DOCTORAL THESIS

Redshift distribution uncertainty in weak lensing cosmology

Author:
Juan P. Cordero

Supervisor:
Prof. Sarah L. Bridle
Dr. Ian Harrison

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy

in the

Department of Physics and Astronomy in the School of Natural Sciences
Faculty of Science and Engineering

2021

Contents

List of Abbreviations

Astronomical and Astrophysical:	
ΛCDM	**Λ** Cold **D**ark **M**atter
CMB	**C**osmic **M**icrowave **B**ackground
BBN	**B**ig-**B**ang **N**ucleosynthesis
BAO	**B**aryon **A**coustic **O**scillations
MOND	**MO**dified **N**ewtonian **D**ynamics
SED	**S**pectral **E**nergy **D**istribution
PSF	**P**oint **S**pread **F**unction
QSO	**Q**uasi-**S**tellar **O**bject
Surveys:	
DES	**D**ark **E**nergy **S**urvey
KiDS	**Ki**lo **D**egree **S**urvey
CFHTLenS	**C**anada-**F**rance-**H**awaii **T**elescope **Len**sing **S**urvey
LSST	**L**egacy **S**urvey of **S**pace and **T**ime
HST	**H**ubble **S**pace **T**elescope
SKA	**S**quare **K**ilometer **A**rray
SDSS	**S**loan **D**igital **S**ky **S**urvey
Software and algorithms:	
MCMC	**M**onte**C**arlo **M**arkov **C**hain
SOM	**S**elf-**O**rganizing **M**aps
PZ	**P**hotometric **Redshift**
SOMPZ	**S**elf-**O**rganizing **M**ap **P**hotometric **Redshifts**
WZ	**Clustering Redshift**
SR	**S**hear **R**atios
3sDir	**3**-**s**tep **Dir**richlet
HBM	**H**ierarchical **B**ayesian **M**odel
HMC	**H**ierarchical **M**onte**C**arlo
Dark Energy Survey specific:	
DES SV	**DES S**cience **V**erification
DES Y1	**DES Y**ear **1**
DES Y3	**DES Y**ear **3**

THE UNIVERSITY OF MANCHESTER

Abstract

Faculty of Science and Engineering
Department of Physics and Astronomy in the School of Natural Sciences

Doctor of Philosophy

Redshift distribution uncertainty in weak lensing cosmology

by Juan P. Cordero

Cosmological information from weak lensing surveys is maximised by dividing source galaxies into tomographic sub-samples for which the redshift distributions are estimated. Uncertainties on these redshift distributions must be correctly propagated into the cosmological results to fully account for statistical and systematic errors on the estimations of these distributions. In this thesis I present a new method for marginalising over redshift distributions in gravitational weak lensing and clustering cosmological analyses, called HYPERRANK, which allows discrete samples from the space of possible redshift distributions to be used, meaning the full uncertainty can be explored. In HYPERRANK the set of proposed redshift distributions is ranked according to a small (~ 1) number of summary values, which are then sampled as hyper-parameters along with other nuisance parameters and cosmological parameters in the Monte Carlo chain used for inference. This work focuses on the case of weak lensing cosmic shear analyses and demonstrate our method using simulations made for the Dark Energy Survey. HYPERRANK is compared to the common mean-shifting method of marginalising over redshift uncertainty, its numerical performance assessed and the resulting confidence contours used to validate its use in the DES Year 3 cosmology results. I also introduce the process to estimate and calibrate the distribution of source galaxy redshifts for the Dark Energy Survey Y3 analysis, including details of the SOMPZ scheme, a machine-learning algorithm to leverage deep field photometry to constrain the color-redshift relation of the wide survey galaxies. We describe the process to estimate the uncertainty associated to the different systematic effects involved in the estimation of the source redshift distributions, with an emphasis on the effect of sample variance and the stochastic nature of the SOMPZ training phase.

Declaration of Authorship

I, Juan P. Cordero, declare that this thesis titled, "Redshift distribution uncertainty in weak lensing cosmology" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Copyright Statement

"Nothing compares to the measurement of the Hubble constant in bringing out the worst in astronomers."

Christopher Kochanek

Acknowledgements

This thesis is the combination of a personal and collective effort to analyse and understand the Universe, and could have not been possible without the direct and indirect support, understanding, mentorship, friendship and patience from multiple people I met here in Manchester and around the world over the last four years.

First and foremost, many thanks to Sarah Bridle for her guidance, support and commitment to put the human aspect of work and well-being of people above everything else. I will always admire your extraordinary ability to ask the right questions to help me understand the challenges of weak lensing cosmology.

Also, many thanks to Ian Harrison for helping me navigate through the difficulties of our project, for your thoughtful advice and friendship.

My sincere thanks also goes to my examiners Jens Chluba and David Bacon, for a very entertaining viva and their useful comments and encouragement for the future.

