and (K.32).

K.5.2 Monte Carlo exact parameter shifts

Having an MCMC parameter estimation in the case of parameter duplication presents us with the additional possibility of computing parameter shifts as a Monte Carlo integral, as we discuss in this section.

We first consider the joint posterior probability density function of the two parameter copies $P(\theta^p_{1C}, \theta^p_{2C})$. We can then calculate the distribution of their difference, denoted by $\Delta\theta_C \equiv \theta^p_{1C} - \theta^p_{2C}$, as the N-dimensional convolution integral:

$$P(\Delta\theta_C) = \int_{V_p} P(\theta^p_{1C}, \theta^p_{1C} - \Delta\theta_C)\, d\theta^p_{1C}\,, \tag{K.36}$$

over the whole parameter space volume V_p. Note that this equation is general and describes the probability to observe a parameter shift $\Delta\theta_C$ without assuming the parameters to be independent. In the limit of uncorrelated data sets the joint probability distribution in the above expression reduces to $P(\theta^p_{1C}, \theta^p_{2C}) = P_1(\theta^p_{1C})P_2(\theta^p_{2C})$.

To compute the statistical significance of a shift in parameters we then evaluate the integral:

$$S = \int_{P(\Delta\theta_C)>P(0)} P(\Delta\theta_C)\, d\Delta\theta_C \ , \tag{K.37}$$

where the volume of integration is defined as the region of parameter space where the probability to get a shift $\Delta\theta_C$ is above the isocontour of no shift, $\Delta\theta_C = 0$.

To form the MCMC chain of parameter differences in the case of correlated data sets we can take, sample by sample, the difference between the first and second copy of the parameters, without changing the weights of the samples. The result would be the MCMC estimate of the convolution integral in Eq. (K.36). Since the parameter duplication chain is run to convergence in the full $2N$ dimensional space the parameter difference chain is appropriately sampled.

Once we have the samples from the parameter difference probability we can compute the integral in Eq. (K.37) with a mixture of kernel density estimation (KDE) and MCMC techniques.

The probability of a difference in parameter, for every sample in the difference chain, is estimated through KDE with a Gaussian smoothing kernel that uses the scaled parameter difference covariance. The smoothing kernel is explicitly given by:

$$K(\Delta\theta_{1C}, \Delta\theta_{2C}) = \mathcal{N}(\Delta\theta_{1C}; \Delta\theta_{2C}, \Lambda\, \mathcal{C}(\Delta\theta_C)\, \Lambda) \ , \tag{K.38}$$

where $\Lambda = \mathrm{diag}(\sqrt{\lambda})$ is a scaling matrix with λ as the smoothing scaling parameter.

We fix this parameter using Silverman's rule of thumb Wand & Jones (1994) to:

$$\sqrt{\lambda} = \left(\frac{4}{n_{\rm s}(N+2)} \right)^{\frac{1}{N+4}} , \tag{K.39}$$

where $n_{\rm s}$ is the number of samples in the chain and N is the number of parameters.

For a given MCMC sample j the KDE probability of a shift is given by:

$$P(\Delta\theta_{jC}) = \frac{1}{\sum_{i=1}^{n_{\rm s}} w_i} \sum_{i=1}^{n_{\rm s}} w_i \, K(\Delta\theta_{jC}, \Delta\theta_{iC}) , \tag{K.40}$$

where w_i denotes the weights of the samples and given that the smoothing kernel is normalized. Eq. (K.40) is also computed for the zero shift so that the MCMC estimate of the integral in Eq. (K.37) is given by the number of samples that have a KDE probability of shift above the KDE probability of zero over the total number of samples.

This approach has several advantages. First, the combination of MCMC and KDE makes the estimate weakly sensitive to the choice of the smoothing kernel. The amount of over/under smoothing that the kernel might be doing is balanced by the fact that that would also happen for the zero shift estimate and would drop in the difference. In other words we never just use directly the probability of a zero shift, as obtained from Eq. (K.40), that would largely depend on the smoothing kernel in general, but rather compute how many samples from the distribution are above that probability. The second advantage is that this parameter shift estimate is now completely accounting for all possible non-Gaussianities in the parameter posterior.

The challenge in using this estimate is that, for statistically significant results, the estimate is likely to be noisy due to the fact that the MCMC chain would have very few samples in the tail of the distribution.

This sampling error can, however, be estimated in two ways. The first is given by a shot noise estimate, by taking the square root of the number of MCMC samples in the

smallest probability tail to account for both tensions and confirmation results. The second is estimated as the variance of the result across multiple MCMC chains of the same distribution. In this case we have $n_{\rm chains}$ chains and we compute the shift probability for each of them and then estimate the error as the ratio of the variance across chains, weighted by the number of chains, $\sigma_S^2 = {\rm var}(S)_{\rm chains}/n_{\rm chains}$, since the fiducial result uses all of them and given that different chains are independent. The two error estimates are usually in good agreement for well converged chains.

K.5.3 Goodness of fit loss

The last CDE that we discuss is goodness-of-fit loss with the parameter split approach. In contrast to the data split case, when considering parameter splits $Q^C_{\rm DMAP}$ becomes easy to compute, as we discuss below.

At the posterior level the goodness-of-fit loss statistics is defined as:

$$Q^C_{\rm DMAP} \equiv 2\ln\mathcal{L}_J(\theta^p_C) - 2\ln\mathcal{L}_J(\theta^p_J)\,. \tag{K.41}$$

Shortly below we discuss in detail its exact distribution as a linear combination of chi-squared variables. In practice the distribution of $Q^C_{\rm DMAP}$ can be approximated by that of a single chi-squared distribution, matching the mean of the exact distribution, with degrees of freedom:

$$\langle Q^C_{\rm DMAP}\rangle = N^C_{\rm eff} - N^J_{\rm eff} - {\rm tr}[\mathcal{C}^{-1}_{\Pi C}\mathcal{C}_{pC}(\mathbb{I}_{2N} - \mathcal{C}^{-1}_{\Pi C}\mathcal{C}_{pC})\mathbb{J}_{2N}]\,, \tag{K.42}$$

where we have defined $\mathbb{J}_{2N} \equiv D_C D_C^T - \mathbb{I}_{2N}$.

As we can see, the statistics of $Q^C_{\rm DMAP}$ can be easily computed from the posterior MCMC samples. In the uncorrelated case it also reduces to the $Q_{\rm DMAP}$ statistics discussed in Raveri & Hu (2019). Compared to the uncorrelated case, we notice that in the correlated case there

is an extra term that is present in the mean of the exact $Q^C_{\rm DMAP}$ distribution, as shown in Eq. (K.42), in addition to the difference in the number of effective parameters. Notice that this term vanishes for fully data or prior constrained directions. Its appearance is associated with the mismatch of assuming the data is drawn from a single parameter and prior but analyzed with split parameters and independent priors.

At maximum likelihood level the statistics of goodness-of-fit loss is chi-squared distributed as a consequence of the fact that the parameter copies decompose the joint parameter estimate, as we show below.