As a proud member of the Bureau of Farming and Cosmology I'm grateful to my colleagues in 3.106: Nicolas, Lucía, Ian, Alana, Eleonora, and teammates on the *Bridle Party* triathlon team, Joe and Richard.

Special thanks to my Chilean-Mancunian family: Carlos, Diego, Pía, Camila, Pepe, Cristián, Emma, Aurora, Mónica, Pal, Natalia and Miguel; for making Manchester feel just like home.

Finally, thanks to my family and specially my parents Tere and Julio, for their unconditional support and love and for giving me the freedom, courage and strength to embark in this crazy adventure.

I also acknowledge and thank the support granted by Agencia Nacional de Investigación y Desarrollo (ANID) DOCTORADO BECAS CHILE/2016 - 72170279.

Supporting publications

Part of the work presented in this thesis includes previously published collaborative work or work currently in preparation, as described below:

- Chapter 3 contains an adaptation of Myles et al. (2020), and references to Hartley et al. (2020a) part of the DES Y3 batch of papers in which the author of this thesis is a collaborator and co-author. The contributions of the author of this thesis to this publication include: co-authoring of the SOMPZ code, contributions to the estimations of uncertainty associated to sample variance and the random training of the SOMPZ scheme, and contribution to the Deep field imaging pipeline.

- Chapters 4 and 5 describe the method and validation tests presented in Cordero et al. (2020), to be submitted to the MNRAS and led by this author as part of the DES Y3 batch of papers. As main author of the paper, the author of this thesis designed the code, implemented the one- and multi-dimensional HYPERRANK approaches as a CosmoSIS module and designed and performed the validation and performance tests. Also, some of the code developed and presented in this chapter was included in the SkyPy software package (SkyPy Collaboration et al., 2021), developed by the SkyPy collaboration of which the author of this thesis is a member and contributor. This includes a CosmoSIS module to implement the Eisenstein & Hu linear matter power spectrum approximation.

- Chapter 6 presents results on real data applied to the third year of observations of the Dark Energy Survey. These results will be presented in Amon et al. (2020) and Secco et al. (2020) for the Cosmic shear analysis, and in DES Collaboration et al. (2020) for the 3x2pt case, in all of which the author of this thesis is a collaborator and co-author. The author contributed with the configuration of the CosmoSIS runs, including the fiducial analysis using Δz and the alternative using HYPERRANK, obtaining the 3SDIR redshift distributions, and running chains for the final published results.

Chapter 1

Principles of Modern Cosmology

Cosmology is perhaps one of the oldest scientific disciplines, given the questions it typically tries to answer. As such, it has evolved from a philosophical and speculative affair to a highly technical branch of modern science where multiple independent and detailed analysis of large datasets converge to help understanding the details of complex models used to describe the Universe and its constituents. The current era of Cosmology, usually referred to as the era of *precision Cosmology*, is characterized by the analysis of large[1] surveys where different probes are used to test the models believed to describe the large scale structure of the Universe and its evolution.

Because of the large size of these datasets, the dominant source of uncertainty in the measurements is quickly migrating from statistical uncertainty to model and methodological uncertainty, caused by our effort to measure various continuous properties of the Universe using sets of discrete data with increasingly complex tools. Another contributing factor to this paradigm is the fact that many of the current experiments are carried out by large collaborations with hundreds of members. This requires a level of coordination and trust that can become a systematic itself; blinded analyses have become a necessary step to shield us from confirmation biases.

We are in an era where we spend more and more time trying to identify and account for statistical and systematic errors associated to our observations and analysis tools and choices. Still, there are several sources of tension which must be resolved either by identifying the systematic effects causing them or adding complexity to our models. Discrepancy between low and high redshift probes, for instance, is one big open question which new surveys will aim to answer in the next decades. Nevertheless, the current model has been tremendously successful at predicting the observations made at large scale ranges and evolutionary stages of the Universe. In this introduction we will describe the current standard cosmological model and give a qualitative description of the main observational probes used in modern experiments. We will start by introducing the standard model of gravity in section 1.1 which describes the interplay between the components of the Universe and the expansion of space-time and how structures form

[1] Here, *large* refers to the size of datasets, which can be achieved by either surveying a large fraction of the sky, observing it with more sensitivity, or both.

from primordial fluctuations. Then, in section 1.3 we will provide a review of the current observational probes to constrain the parameters of the standard model and finalize with a brief description of future planned cosmological surveys in section 1.4.