We first consider the statistics of ML ratios (DML). We focus on the distribution of the DML statistic between the joint chain and the one with the duplicated parameter space. To do so, we begin by considering the ML parameter split determination, $\theta^{\rm ML}_C = (\theta^{\rm ML}_{1C}, \theta^{\rm ML}_{2C})^T$, and the joint ML parameters, $\theta^{\rm ML}_J$. We then use them to define the difference in joint log-likelihood at the ML point as:

$$Q^C_{\rm DML} \equiv -2\ln\mathcal{L}_J(\theta^{\rm ML}_J) + 2\ln\mathcal{L}_J(\theta^{\rm ML}_C)\ . \tag{K.43}$$

Note that in the limit of uncorrelated data this reduces to $Q^C_{\rm DML} = -2\ln\mathcal{L}_J(\theta^{\rm ML}_J) + 2\ln\mathcal{L}_1(\theta^{\rm ML}_1) + 2\ln\mathcal{L}_2(\theta^{\rm ML}_2)$ which is similar to the expressions used in Raveri & Hu (2019).

In the GLM, it can be shown that, up to constants which are not important for our purpose, we get the following quadratic form in data space:

$$\begin{aligned} Q^C_{\rm DML} &= X^T\left[(\mathbb{I}-\mathbb{P}_J)^T\,\Sigma_J^{-1}\,(\mathbb{I}-\mathbb{P}_J) - (\mathbb{I}-\mathbb{P}_C)^T\,\Sigma_J^{-1}\,(\mathbb{I}-\mathbb{P}_C)\right]X \\ &\equiv X^T A^C_{\rm DML} X\ , \end{aligned} \tag{K.44}$$

where we have used the joint projector $\mathbb{P}_J = M_J\mathcal{C}_J M_J^T\Sigma_J^{-1}$ and the projector under parameter duplication written as $\mathbb{P}_C = M_C\mathcal{C}_C M_C^T\Sigma_J^{-1}$. Now, we can rewrite the joint projector in

the following way:

$$\mathbb{P}_J = M_C D_C (D_C^T \mathcal{C}_C^{-1} D_C)^{-1} D_C^T M_C^T \Sigma_J^{-1} \ , \tag{K.45}$$

while the projector in the case of parameter duplication can be expressed as:

$$\mathbb{P}_C = M_C (M_C^T \Sigma_J^{-1} M_C)^{-1} M_C^T \Sigma_J^{-1} \ . \tag{K.46}$$

Then, using the above expressions it is straightforward to show that the joint set of parameters is a subset of the duplicate set, since $\mathbb{P}_J \mathbb{P}_C = \mathbb{P}_C \mathbb{P}_J = \mathbb{P}_J$. Therefore, we can use theorem (5.2.5) in A.M. Mathai (1992) to show that, at the ML level,

$$\begin{aligned} Q_{\rm DML}^C &\sim \chi^2(\text{rank}(\mathbb{I} - \mathbb{P}_J) - \text{rank}(\mathbb{I} - \mathbb{P}_C)) \\ &= \chi^2(N_C - N_J) \ , \end{aligned} \tag{K.47}$$

where N_C and N_J are the number of parameter duplicates and the number of joint parameters respectively. Note that, in the limit of uncorrelated data sets $N_C = N_1 + N_2$, where N_1 and N_2 are the number of relevant parameters for the first and second data sets respectively.

In contrast with the case of data split the exact statistics of the parameter split DML estimator is a chi square, which also means that $Q_{\rm DML}^C$ is optimal.

The exact statistics of $Q_{\rm DML}^C$ can also be obtained by explicitly computing the eigenvalues of $A_{\rm DML}^C \mathcal{S}_J = \mathbb{P}_C^T - \mathbb{P}_J^T$, where $\mathcal{S}_J = \Sigma_J + M_J \mathcal{C}_\Pi M_J^T$. Notice that $M_J \mathcal{C}_\Pi M_J^T$ represents a prior that is fully correlated between the split parameters, whereas our parameter split analysis assumes separate priors that are uncorrelated. This is necessary since otherwise the split parameters would be expected to vary according to $\mathcal{C}_\Pi$ leading to a counterfactually large expected improvement from fitting them separately. Conversely, the split parameter technique cannot employ fully correlated priors because no matter how weak such a prior is,

it would force the split parameter posterior means to the same values.

We can now turn to the distribution of $Q_{\rm DMAP}$, with parameter copies. This can be written as:

$$\begin{aligned} Q^C_{\rm DMAP} &\equiv -2\ln\mathcal{L}_J(\theta^p_J) + 2\ln\mathcal{L}_J(\theta^p_C) \\ &= Q^C_{\rm DML} + X^T_J\left[\tilde{M}^T_J\mathcal{C}^{-1}_\Pi\mathcal{C}_{pJ}\mathcal{C}^{-1}_J\mathcal{C}_{pJ}\mathcal{C}^{-1}_\Pi\tilde{M}_J\right. \\ &\quad \left. -\tilde{M}^T_C\mathcal{C}^{-1}_{\Pi C}\mathcal{C}_{pC}\mathcal{C}^{-1}_C\mathcal{C}_{pC}\mathcal{C}^{-1}_{\Pi C}\tilde{M}_C\right]X_J \\ &\equiv X^T_J A^C_{\rm DMAP}X_J\ , \end{aligned} \tag{K.48}$$

where we have used that for Gaussian priors the likelihood at the point of maximum posterior is given by $-2\ln\mathcal{L}(\theta_{pJ}) = X^T[(\mathbb{I}-\mathbb{P}_J)^T\Sigma^{-1}_J(\mathbb{I}-\mathbb{P}_J) + \tilde{M}^T_J\mathcal{C}^{-1}_\Pi\mathcal{C}_{pJ}\mathcal{C}^{-1}_J\mathcal{C}_{pJ}\mathcal{C}^{-1}_\Pi\tilde{M}_J]X$ for the joint, and similarly for the parameter copy case. In the above, the copy prior covariance is defined as $\mathcal{C}_{\Pi C} = {\rm diag}(\mathcal{C}_\Pi, \mathcal{C}_\Pi)$.

To calculate the exact distribution of $Q^C_{\rm DMAP}$, we follow the same procedure as in the case of data splits, at the end of in Section K.4. Therefore, we start with the computation of the matrix $A^C_{\rm DMAP}\mathcal{S}_J$ whose spectrum completely specifies the distribution of $Q^C_{\rm DMAP}$ as a sum of independent Gamma distributed variables. It can be shown that this matrix reduces to:

$$\begin{aligned} A^C_{\rm DMAP}\mathcal{S}_J = \mathbb{P}^T_C - \mathbb{P}^T_J + \tilde{M}^T_J\mathcal{C}^{-1}_\Pi\mathcal{C}_{pJ}M^T_J \\ - \tilde{M}^T_C\mathcal{C}^{-1}_{\Pi C}\mathcal{C}_{pC}\mathcal{C}^{-1}_C\mathcal{C}_{pC}\left(\mathcal{C}^{-1}_{\Pi C}\mathcal{C}_C + D_C D^T_C\right)M^T_C\ . \end{aligned} \tag{K.49}$$

It can then be also show that the non-zero eigenvalues of Eq. (K.49) are also the eigenvalues

of the matrix:

$$\begin{aligned}\mathcal{A}^{C}_{\rm DMAP} &= \mathbb{I}_{2N} + D_C\mathcal{C}_{\Pi}^{-1}\mathcal{C}_{pJ}D_C^T - \mathcal{C}_{pC}^{-1}D_C\mathcal{C}_{pJ}D_C^T \\ &\quad -\mathcal{C}_{\Pi C}^{-1}\mathcal{C}_{pC}D_CD_C^T + \mathcal{C}_{\Pi C}^{-1}\mathcal{C}_{pC}\mathcal{C}_{\Pi C}^{-1}\mathcal{C}_{pC}\mathbb{J}_{2N}\ ,\end{aligned} \tag{K.50}$$

where we have defined, for convenience, the exchange matrix $\mathbb{J}_{2N} \equiv D_CD_C^T - \mathbb{I}_{2N}$ that exchanges the off diagonal blocks with the diagonal ones. Note that the above expression is written in terms of quantities that can be obtained from MCMC samples of the posterior of both the parameter copy and joint chains.

Either one can use Eq. (K.50) to compute the exact distribution or one can approximate it by a chi squared distribution matching the mean of the exact distribution as a first order Patnaiks' approximation Patnaik (1950). The mean of the exact distribution and the number of degrees of freedom of the chi squared approximation is given by:

$$\begin{aligned}\mathrm{tr}[\mathcal{A}^{C}_{\rm DMAP}] &= N + \mathrm{tr}[\mathcal{C}_{\Pi}^{-1}\mathcal{C}_{pJ}] - \mathrm{tr}[\mathcal{C}_{\Pi C}^{-1}\mathcal{C}_{pC}] \\ &\quad -\mathrm{tr}[\mathcal{C}_{\Pi C}^{-1}\mathcal{C}_{pC}(\mathbb{I}_{2N} - \mathcal{C}_{\Pi C}^{-1}\mathcal{C}_{pC})\mathbb{J}_{2N}] \\ &= N^{C}_{\rm eff} - N^{J}_{\rm eff} + \mathrm{tr}[\mathcal{C}_{\Pi}^{-1}(\mathcal{C}_{p1C} + \mathcal{C}_{p2C} - \mathcal{C}_{\Pi})\mathcal{C}_{\Pi}^{-1}(\mathcal{C}_{p12C} + \mathcal{C}_{p21C})]\ .\end{aligned} \tag{K.51}$$

We can furthermore calculate the variance of the distribution as it is proportional to the trace of the matrix $(\mathcal{A}^{C}_{\rm DMAP})^2$. This, however, does not significantly simplify and in practical applications it is significantly easier to compute the variance numerically.

All the results above agree with the results in Raveri & Hu (2019) in the uncorrelated limit.

K.6 Arbitrarily split parameters

In this section we generalize the discussion of the parameter split estimators to the case where we consider parameters that are multiply split or not split at all.

Quantities associated with the split and unsplit part of the parameter space will be denoted by the subscripts "C" and "U" respectively. We denote the unsplit posterior parameters with θ^p_U and the n posterior parameter copies with $\theta^p_{SC} = (\theta^p_{1C}, \theta^p_{2C}, \ldots, \theta^p_{nC})^T$. Therefore, the full posterior parameter vector can be written as $\theta^p_C = (\theta^p_{SC}, \theta^p_U)^T$. In what follows, the total number of copy parameters will be nN_C, where N_C is the number of split parameters, and the number of unsplit parameters will be N_U; therefore, $N_p = nN_C + N_U$ is the total number of parameters in the final parameter vector.

Note also that the joint analysis deals with the original $N_J = N_C + N_U$ parameters in total. The joint parameter vector will be denoted as $\theta^p_J = (\theta^p_{CJ}, \theta^p_{UJ})^T$, where the two parts θ_{CJ} and θ_{UJ} correspond to the parameter subspaces that are split and unsplit in the parameter split methodology, respectively.

The design matrix D_C relates the joint quantities with the copy ones. Constructing this appropriately is then enough to generalize our analysis as described in the previous sections. Let D_{SC} be the $nN_C \times N_C$ dimensional design matrix related to the part of parameter space that is being copied n times; thus, $D^T_{SC} = (\mathbb{I}_C, \ldots, \mathbb{I}_C)$ with n instances of the identity matrix $\mathbb{I}_C$ of dimensions $N_C \times N_C$. Let also $\mathbb{I}_U$ be the $N_U \times N_U$ identity matrix related to the N_U unsplit parameters. Then, the design matrix takes the form

$$D_C = \begin{pmatrix} D_{SC} & \mathbb{O} \\ \mathbb{O} & \mathbb{I}_U \end{pmatrix}, \tag{K.52}$$

where $\mathbb{O}$ is the vector with the appropriate number of zeros in each case. Thus, the full design matrix has dimensions of $(nN_C + N_U) \times (N_C + N_U)$.

We first consider parameter shifts of the form $\theta^p_{iC} - \theta^p_{jC}$ between the i-th and j-th copies,

where $i, j = 1, 2, \ldots n$ run over all the n parameter copies. Then, we can express the general form of the covariance between the two parameter differences $\theta^p_{iC} - \theta^p_{jC}$ and $\theta^p_{kC} - \theta^p_{lC}$, with $k, l = 1, 2, \ldots, n$, as

$$\langle (\theta^p_{iC} - \theta^p_{jC})(\theta^p_{kC} - \theta^p_{lC})^T \rangle = \mathcal{C}_{pikC} + \mathcal{C}_{pjlC} - \mathcal{C}_{pilC} - \mathcal{C}_{pjkC} \ . \tag{K.53}$$

These matrices then construct the covariance that is associated with the split copy part of the parameter space. Since there is no shift in the unsplit parameters, the parameter differences and covariances associated with the unsplit part of the parameter space is zero.

We now turn to the discussion of update parameter differences. In this case, we consider differences between the posterior parameters from a joint analysis, namely θ^p_J, and the copy parameter vector $(\theta^p_{iC}, \theta^p_U)^T$ which includes the unsplit copy parameters as well as the i-th copy parameter set. We thus form the parameter differences in update form as

$$\Delta\theta^U_C \equiv (\theta^p_{iC}, \theta^p_U)^T - \theta^p_J = (\theta^p_{iC} - \theta^p_{CJ}, \theta^p_U - \theta^p_{UJ})^T \ . \tag{K.54}$$

Note that θ^p_{UJ} are generally different from θ^p_U, and that the unsplit parameters can be correlated with the split parameters. We can then explicitly calculate the parts of the covariance between such update parameter differences.

We begin by considering the covariance of the split parameter differences, which results in:

$$\langle (\theta^p_{iC} - \theta^p_{CJ})(\theta^p_{jC} - \theta^p_{CJ})^T \rangle = \mathcal{C}_{pijC} - \mathcal{C}_{pCJ} \ , \tag{K.55}$$

where $\mathcal{C}_{pCJ} = \langle (\theta^p_{CJ})(\theta^p_{CJ})^T \rangle - \langle (\theta^p_{CJ}) \rangle \langle (\theta^p_{CJ})^T \rangle$ is the covariance of the parameters in the split part of the joint set. We have used the fact that $\langle (\theta^p_{iC})(\theta^p_{CJ})^T \rangle - \langle (\theta^p_{iC}) \rangle \langle (\theta^p_{CJ})^T \rangle = \mathcal{C}_{pCJ}$.