1.1 The standard cosmological model

The Universe and the structures that exist in it evolve over time as a consequence of the interaction between its fundamental constituents. The framework describing these relations is called the **Standard Cosmological Model**, which encompasses not only the physical phenomena in a mathematical form, characterized by a set of parameters which establish the limits, intensities, relative importance and general properties of this mathematical description. While the employed parameterisation for an accurate model can use an arbitrarily large number of parameters, it is often desirable to being able to describe it with as little parameters as possible. Not only this is more easily manageable from a practical point of view, but helps accepting the idea that multiple aspects of the model can be explained from a simpler and more fundamental origin.

1.1.1 A expanding Universe

The current standard model of the Universe is called the ΛCDM model, and it assumes the three components, radiation, matter, and Dark Energy, interact with each other according to General Relativity, via the Einstein field equations

$$\begin{aligned} G_{\mu\nu} + \Lambda g_{\mu\nu} &= \frac{8\pi G}{c^4} T_{\mu\nu} \\ R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} &= \frac{8\pi G}{c^4} T_{\mu\nu} \end{aligned} \tag{1.1}$$

Einstein Field Equations describe how spacetime, a four-dimensional representation of the physical and temporal coordinates of space, curves as a consequence of the mass, energy and momentum of the components residing on it. $G_{\mu\nu}$ is the Einstein tensor describing the curvature of space, $R_{\mu\nu}$ is the Ricci curvature tensor which can be broadly interpreted as the deviation of an Euclidean geometry, $T_{\mu\nu}$ is the stress–energy tensor which describes the energy and momentum of the components, which are assumed to form a perfect fluid with isotropic pressure p, Λ is the cosmological constant, G is the gravitational constant and R is the scalar curvature. These equations solve for the metric tensor $g_{\mu\nu}$ which describes geometry of spacetime in the presence of mass, energy and momentum, can be used to describe the *geodesic* trajectories, which describe the motion of inertial bodies and photons. Einstein field equations relates this set of symmetric 4×4 tensors, which reduce to a set of 10 partial differential equations. In this framework, the space is distorted and can expand or contract as a consequence of the presence of the different components of the Universe. To accommodate for this potential expansion or contraction, we start by defining a system of reference where the physical distances $\mathbf{r}$

between points in space can be described by

$$\mathbf{r}(\chi, t) = a(t)\chi, \tag{1.2}$$

where χ is the distance the points would have in the absence of any contraction or expansion, called comoving distance; and $a(t)$ is the scale factor that describes the the expansion factor of the coordinates of space at a given time t. This scale factor is normalized such that it becomes unity today ($t = t_0$). Because of the relation between proper and comoving distance given in equation 1.2, we can infer that the distance to a source in an expanding universe will be proportional to the rate at which the space expands.

$$\begin{aligned} \frac{d}{dt}\mathbf{r}(\chi, t) &= \frac{d}{dt}a(t)\chi + \mathbf{v_{pec}} \\ \dot{\mathbf{r}}(\chi, t) &= \frac{\dot{a}(t)}{a(t)}\mathbf{r}(\chi, t) + \mathbf{v_{pec}} \\ \dot{\mathbf{r}}(\chi, t) &= H(t)\mathbf{r}(\chi, t) + \mathbf{v_{pec}} \end{aligned} \tag{1.3}$$

where $H(t)$ is called the Hubble parameter, typically expressed in units of $\mathrm{km s^{-1}\,Mpc^{-1}}$, which indicates how much the space is expanding per unit time, per unit distance, and $\mathbf{v_{pec}}$ is the peculiar velocity of the source with respect to its local environment, which for cosmological distances is typically very small and negligible. The value of this parameter at the present time, $H_0 = H(t)|_{t_0}$ is referred to as the **Hubble constant** and its value has been measured to be around 70 $\mathrm{km s^{-1}\,Mpc^{-1}}$ (Planck Collaboration et al., 2020; de Jaeger et al., 2020; Riess et al., 2019). Two useful quantities can be derived from H_0: the *Hubble time*, H_0^{-1}, which can be interpreted as the age of the Universe as inferred from its current size and assuming a constant expansion given by H_0; and the *Hubble radius*, c/H_0, which determines the size of the observable Universe, had it existed for a period of time equal to the Hubble time.

1.1.2 Friedmann equations in a FLRW metric

A metric that describes distances between two points in space-time considering the potential expansion or contraction of the Universe is the Friedmann-Lemaitre-Robertson-Walker metric,

$$dr^2 = c^2dt^2 - a^2(t)\left[d\chi^2 + S_K^2(\chi)(d\theta^2 + d\phi^2\sin^2\theta)\right], \tag{1.4}$$

which introduces the remaining coordinates of a spherically symmetric system, (θ, ϕ), in addition to the comoving distance χ. This metric describes a solution to the Einstein field equations when the distribution of mass, energy and momentum is isotropic and homogeneous. While this is clearly not the case on small scales (the solar system, for example, is **very** inhomogeneous and non-isotropic), it holds on large cosmological scales. Because of the expansion (or contraction) of space, distances between points depend on the induced curvature which is parameterised by the parameter K. $S_K(\chi)$ is defined to

account for the three possible geometries of the isotropic homogeneous universe

$$S_k(\chi) = \begin{cases} \frac{1}{\sqrt{K}} \sinh\left(\sqrt{K}\chi\right) & K > 0;\ \text{Positive curvature} \\ \chi & K = 0;\ \text{No curvature} \\ \frac{1}{\sqrt{|K|}} \sin\left(\sqrt{|K|}\chi\right) & K < 0;\ \text{Negative curvature} \end{cases} \tag{1.5}$$

describing an spherical, flat, and hyperbolic spacetime, respectively.