Similarly to the above, we can calculate the covariance of the unsplit parameter differences

as:

$$\langle(\theta^p_U - \theta^p_{UJ})(\theta^p_U - \theta^p_{UJ})^T\rangle = \mathcal{C}_{pU} - \mathcal{C}_{pUJ} \ , \tag{K.56}$$

where the covariance matrices $\mathcal{C}_{pU}$ and $\mathcal{C}_{pUJ}$ correspond to the unsplit part of the copy and joint parameter sets respectively in the same manner as for the split parameters above. Finally, we can calculate the covariance between split and unsplit parameter differences, which yields:

$$\langle(\theta^p_{iC} - \theta^p_{CJ})(\theta^p_U - \theta^p_{UJ})^T\rangle = \mathcal{C}_{piU} - \mathcal{C}_{pCUJ} \ . \tag{K.57}$$

In the above we have defined the covariance $\mathcal{C}_{piU}$ between the copy i and unsplit copy parameters and $\mathcal{C}_{pCUJ}$ between the split and unsplit joint parameters, again as above.

We can now comment on the relation between the $Q^C_{\rm DM}$ and $Q^C_{\rm UDM}$ estimators and their statistical significance. As in the case of two parameter copies without unsplit parameters, which is discussed in Sec. K.5.1, their significance is the same for the maximum likelihood parameters, since

$$\begin{pmatrix} \theta^{\rm ML}_{iC} \\ \theta^{\rm ML}_U \end{pmatrix} - \theta^{\rm ML}_J = \mathcal{C}_J D^T_C \mathcal{C}^{-1}_C \begin{pmatrix} D_{SC}\theta^{\rm ML}_{iC} - \theta^{\rm ML}_{SC} \\ \mathbf{O} \end{pmatrix}$$

where $\theta^{\rm ML}_{SC} = (\theta^{\rm ML}_{1C}, \theta^{\rm ML}_{2C}, \ldots, \theta^{\rm ML}_{nC})^T$ and the zero vector $\mathbf{O}$ has length N_U.

At the maximum posterior level the two statistics can differ, however, since the update parameter shifts contain only one copy of the prior in the joint but the prior is applied once to each set in the split analysis. As we did in Sec. K.5.1 here we can also define the joint parameter estimate $\tilde{\theta}^p_J = (\tilde{\theta}^p_{CJ}, \theta^p_{UJ})^T$ that counts the prior n times and has covariance $\tilde{\mathcal{C}}^{-1}_{pJ} = \mathcal{C}^{-1}_J + D^T_C \mathcal{C}^{-1}_{\Pi C} D_C = D^T_C \mathcal{C}^{-1}_{pC} D_C$. Then, the update parameter shifts would be defined as

$$\begin{pmatrix} \theta^p_{iC} \\ \theta^p_U \end{pmatrix} - \tilde{\theta}^p_J = \tilde{\mathcal{C}}_{pJ} D^T_C \mathcal{C}^{-1}_{pC} \begin{pmatrix} D_{SC}\theta^p_{iC} - \theta^p_{SC} \\ \mathbf{O} \end{pmatrix} \tag{K.58}$$

where $\mathbf{O}$ is the zero vector of length N_U. Therefore, the statistical significance of the update differences $(\theta^p_{iC}, \theta^p_U)^T - \tilde{\theta}^p_J$ is the same as that of the parameter shifts $\theta^p_C - D_{SC}\theta^p_{iC}$ since they are related by a linear and invertible transformation. We can then always use $\tilde{\theta}^p_J$ to rewrite the update parameter difference as:

$$\begin{pmatrix} \theta^p_{iC} \\ \theta^p_U \end{pmatrix} - \theta^p_J = \left[\begin{pmatrix} \theta^p_{iC} \\ \theta^p_U \end{pmatrix} - \tilde{\theta}^p_J \right] + (\tilde{\theta}^p_J - \theta^p_J) \ . \tag{K.59}$$

To complete the generalization in the case of n parameter copies with unsplit parameters, we now discuss how the statistics of goodness-of-fit loss both at the ML level, through $Q^C_{\text{DML}} \equiv X^T_J A^C_{\text{DML}} X_J$, and at the level of MAP, through $Q^C_{\text{DMAP}} \equiv X^T_J A^C_{\text{DMAP}} X_J$, can be computed. To do so we construct the matrices $A^C_{\text{DML}}\mathcal{S}_J$ and $A^C_{\text{DMAP}}\mathcal{S}_J$, respectively for ML and MAP, where we define the covariance matrix $\mathcal{S}_J \equiv \Sigma_J + M_J \mathcal{C}_\Pi M^T_J$ for Gaussian priors.

Doing so it can be shown that it is still true that the joint projector is a subset of the copy one, and thus Q^C_{DML} is chi-squared distributed with $\text{rank}(\mathbb{I}-\mathbb{P}_J) - \text{rank}(\mathbb{I}-\mathbb{P}_C)$ degrees of freedom.

Furthermore, Q_{DMAP} can be approximated by a chi-square distribution by matching moments of the approximate and exact distributions; the mean will then be given by the equivalent of Eq. (K.51) if we define $N^C_{\text{eff}} = nN_C + N_U - \text{tr}[\mathcal{C}^{-1}_{\Pi C}\mathcal{C}_{pC}]$.

At last we highlight that, with the MCMC chain of multiple parameter copies we can easily construct the distribution of parameter differences and proceed with the statistical significance calculation as in Sec. K.5.2 to compute the overall statistical significance of multiple parameter shifts.

REFERENCES

Abate, A. & Lahav, O. 2008, Mon. Not. Roy. Astron. Soc., 389, 47

Abazajian, K., Zheng, Z., Zehavi, I., et al. 2005, ApJ, 625, 613

Abbas, U., de la Torre, S., Le Fèvre, O., et al. 2010, Monthly Notices of the Royal Astronomical Society, 406, 1306

Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. 2018, The Astrophysical Journal Supplement Series, 239, 18

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, PhRvD, 98, 043526

Abbott, T. M. C., Aguena, M., Alarcon, A., et al. 2022, PhRvD, 105, 023520

Addison, G. E., Huang, Y., Watts, D. J., et al. 2016, Astrophys. J., 818, 132

Addison, G. E., Watts, D. J., Bennett, C. L., et al. 2018, Astrophys. J., 853, 119

Adhikari, S. & Huterer, D. 2019, Journal of Cosmology and Astroparticle Physics, 2019, 036

Aihara, H., Armstrong, R., Bickerton, S., et al. 2017, Publications of the Astronomical Society of Japan, 70, https://academic.oup.com/pasj/article-pdf/70/SP1/S8/23692373/psx081.pdf, s8

Aitken, A. C. 1936, Proceedings of the Royal Society of Edinburgh, 55, 42–48

A.M. Mathai, S. B. P. 1992, Quadratic Forms in Random Variables (Taylor and Francis)

Amendola, L., Marra, V., & Quartin, M. 2013, Mon. Not. Roy. Astron. Soc., 430, 1867

Baldauf, T., Seljak, U., Desjacques, V., & McDonald, P. 2012, PhRvD, 86, 083540

Baldauf, T., Smith, R. E., Seljak, U., & Mandelbaum, R. 2010, PhRvD, 81, 063531

Baldry, I. K., Robotham, A. S. G., Hill, D. T., et al. 2010, MNRAS, 404, 86

Bartelmann, M. & Maturi, M. 2017, Scholarpedia, 12, 32440

Baxter, E., Clampitt, J., Giannantonio, T., et al. 2016, MNRAS, 461, 4099

Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013, ApJ, 770, 57

Benítez, N. 2000, ApJ, 536, 571

Bennett, C. L., Larson, D., Weiland, J. L., & Hinshaw, G. 2014, Astrophys. J., 794, 135

Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M., & Lacey, C. G. 2000, MNRAS, 311, 793

Berlind, A. A. & Weinberg, D. H. 2002, ApJ, 575, 587

Bernal, J. L., Verde, L., & Cuesta, A. J. 2016, JCAP, 1602, 059

Bernstein, G. M. 2009, The Astrophysical Journal, 695, 652

Bhattacharya, S., Habib, S., Heitmann, K., & Vikhlinin, A. 2013, ApJ, 766, 32

Bilicki, M., Dvornik, A., Hoekstra, H., et al. 2021, arXiv e-prints, arXiv:2101.06010

Bird, S., Viel, M., & Haehnelt, M. G. 2012, MNRAS, 420, 2551

Blazek, J., Vlah, Z., & Seljak, U. 2015, Journal of Cosmology and Astroparticle Physics, 2015, 015

Brainerd, T. G., Blandford, R. D., & Smail, I. 1996, ApJ, 466, 623

Bridle, S. & King, L. 2007, New Journal of Physics, 9, 444

Bridle, S. L., Crittenden, R., Melchiorri, A., et al. 2002, MNRAS, 335, 1193

Brown, M. J. I., Zheng, Z., White, M., et al. 2008, ApJ, 682, 937

Cacciato, M., Van Den Bosch, F. C., More, S., et al. 2009, MNRAS, 394, 929

Cacciato, M., van den Bosch, F. C., More, S., Mo, H., & Yang, X. 2013, MNRAS, 430, 767

Carlberg, R. G., Yee, H. K. C., & Ellingson, E. 1997, ApJ, 478, 462

Carretero, J., Castander, F. J., Gaztañaga, E., Crocce, M., & Fosalba, P. 2015, MNRAS, 447, 646

Carroll, S. M., Press, W. H., & Turner, E. L. 1992, Annual Review of Astronomy and Astrophysics, 30, 499

Choi, A., Tyson, J. A., Morrison, C. B., et al. 2012, ApJ, 759, 101

Chu, M. & Knox, L. 2005, Astrophys. J., 620, 1

Clampitt, J., Sánchez, C., Kwan, J., et al. 2017, MNRAS, 465, 4204

Coil, A. L. 2013, in Planets, Stars and Stellar Systems. Volume 6: Extragalactic Astronomy and Cosmology, ed. T. D. Oswalt & W. C. Keel, Vol. 6, 387

Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039

Conroy, C. & Wechsler, R. H. 2009, ApJ, 696, 620

Conroy, C., Wechsler, R. H., & Kravtsov, A. V. 2006, ApJ, 647, 201

Cooray, A. 2006, MNRAS, 365, 842

Cooray, A., Hu, W., & Miralda-Escudé, J. 2000, ApJL, 535, L9

Cooray, A. & Sheth, R. 2002, Physics Reports, 372, 1

Coupon, J., Kilbinger, M., McCracken, H. J., et al. 2012, A&A, 542, A5

Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P., & Carretero, J. 2015, MNRAS, 453, 1513

Davis, M., Huchra, J., Latham, D. W., & Tonry, J. 1982, ApJ, 253, 423

de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., Valentijn, E. A., & KiDS and Astro-WISE Consortiums. 2013, Experimental Astronomy, 35, 25

de Jong, J. T. A., Verdoes Kleijn, G. A., Boxhoorn, D. R., et al. 2015, A&A, 582, A62

De Vicente, J., Sánchez, E., & Sevilla-Noarbe, I. 2016, MNRAS, 459, 3078

DeRose, J., Wechsler, R. H., Becker, M. R., et al. 2019, arXiv e-prints, arXiv:1901.02401

DES Collaboration. 2016, Mon. Not. R. Astron. Soc., 460, 1270

DES Collaboration. 2021, arXiv e-prints, arXiv:2105.13549

DESI Collaboration. 2016, arXiv e-prints, arXiv:1611.00036

Desjacques, V., Jeong, D., & Schmidt, F. 2018, PhR, 733, 1

Diemer, B., More, S., & Kravtsov, A. V. 2013, ApJ, 766, 25

Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971

Drlica-Wagner, A., Sevilla-Noarbe, I., Rykoff, E. S., et al. 2018, The Astrophysical Journal Supplement Series, 235, 33

Drory, N., Bundy, K., Leauthaud, A., et al. 2009, ApJ, 707, 1595

Dvornik, A., Hoekstra, H., Kuijken, K., et al. 2018, MNRAS, 479, 1240

Eisenstein, D. J. & Hu, W. 1998, ApJ, 496, 605

Elvin-Poole, J. et al. 2021, To be submitted to MNRAS

Erben, T., Hildebrandt, H., Miller, L., et al. 2013, MNRAS, 433, 2545

Fang, X., Eifler, T., & Krause, E. 2020, MNRAS, 497, 2699

Feroz, F., Allanach, B. C., Hobson, M., et al. 2008, JHEP, 10, 064

Feroz, F. & Hobson, M. P. 2008, MNRAS, 384, 449

Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601

Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, The Open Journal of Astrophysics, 2, 10

Flaugher, B. 2005, International Journal of Modern Physics A, 20, 3121

Flaugher, B., Diehl, H. T., Honscheid, K., et al. 2015, ApJ, 150, 150

Fortuna, M. C., Hoekstra, H., Joachimi, B., et al. 2021, MNRAS, 501, 2983

Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015, MNRAS, 448, 2987

Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2015, MNRAS, 447, 1319

Friedrich, O., Seitz, S., Eifler, T. F., & Gruen, D. 2016, MNRAS, 456, 2662

Friedrich, O., Andrade-Oliveira, F., Camacho, H., et al. 2020, arXiv e-prints, arXiv:2012.08568

Gaia Collaboration. 2021, A&A, 649, A1

Gatti, M., Sheldon, E., Amon, A., et al. 2020, arXiv e-prints, arXiv:2011.03408

Georgiou, C., Chisari, N. E., Fortuna, M. C., et al. 2019, A&A, 628, A31

Giblin, B., Heymans, C., Asgari, M., et al. 2021, A&A, 645, A105

Gillis, B. R., Hudson, M. J., Erben, T., et al. 2013, MNRAS, 431, 1439

Gneiting, T., Balabdaoui, F., & Raftery, A. E. 2007, Journal of the Royal Statistical Society Series B: Statistical Methodology, 69, 243

Grandis, S., Rapetti, D., Saro, A., Mohr, J. J., & Dietrich, J. P. 2016, Mon. Not. Roy. Astron. Soc., 463, 1416