By assuming a homogeneous and isotropic universe via 1.4, and that it is filled by an ideal, collisionless fluid the Einstein field equations 1.1 greatly simplify from a set of 10 non-linear differential equations to two independent differential equations describing the behaviour of the scale factor $a(t)$, called the Friedmann equations

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

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

where $\rho(t)$ and $p(t)$ describe the density and pressure of the perfect fluid. From this equation we can separate three cases for the three possible curvature values K which describe an eternally expanding universe ($K < 0$), a universe which initially expands and then collapses on itself ($K > 0$), and a critical case where expansion asymptotically halts as $t \to \infty$ for $K = 0$. The critical mass density required for such scenario, ρ_{cr} is found by setting $t = t_0$ and defining

$$\rho_{\text{cr}} = \frac{3H_0^2}{8\pi G}. \tag{1.8}$$

We can define the relative energy density of all components with respect to ρ_{cr} as

$$\Omega_X = \frac{\rho_X}{\rho_{\text{cr}}} \tag{1.9}$$

1.1.3 Cosmological redshift

The relation between proper distance and the rate at which cosmological objects move away from each other is a fundamental tool used to compute distances since it provides an accurate estimation in the regime where the peculiar velocities of objects is small compared to cosmic expansion. To estimate this distance we typically measure the scale factor a by using the concept of *redshift* z, which measures the change in wavelength of radiation emitted from sources caused by the expansion of the Universe. To find this relation, we assume that light emitted from a source travels radially along a null geodesic ($dr^2 = 0$) on an homogeneous, isotropic universe. Hence, it stays at a fixed set of angular coordinates (θ, ϕ). From equation 1.4 we have for a radial trajectory towards the observer that

$$c\frac{dt}{a(t)} = d\chi. \tag{1.10}$$

If we integrate the total proper distance traveled by radiation emitted from the origin at $t = t_e$ until observed at distance χ at the time $t = t_0$, and then repeat for the same type of radiation emitted one wave period after the first one, $\Delta t = \lambda/c$, because of the expansion (or contraction) of the Universe the difference in arrival time between these two waves will not necessarily be the same as the difference in emission time, but in all cases the proper distance traveled will be the same. Hence, we have

$$\int_{t_e}^{t_0} \frac{dt}{a(t)} = \int_{t_e+\lambda_e/c}^{t_0+\lambda_0/c} \frac{dt}{a(t)}. \tag{1.11}$$

Subtracting $\int_{t_e+\lambda_e/c}^{t_0} \frac{dt}{a(t)}$ to both sides and assuming that the scale factor does not change significantly in the period of time Δt, then we get

$$\frac{1}{a(t_e)} \int_{t_e}^{t_e+\lambda_e/c} dt = \frac{1}{a(t_0)} \int_{t_0}^{t_0+\lambda_0/c} dt \tag{1.12}$$

which reduces to

$$\frac{\lambda_e}{a(t_e)} = \frac{\lambda_0}{a(t_0)}. \tag{1.13}$$

This last equation relates the wavelengths from radiation emitted and observed to the scale factors of the Universe for those two times. Redshift z is then defined as the fractional change in wavelength between emitted and observed radiation,

$$z = \frac{\lambda_0 - \lambda_e}{\lambda_e},$$

which leads to

$$1 + z = \frac{1}{a(t_e)}. \tag{1.14}$$

This useful relation allows us to interpret redshift as a measure of distance and time, since we can relate it to a specific evolutionary state of the Universe described by the scale factor $a(t)$. ~~All of the time dependent quantities in the Friedmann equation can be written as a function of redshift instead.~~

Since the speed of light c is finite, and in theory any interaction or information travels slower that c, the sphere of influence of the environment around any point in space is determined by the distance light has had the time to travel from that point. We can define the horizon distance at a time t (or redshift z), $r_{H,\mathrm{com}}$ as the comoving distance traveled by light from $t = 0$ up to that instant

$$r_{H,\mathrm{com}} = \int_0^t \frac{cdt}{a(t)}$$
$$r_{H,\mathrm{com}} = \int_0^{1/(1+z)} \frac{cda}{a^2 H(a)}. \tag{1.15}$$

Where we have used the definition of the Hubble parameter, $H = \dot{a}/a$ and 1.14. This definition will be useful when we describe the growth of structures in the Universe and how different physical scales become causally connected as they enter the horizon.