Griffen, B. F., Ji, A. P., Dooley, G. A., et al. 2016, The Astrophysical Journal, 818, 10

Guo, H., Zheng, Z., Zehavi, I., et al. 2014, Monthly Notices of the Royal Astronomical Society, 446, 578

Guo, H., Zheng, Z., Behroozi, P. S., et al. 2016, MNRAS, 459, 3040

Guo, Q., White, S., Li, C., & Boylan-Kolchin, M. 2010, MNRAS, 404, 1111

Habib, S., Heitmann, K., Higdon, D., Nakhleh, C., & Williams, B. 2007, PhRvD, 76, 083503

Han, J., Eke, V. R., Frenk, C. S., et al. 2015, MNRAS, 446, 1356

Handley, W. & Lemos, P. 2019, Phys. Rev., D100, 043504

Hansen, S. M., McKay, T. A., Wechsler, R. H., et al. 2005, ApJ, 633, 122

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399

Hartley, W. G., Choi, A., Amon, A., et al. 2022, MNRAS, 509, 3547

Hayashi, E. & White, S. D. M. 2008, MNRAS, 388, 2

Heitmann, K., Higdon, D., Nakhleh, C., & Habib, S. 2006, The Astrophysical Journal, 646, L1

Heitmann, K., Higdon, D., White, M., et al. 2009, ApJ, 705, 156

Heitmann, K., Lawrence, E., Kwan, J., Habib, S., & Higdon, D. 2013, The Astrophysical Journal, 780, 111

Heitmann, K., White, M., Wagner, C., Habib, S., & Higdon, D. 2010, The Astrophysical Journal, 715, 104

Heitmann, K., Bingham, D., Lawrence, E., et al. 2016, ApJ, 820, 108

Herbel, J., Kacprzak, T., Amara, A., et al. 2017, Journal of Cosmology and Astroparticle Physics, 2017, 035

Heymans, C., White, M., Heavens, A., Vale, C., & Van Waerbeke, L. 2006, Monthly Notices of the Royal Astronomical Society, 371, 750

Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012, MNRAS, 427, 146

Heymans, C., Tröster, T., Asgari, M., et al. 2021, A&A, 646, A140

Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020, Astronomy & Astrophysics, 633, A69

Hirata, C. M. & Seljak, U. 2004, PhRvD, 70, 063526

Hoekstra, H., Yee, H. K. C., & Gladders, M. D. 2004, ApJ, 606, 67

Hotelling, H. 1936, Biometrika, 28, 321

Hoyle, B., Gruen, D., Bernstein, G. M., et al. 2018, Monthly Notices of the Royal Astronomical Society, 478, 592

Hu, W. & Jain, B. 2004, PhRvD, 70, 043009

Huang, Y., Addison, G., & Bennett, C. 2019, arXiv:1904.10521 [astro-ph.CO]

Huang, Y., Addison, G. E., Weiland, J. L., & Bennett, C. L. 2018, Astrophys. J., 869, 38

Hubble, E. 1934, ApJ, 79, 8

Hubble, E. P. 1926, ApJ, 64, 321

Hudson, M. J., Gillis, B. R., Coupon, J., et al. 2014, MNRAS, 447, 298

Huff, E. & Mandelbaum, R. 2017, arXiv e-prints, arXiv:1702.02600

Jain, B. & Taylor, A. 2003, Phys. Rev. Lett., 91, 141302

Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338

Jarvis, M., Bernstein, G. M., Amon, A., et al. 2020, arXiv e-prints, arXiv:2011.03409

Joachimi, B. & Bridle, S. L. 2010, A&A, 523, A1

Joachimi, B., Semboloni, E., Hilbert, S., et al. 2013, MNRAS, 436, 819

Joachimi, B., Lin, C. A., Asgari, M., et al. 2021, A&A, 646, A129

Johnston, D. E., Sheldon, E. S., Wechsler, R. H., et al. 2007, arXiv e-prints, arXiv:0709.1159

Joudaki, S., Blake, C., Johnson, A., et al. 2018, MNRAS, 474, 4894

Karpenka, N. V., Feroz, F., & Hobson, M. P. 2015, Mon. Not. Roy. Astron. Soc., 449, 2405

Kaufman, G. M. 1967, Center for Operations Research and Econometrics Discussion Paper, 44

Kerscher, M. & Weller, J. 2019, arXiv:1901.07726 [astro-ph.CO]

Kirshner, R. P., Oemler, A., J., & Schechter, P. L. 1978, AJ, 83, 1549

Krause, E. & Eifler, T. 2017, MNRAS, 470, 2100

Krause, E., Eifler, T. F., Zuntz, J., et al. 2017, arXiv e-prints, arXiv:1706.09359

Krause, E., Fang, X., Pandey, S., et al. 2021a, arXiv e-prints, arXiv:2105.13548

Krause, E. et al. 2021b, To be submitted to MNRAS

Kravtsov, A. V., Berlind, A. A., Wechsler, R. H., et al. 2004, ApJ, 609, 35

Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500

Kwan, J., Bhattacharya, S., Heitmann, K., & Habib, S. 2013, ApJ, 768, 123

Kwan, J., Heitmann, K., Habib, S., et al. 2015, The Astrophysical Journal, 810, 35

Kwan, J., Sánchez, C., Clampitt, J., et al. 2016, MNRAS, 464, 4045

Köhlinger, F., Joachimi, B., Asgari, M., et al. 2019, Mon. Not. Roy. Astron. Soc., 484, 3126

Landy, S. D. & Szalay, A. S. 1993, ApJ, 412, 64

Lange, J. U., Leauthaud, A., Singh, S., et al. 2021, MNRAS, 502, 2074

Lange, J. U., Yang, X., Guo, H., Luo, W., & van den Bosch, F. C. 2019, MNRAS, 488, 5771

Larson, D., Weiland, J. L., Hinshaw, G., & Bennett, C. L. 2015, Astrophys. J., 801, 9

Lawrence, E., Heitmann, K., Kwan, J., et al. 2017, The Astrophysical Journal, 847, 50

Leauthaud, A., Tinker, J., Behroozi, P. S., Busha, M. T., & Wechsler, R. H. 2011, ApJ, 738, 45

Leauthaud, A., Saito, S., Hilbert, S., et al. 2017, MNRAS, 467, 3024

Leauthaud, A., Amon, A., Singh, S., et al. 2022, MNRAS, 510, 6150

Lee, S., Troxel, M. A., Choi, A., et al. 2022, MNRAS, 509, 2033

Leistedt, B., Peiris, H. V., Elsner, F., et al. 2016, ApJS, 226, 24

Lemos, P., Köhlinger, F., Handley, W., et al. 2019, arXiv:1910.07820 [astro-ph.CO]

Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473

Lin, W. & Ishak, M. 2019, arXiv:1909.10991 [astro-ph.CO]

Lin, Y.-T., Mohr, J. J., & Stanford, S. A. 2004, ApJ, 610, 745

Luo, R., Fu, L., Luo, W., et al. 2022, A&A, 668, A12

Lupton, R. H., Gunn, J. E., & Szalay, A. S. 1999, AJ, 118, 1406

Ma, C.-P. & Fry, J. N. 2000, ApJ, 543, 503

MacCrann, N., Blazek, J., Jain, B., & Krause, E. 2020a, MNRAS, 491, 5498

MacCrann, N., Becker, M. R., McCullough, J., et al. 2020b, arXiv e-prints, arXiv:2012.08567