Component	Ω_X	ω_X	Evolution
Radiation	10^{-4}	$1/3$	$(1+z)^4$
Baryonic Matter	0.049	0	$(1+z)^3$
Cold Dark Matter	0.264	0	$(1+z)^3$
Dark Energy	0.684	~ -1	constant?

TABLE 1.1: Summary of relative energy densities Ω_X with respect to the critical density ρ_c, equation of state ω_X and scale factor / redshift dependency of the energy density for each identified component of the standard model. Values shown here correspond to the reported TT,TE,EE+lowE+lensing values reported in Planck Collaboration et al. (2020)

1.1.4 Equations of state and expansion history

The Friedmann equations (1.6) cannot fully describe the dynamics of the expansion of space-time on their own, since they depend on three separate quantities, $a(t)$, $\rho(t)$ and $p(t)$, and only provide a system of two equations. The standard model assumes there are three main components in the Universe, each described by its own **equation of state** which relates the energy density ρ and the excerpted pressure p in the form

$$p_X = \omega_X \rho_X c^2, \tag{1.16}$$

where X refers to a particular fluid or *component* of the Universe. We can find the evolution of the energy density by assuming first that in a homogeneous universe, the perfect fluid expands adiabatically. For a sphere of fluid of comoving radius r_s, its volume will be $V = \frac{4\pi r_s^3}{3} a(t)$ and its total energy will be $E = V\rho c^2$. Energy conservation tells us that

$$\begin{aligned} dQ = dE + PdV = 0 \\ \dot{E} + P\dot{V} = 0 \\ \dot{V}\rho c^2 + V\dot{\rho}c^2 + P\dot{V} = 0 \end{aligned}$$

From the expression for the volume of the sphere we find that

$$\dot{V} = 3\frac{\dot{a}}{a} V$$

Replacing in the above expression for the energy conservation, we find the *fluid equation*

$$\rho c^2 + 3\frac{\dot{a}}{a}\left(\rho c^2 + \omega \rho c^2\right)$$

which has solutions of the form

$$\rho_X(z) = \rho_{X,0}(1+z)^{3(1+\omega_X)}, \tag{1.17}$$

where $\rho_{X,0}$ is the present day mean energy density for component X.

Radiation

Strictly speaking, radiation refers to any kind of particle moving at relativistic speeds, with radiation in the form of photons being the most common. A very small fraction of this radiation comes from stellar or *baryonic* sources, while the vast majority corresponds to the relic radiation of the cosmic microwave background (CMB). Radiation pressure is related to its energy density ρ_γ via

$$P_\gamma = \frac{1}{3}\rho_\gamma c^2 \tag{1.18}$$

which allows us to infer the equation state of radiation is characterized by $\omega_\gamma = 1/3$. From this, we can obtain the dependence of the energy density on the scale parameter by substituting ω_γ in 1.16 to obtain

$$\rho_\gamma \propto (1+z)^4 \propto a^{-4}. \tag{1.19}$$

The exponent on the scale factor dependency comes from its spatial dependency which contributes a factor a^{-3}, similar to non-relativistic matter and the stretching of wavelength λ due to cosmological expansion which contributes a factor a^{-1} originating in the photon energy equation $E = hc/\lambda$.

Neutrinos are another type of particle that also fits the definition of radiation, and are predicted by the standard model of particle physics. They exist in three *flavors*, $|\nu_e\rangle$, $|\nu_\mu\rangle$ and $|\nu_\tau\rangle$, each associated to the decay of its associated leptons; electrons, muons and τ leptons respectively. Unlike photons, neutrinos interact via gravity and weak nuclear force with other species but not via electromagnetism, but like them, their equation of state is also characterized by $\omega_\nu = 1/3$, which means their energy density evolves as $\rho_n u \propto a^{-4}$.

Because of the way they interact with other species, they directly impact cosmological observables such as the Cosmic Microwave Background and the Matter Power Spectrum. At early epochs, when moving at relativistic speeds they enhance the radiation density, moving the radiation-matter equality to lower redshift, suppressing CMB peaks as the period of expansion during radiation dominated epoch is longer. When the temperature of the neutrino component decreases to non-relativistic levels, the neutrino component enhances the dark matter component as they do not interact with radiation in the same way as baryonic matter. This produces an enhanced formation of small scale features in the matter power spectrum. The neutrino component is typically characterized by the effective number of species and the sum of the masses of the different neutrino species. CMB and Matter Power Spectrum measurements can be used to constrain both parameters of the neutrino components.