Magliocchetti, M. & Porciani, C. 2003, MNRAS, 346, 186

Mandelbaum, R., Seljak, U., Cool, R. J., et al. 2006a, MNRAS, 372, 758

Mandelbaum, R., Seljak, U., Kauffmann, G., Hirata, C. M., & Brinkmann, J. 2006b, MNRAS, 368, 715

Mandelbaum, R., Slosar, A., Baldauf, T., et al. 2013, MNRAS, 432, 1544

Mandelbaum, R., Tasitsiomi, A., Seljak, U., Kravtsov, A. V., & Wechsler, R. H. 2005, MNRAS, 362, 1451

March, M. C., Trotta, R., Amendola, L., & Huterer, D. 2011, Mon. Not. Roy. Astron. Soc., 415, 143

Marshall, P., Rajguru, N., & Slosar, A. 2006, Phys. Rev., D73, 067302

Martin, J., Ringeval, C., Trotta, R., & Vennin, V. 2014, Phys. Rev., D90, 063501

McDonald, P. & Roy, A. 2009, Journal of Cosmology and Astroparticle Physics, 2009, 020

McKay, T. A., Sheldon, E. S., Racusin, J., et al. 2001, arXiv e-prints, astro

Mead, A. J., Brieden, S., Tröster, T., & Heymans, C. 2021, MNRAS, 502, 1401

Moster, B. P., Naab, T., & White, S. D. M. 2013, MNRAS, 428, 3121

Moster, B. P., Somerville, R. S., Maulbetsch, C., et al. 2010, The Astrophysical Journal, 710, 903

Motloch, P. & Hu, W. 2018, Phys. Rev., D97, 103536

Motloch, P. & Hu, W. 2019, Phys. Rev., D99, 023506

Mucesh, S. et al. 2020, arXiv:2012.05928 [astro-ph.GA]

Muir, J., Bernstein, G. M., Huterer, D., et al. 2020, MNRAS, 494, 4454

Myles, J., Alarcon, A., Amon, A., et al. 2020, arXiv e-prints, arXiv:2012.08566

Nagai, D. & Kravtsov, A. V. 2005, ApJ, 618, 557

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Nelson, D., Springel, V., Pillepich, A., et al. 2019, Computational Astrophysics and Cosmology, 6, 2

Newman, J. A. 2008, ApJ, 684, 88

Newman, J. A. & Gruen, D. 2022, ARA&A, 60, 363

Neyman, J. & Scott, E. L. 1952, ApJ, 116, 144

Neyman, J., Scott, E. L., & Shane, C. D. 1953, ApJ, 117, 92

Nicola, A., Amara, A., & Refregier, A. 2019, JCAP, 1901, 011

Okamura, S. 2019, Proceedings of the International Astronomical Union, 15, 1–11

Padmanabhan, T. 1993, Structure Formation in the Universe (Cambridge University Press)

Pandey, S., Baxter, E. J., Xu, Z., et al. 2019, Phys. Rev. D, 100, 063519

Pandey, S., Krause, E., Jain, B., et al. 2020, Physical Review D, 102

Pandey, S., Krause, E., DeRose, J., et al. 2022, PhRvD, 106, 043520

Park, Y., Rozo, E., & Krause, E. 2021, PhRvL, 126, 021301

Park, Y. et al. 2015

Park, Y., Krause, E., Dodelson, S., et al. 2016, Physical Review D, 94

Patnaik, P. B. 1950, Biometrika, 37, 78

Peacock, J. A. & Smith, R. E. 2000a, MNRAS, 318, 1144

Peacock, J. A. & Smith, R. E. 2000b, MNRAS, 318, 1144

Peebles, P. J. E. 1984, ApJ, 284, 439

Peebles, P. J. E. & Groth, E. J. 1975, ApJ, 196, 1

Peebles, P. J. E. & Ratra, B. 2003, Rev. Mod. Phys., 75, 559

Planck Collaboration. 2020, A&A, 641, A6

Porredon, A., Crocce, M., Fosalba, P., et al. 2021, PhRvD, 103, 043503

Porredon, A., Crocce, M., Elvin-Poole, J., et al. 2022, PhRvD, 106, 103530

Porredon, A. et al. in prep., To be submitted to PRD

Prat, J., Sánchez, C., Miquel, R., et al. 2017, MNRAS, 473, 1667

Prat, J., Sánchez, C., Fang, Y., et al. 2018, PhRvD, 98, 042005

Prat, J., Baxter, E., Shin, T., et al. 2019, MNRAS, 487, 1363

Prat, J., Blazek, J., Sánchez, C., et al. 2022, PhRvD, 105, 083528

Prat, J., Zacharegkas, G., Park, Y., et al. 2023, MNRAS, 522, 412

Press, W. H. & Schechter, P. 1974, ApJ, 187, 425

Raveri, M. 2016, Phys. Rev., D93, 043522

Raveri, M. & Hu, W. 2019, Phys. Rev., D99, 043506

Raveri, M., Zacharegkas, G., & Hu, W. 2020, Phys. Rev. D, 101, 103527

Reid, B. A., Seo, H.-J., Leauthaud, A., Tinker, J. L., & White, M. 2014, Monthly Notices of the Royal Astronomical Society, 444, 476

Rozo, E., Rykoff, E. S., Abate, A., et al. 2016, Monthly Notices of the Royal Astronomical Society, 461, 1431

Ruiz, E. J. & Huterer, D. 2015, Phys. Rev., D91, 063009

Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, ApJ, 785, 104

Rykoff, E. S., Rozo, E., Hollowood, D., et al. 2016, The Astrophysical Journal Supplement Series, 224, 1

Saito, S., Leauthaud, A., Hearin, A. P., et al. 2016, Monthly Notices of the Royal Astronomical Society, 460, 1457

Samuroff, S., Mandelbaum, R., & Di Matteo, T. 2019, Monthly Notices of the Royal Astronomical Society, 491, 5330

Sánchez, C., Prat, J., Zacharegkas, G., et al. 2022a, PhRvD, 105, 083529

Sánchez, C., Alarcon, A., Bernstein, G. M., et al. 2022b, arXiv e-prints, arXiv:2211.16593

Schaan, E., Ferraro, S., Amodeo, S., et al. 2021, Phys. Rev. D, 103, 063513

Schaye, J., Vecchia, C. D., Booth, C. M., et al. 2010, Monthly Notices of the Royal Astronomical Society, 402, 1536

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525

Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, ApJ, 546, 20

Seehars, S., Amara, A., Refregier, A., Paranjape, A., & Akeret, J. 2014, Phys. Rev., D90, 023533

Seehars, S., Grandis, S., Amara, A., & Refregier, A. 2016, Phys. Rev., D93, 103507