Baryonic matter

In the standard model of particle physics *baryons* are particles formed by three or more odd number of quarks, and part of the *hadron* family of particles. Out of the many compatible combinations of quarks which can result in a baryon only electrons, protons and neutrons are stable, and the last two are the most massive, with the mass of the electron being a small fraction to that of the proton. Baryons combine to form atoms, with a vast majority of the baryonic matter of the Universe being made of Hydrogen ($\sim$ 75%) and Helium $\sim$ 25%. A small fraction of other stable elements like Deuterium and Lithium were formed shortly after the Big Bang in a process called **Big-Bang nucleosynthesis** (BBN), and the rest of the heavy elements are formed predominantly in stellar interiors, Supernovae, Neutron star collisions (Chakrabarti et al., 1987) and in the accretion disks of black holes (Berger et al., 2013). An important aspect of baryons is that they absorb and emit photons which means their dynamics are closely connected to that of radiation. Current best estimates of the baryon density from BBN and cosmic microwave experiments (Planck Collaboration et al., 2020) yield $\Omega_b \sim 0.049$, although surveys based on stellar distributions of galaxies, interstellar and intergalactic medium and small compact objects only account for a fraction of it. The remaining fraction is believed to exist in highly hot and diffuse gas barely visible in the X-rays in halos and filaments of the large-scale structure (e.g. Gupta et al., 2012). From 1.16 and the assumption that the matter component can be approximated as a collisionless, pressureless fluid, we have that $\omega_m = 0$ and hence the energy density of matter scales as

$$\rho_m \propto (1+z)^3 \propto a^{-3}. \tag{1.20}$$

Dark Matter

We mentioned above that the surveys to quantify the sources of baryonic matter barely add up to the values expected from BBN and CMB experiments. Furthermore, even if we were able to find all the sources of matter to account for the primordial Ω_b we would still have a problem, since a series of basic observations made in the early twentieth century suggested that the total matter density of the Universe is almost an order of magnitude larger.

- **Velocity dispersion of cluster galaxies** Virial theorem can be used to estimate the mass of galaxy clusters and globular clusters from the dispersion of velocities of its members, which can be estimated from its radial velocities using redshift (Zwicky, 1933). Comparison with the estimated mass of the luminous component usually results in a $\sim$ 80% deficiency with respect to the virial masses, which suggest a large fraction of the mass is in non radiation emitting form.

- **Galaxy rotation curves** Similarly to the velocity dispersion argument, the mass of rotating galaxies can be estimated from tracing the luminous components, like gas and stars. For a spiral galaxy with a mass distribution traced by this component it is

expected that the tangential speed $V_\perp$ of stars around the galactic centre decreases at high radii as $\sim 1/\sqrt{r}$. This can derived from assuming that the acceleration of a particle in circular orbit of radius r, enclosing a mass $M(< r)$ is

$$a_\perp = \frac{V_\perp^2}{r} = \frac{GM(< r)}{r^2}$$

where $a_\perp$ is the centripetal acceleration and $V_\perp$ is the tangential velocity of the particle in circular orbit. At large radii the enclosed mass $M(< r)$ is essentially constant, hence $V_\perp \propto 1/\sqrt{r}$ Several observations suggest that the rotation curves of many spiral galaxies, including Andromeda, are flat up to very large values of r, well past its stellar component, which suggests galaxies are embedded in a large extended halo of invisible matter (See e.g.: Babcock, 1939; Rubin et al., 1980; Persic et al., 1996).

If we then assume that this missing matter has the same properties as baryonic matter, we then enter in conflict with the estimates of BBN and CMB. While General Relativity is widely considered as the most successful theory of gravity, alternative models based on modifications of the classical Newtonian dynamics exist to explain some these phenomena (MOND, see Famaey & McGaugh (2012) for a review), usually based on the assumption that Newtonian dynamics behave differently at different scales. Under the assumption of General Relativity, there is a massive component of the Universe, which accounts up to 27% of the critical density $\Omega \sim 0.27$, interacts gravitationally with the rest of the components but does not interact with radiation at all. We call it dark matter, almost as a recognition of our current ignorance on its exact nature. Matter in general is assumed to be pressureless when moving to speeds $v \ll c$. We usually refer to it as *cold dark matter*. For this reason, the equation of state parameter of Dark Matter is assumed to be very close to zero, similar to baryonic matter. Hence, the evolution of the Dark Matter energy density follows the same dependency with the scale factor shown in equation 1.20.

Dark Energy

During the early twentieth century Einstein pictured the Universe to be infinitely old, spatially finite and stationary, which required the addition of a component with negative pressure balancing the effect of gravity. He added this component to his field equations using a constant negative term Λ, known as the cosmological constant. Soon afterwards Edwin Hubble reported the first measurements of the receding velocity of distant objects which when compared against their redshifts gave evidence of a expanding Universe. This constant expansion solved the inconsistencies seen in the field equations, removing any need for a cosmological constant. It wasn't until the discovery of the accelerated expansion of the Universe (See section 1.3.2) that the concept was brought back to life.

A constant scale independent energy density eventually dominates the total energy density as the rest of the components eventually dilute because of their dependency with

the scale factor. In order to cause an expansion of the space-time its equation of state must be described by a negative pressure. These two conditions result in a equation of state parameter $\omega = -1$. In more general terms, the equation of state that allows for an accelerated expansion can allow for a time dependent energy density. A common parameterisation of the equation of state parameter then takes the form (Chevallier & Polarski, 2001)

$$\omega(a) = \omega_0 + \omega_a(1-a). \tag{1.21}$$

allowing for a time-dependent departure from a constant equation of state $\omega = -1$ The simplest explanation for the origin of this component assumes that the vacuum in space posses a base energy level denoted Λ, which when applied to the Einstein's field equations 1.1 predicts an accelerated expansion of the scale factor when it dominates over the matter component. Other alternative theories to explain observations include *Quintessence* models (Ratra & Peebles, 1988) in which Dark Energy corresponds to the potential energy of a dynamic field which can vary as a function of time and space, unlike the cosmological constant models. Because of our ignorance about its true origin, we employ an placeholder name to refer to it: **Dark Energy**. As a consequence of the observed flat curvature of the Universe and the estimations of the matter component, it is estimated that the Dark Energy energy density relative to the critical density is $\Omega_\Lambda \sim 0.7$.

1.1.5 Hubble expansion

The three components, matter (Dark and Baryonic), radiation and Dark Energy are characterized by their equation of state values ω_X, listed in table 1.16. Combining 1.17 and 1.6 using the values from table 1.1 yield the known differential equation for the expansion history

$$\frac{\dot{a}(t)}{a(t)} = H(t) = H_0 \left[\Omega_m a(t)^{-3} + \Omega_r a(t)^{-4} + \Omega_K a(t)^{-2} + \Omega_\Lambda\right]^{1/2}, \tag{1.22}$$

where $\Omega_K =$ is a term describing the curvature of space-time, and Ω_r is the energy density of radiation coming from photons and neutrinos. While this equation can only be solved numerically, a quantitative description of the Hubble expansion can be made by assuming the dominance of the components at different ages of the Universe. Because of the dependency of their energy densities with the scale factor a, the different components dominate the energy budget of the Universe at different redshift ranges, starting with radiation shortly after the big-bang and up until the instant of matter radiation equality. Up to that point, equation 1.22 exhibits a slow growth of $a(t)$ proportional to $t^{1/2}$. Because of the small radiation energy density and its steep decline proportional to a^{-4} the matter domination era starts not long after the big-bang when the Universe is approximately 50,000 years old. Equation 1.22 can be solved for a matter dominated Universe resulting in an expansion proportional to $t^{2/3}$. After dominating for approximately 15 billion years, matter energy density becomes subdominant to Dark Energy, which has an energy density evolution independent of the scale factor. Solutions to equation 1.22 in this era are exponential functions $a(t) \propto e^{2H_0\Omega_\Lambda t}$. Figure 1.1 shows a numerical solution

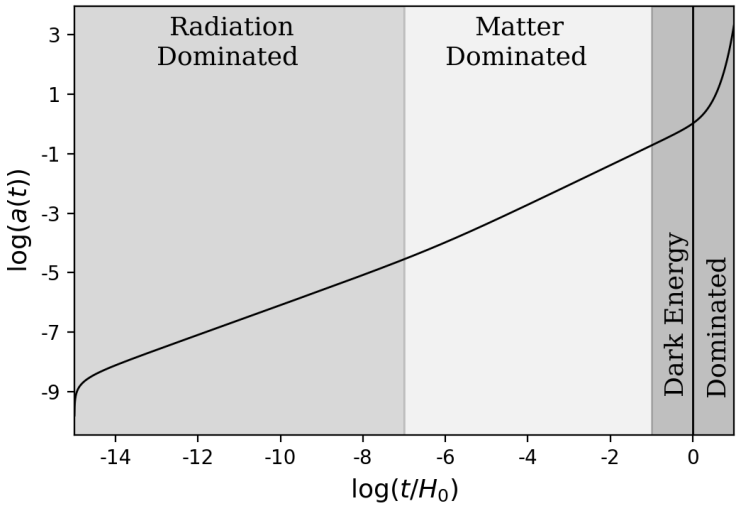

FIGURE 1.1: Scale factor $a(t)$ as a function of cosmic time, as obtained by numerically integrating equation 1.22, using $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $\Omega_\gamma = 10^{-5}$ and $\Omega_K = 1 - \Omega_m - \Omega_\Lambda - \Omega_\gamma$. The three shaded regions separate the three eras on which each component dominates the expansion history, with a growth proportional to $t^{1/2}$, $t^{2/3}$ and $e^{2H_0\Omega_\Lambda t}$ for a radiation, matter, and Dark Energy dominated universe respectively.

to the expansion equation showing the three dominance eras of radiation, matter, and Dark Energy.