Seldner, M., Siebers, B., Groth, E. J., & Peebles, P. J. E. 1977, AJ, 82, 249

Seljak, U. 1998, ApJ, 503, 492

Seljak, U. 2000, MNRAS, 318, 203

Seljak, U., Makarov, A., Mandelbaum, R., et al. 2005, PhRvD, 71, 043511

Sevilla-Noarbe, I., Hoyle, B., Marchã, M. J., et al. 2018, MNRAS, 481, 5451

Sevilla-Noarbe, I., Bechtol, K., Carrasco Kind, M., et al. 2021, ApJS, 254, 24

Shane, C. D. & Wirtanen, C. A. 1967, Publications of the Lick Observatory, 22, Part 1

Shapley, H. & Ames, A. 1932, Annals of Harvard College Observatory, 88, 41

Shectman, S. A., Landy, S. D., Oemler, A., et al. 1996, ApJ, 470, 172

Sheldon, E. S. & Huff, E. M. 2017, ApJ, 841, 24

Sheldon, E. S., Johnston, D. E., Frieman, J. A., et al. 2004, ApJ, 127, 2544

Sifón, C., Cacciato, M., Hoekstra, H., et al. 2015, MNRAS, 454, 3938

Singh, S., Mandelbaum, R., Seljak, U., Rodríguez-Torres, S., & Slosar, A. 2020, MNRAS, 491, 51

Takahashi, R., Sato, M., Nishimichi, T., Taruya, A., & Oguri, M. 2012, ApJ, 761, 152

Tarsitano, F., Hartley, W. G., Amara, A., et al. 2018, Monthly Notices of the Royal Astronomical Society, 481, 2018

Tasitsiomi, A., Kravtsov, A. V., Wechsler, R. H., & Primack, J. R. 2004, ApJ, 614, 533

Tegmark, M. 1997, PhRvD, 55, 5895

Tegmark, M., Hamilton, A. J. S., Strauss, M. A., Vogeley, M. S., & Szalay, A. S. 1998, ApJ, 499, 555

The LSST Dark Energy Science Collaboration. 2018, arXiv e-prints, arXiv:1809.01669

Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709

Tinker, J. L., Norberg, P., Weinberg, D. H., & Warren, M. S. 2007, ApJ, 659, 877

Tinker, J. L., Robertson, B. E., Kravtsov, A. V., et al. 2010, ApJ, 724, 878

Tinker, J. L., Weinberg, D. H., Zheng, Z., & Zehavi, I. 2005, ApJ, 631, 41

Tortorelli, L., Siudek, M., Moser, B., et al. 2021, JCAP, 2021, 013

Troxel, M. A. & Ishak, M. 2015, Physics Reports, 558, 1

Tyson, J. A., Valdes, F., Jarvis, J. F., & Mills, Jr., A. P. 1984, ApJ, 281, L59

Unruh, S., Schneider, P., Hilbert, S., et al. 2020, A&A, 638, A96

Vale, A. & Ostriker, J. P. 2006, MNRAS, 371, 1173

van Daalen, M. P., Schaye, J., Booth, C. M., & Dalla Vecchia, C. 2011, MNRAS, 415, 3649

van den Bosch, F. C., Mo, H. J., & Yang, X. 2003, MNRAS, 345, 923

van den Bosch, F. C., More, S., Cacciato, M., Mo, H., & Yang, X. 2013, Monthly Notices of the Royal Astronomical Society, 430, 725

Van Den Bosch, F. C., Norberg, P., Mo, H. J., & Yang, X. 2004, Monthly Notices of the Royal Astronomical Society, 352, 1302

van Uitert, E., Hoekstra, H., Velander, M., et al. 2011, A&A, 534, A14

van Uitert, E., Cacciato, M., Hoekstra, H., et al. 2016, MNRAS, 459, 3251

van Uitert, E., Joachimi, B., Joudaki, S., et al. 2018, MNRAS, 476, 4662

Velander, M., van Uitert, E., Hoekstra, H., et al. 2013, MNRAS, 437, 2111

Verde, L., Protopapas, P., & Jimenez, R. 2013, Phys. Dark Univ., 2, 166

Verde, L., Treu, T., & Riess, A. G. 2019, in Nature Astronomy 2019

Viola, M., Cacciato, M., Brouwer, M., et al. 2015, MNRAS, 452, 3529

Wand, M. & Jones, M. 1994, Kernel Smoothing, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Taylor & Francis)

Wang, S., Hui, L., May, M., & Haiman, Z. 2007, Phys. Rev., D76, 063503

Wang, Y. & Zhao, G.-B. 2020, Research in Astronomy and Astrophysics, 20, 158

Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11

Wechsler, R. H. & Tinker, J. L. 2018, Annual Review of Astronomy and Astrophysics, 56, 435

Weiland, J. L., Osumi, K., Addison, G. E., et al. 2018, Astrophys. J., 863, 161

White, M., Blanton, M., Bolton, A., et al. 2011, ApJ, 728, 126

Wibking, B. D., Weinberg, D. H., Salcedo, A. N., et al. 2019, Monthly Notices of the Royal Astronomical Society, 492, 2872

Yang, X., Mo, H. J., Jing, Y. P., van den Bosch, F. C., & Chu, Y. 2004, MNRAS, 350, 1153

Yang, X., Mo, H. J., van den Bosch, F. C., et al. 2007, ApJ, 671, 153

Yoo, J. & Seljak, U. 2012, Physical Review D, 86

Yoo, J., Tinker, J. L., Weinberg, D. H., et al. 2006, ApJ, 652, 26

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579

Yuan, S., Eisenstein, D. J., & Leauthaud, A. 2020, Monthly Notices of the Royal Astronomical Society, 493, 5551

Zacharegkas, G., Chang, C., Prat, J., et al. 2022, MNRAS, 509, 3119

Zehavi, I., Weinberg, D. H., Zheng, Z., et al. 2004, ApJ, 608, 16

Zehavi, I., Zheng, Z., Weinberg, D. H., et al. 2011, ApJ, 736, 59

Zhang, J., Hui, L., & Stebbins, A. 2005, Astrophys. J., 635, 806

Zheng, Z. 2004, ApJ, 610, 61

Zheng, Z., Coil, A. L., & Zehavi, I. 2007, ApJ, 667, 760

Zheng, Z., Tinker, J. L., Weinberg, D. H., & Berlind, A. A. 2002, ApJ, 575, 617

Zheng, Z., Zehavi, I., Eisenstein, D. J., Weinberg, D. H., & Jing, Y. P. 2009, ApJ, 707, 554

Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, ApJ, 633, 791

Zu, Y. 2020, arXiv e-prints, arXiv:2010.01143

Zu, Y. & Mandelbaum, R. 2015, MNRAS, 454, 1161

Zu, Y., Weinberg, D. H., Rozo, E., et al. 2014, MNRAS, 439, 1628

Zuntz, J., Paterno, M., Jennings, E., et al. 2015, Astronomy and Computing, 12, 45

Zuntz, J., Paterno, M., Jennings, E., et al. 2015, Astron. Comput., 12, 45

Zuntz, J., Sheldon, E., Samuroff, S., et al. 2018a, Monthly Notices of the Royal Astronomical Society, 481, 1149

Zuntz, J., Sheldon, E., Samuroff, S., et al. 2018b, MNRAS, 481, 1149

Zwicky, F., Herzog, E., Wild, P., Karpowicz, M., & Kowal, C. T. 1961, Catalogue of galaxies and of clusters of galaxies, Vol. I