1.2 Formation of structure

One of the fundamental assumptions of the standard cosmological model is that the Universe is homogeneous and isotropic on large scales. This assumption quickly breaks down on smaller scales where galaxy clusters, super-clusters and voids can be clearly seen on scales of around 100 Mpc or less. The formation of these structures follows a simple argument: small inhomogeneities of the primordial plasma in which the density of matter deviates slightly above the mean density of the surroundings grow in intensity as over-densities attract matter from under-densities because of their larger gravitational potential.

To quantitatively describe the growth of these primordial anisotropies we start by assuming the matter component is a pressureless fluid in which the relative over- and under-densities are small and cen be described by the density contrast

$$\delta(\mathbf{r}, t) = \frac{\rho(\mathbf{r}, t) - \tilde{\rho}(t)}{\tilde{\rho}(t)}, \tag{1.23}$$

where $\tilde{\rho}(t)$ denotes the mean density of matter at time t. Because of the small inhomogeneities, we can use a Newtonian description of gravity, where the behaviour of the

matter field is described by the equations of mass and momentum conservation

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.24}$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla \Phi, \tag{1.25}$$

where $\mathbf{v}$ is the velocity of the fluid field and Φ is the gravitational potential which satisfied the Poisson equation

$$\nabla^2 \Phi = 4\pi G \rho. \tag{1.26}$$

Combining these equations with the Friedmann equations 1.6 one finds that the density contrast satisfies a second-order differential equation

$$\ddot{\delta} + 2H\dot{\delta} = 4\pi G \tilde{\rho} \delta. \tag{1.27}$$

Since the equation does not contain any terms depending on spatial coordinates or derivatives with respect to them, the solutions to this differential equations are of the form

$$\delta(\mathbf{r}, t) = D(t)\tilde{\delta}(\mathbf{r}), \tag{1.28}$$

with $\tilde{\delta}(\mathbf{r})$ being an arbitrary *primordial* density contrast field, and $D(t)$ a time dependent function called the *Growth function*, which satisfies

$$\ddot{D} + 2H\dot{D} - 4\pi G \rho(t) D = 0. \tag{1.29}$$

This means that the density contrast change is *stationary*, with its shape remaining unchanged spatially, and only becoming larger *in-situ*. The solutions to these equations are linear combinations of two families, $D_-(t)$ and $D_+(t)$ which describe monotonically decreasing and increasing functions, respectively. Out of the two, we are only interested in $D_+(t)$ since they become dominant as t increases.

The evolution in time of these inhomogeneities is tightly connected to the cosmological model as the solutions to the growth factor differential equation 1.29 depend on the relative density of dark matter and cosmological constant via the Hubble parameter H. Characterizing the matter component is a fundamental step to constrain the model and estimate its parameters. Now, the model is not able to predict the behaviour at very small scales and it has no business in trying to explain the observed distribution of mass on our vicinity in particular. These peculiarities are dependent on the initial conditions which require infinite information to be described. Instead, we aim to characterize a model capable of reproducing the statistical properties of the mass distribution we observe. Two alternative and equivalent statistical descriptions of the mass distribution of the Universe are the two-point correlation functions and power spectrum.

1.2.1 Two-point correlation functions and the power spectrum

While the two-point correlation functions and the power spectrum do not unambiguously define the exact distribution of matter, they provide a complete statistical description of it as long as the matter distribution follows the properties of a Gaussian random field, in which the probability of finding a value of for the density contrast at any position follows a Gaussian distribution. We can define the spatial two-point correlation function of a quantity $f(\boldsymbol{\theta})$ as (Bartelmann & Schneider, 2001)

$$\begin{aligned} C_f(\theta) &= \langle f(\boldsymbol{\phi}) f(\boldsymbol{\phi}+\boldsymbol{\theta})\rangle \\ &= \langle \bar{f}\,[1+\Delta f(\boldsymbol{\phi})]\,\bar{f}\,[1+\Delta f(\boldsymbol{\phi}+\boldsymbol{\theta})]\rangle \\ &= \bar{f}^2\,[1+\langle \Delta f(\boldsymbol{\phi})\Delta f((\boldsymbol{\phi}+\boldsymbol{\theta}))\rangle]\,, \end{aligned} \tag{1.30}$$

where the $\langle\,\rangle$ brackets denote an average over all possible points, The two-point correlation describes the excess probability of finding two values of a continuous function above an average value $\bar{f}$ expected from a completely random distribution at an angular separation θ. The two-point correlation function between two sets of points A, B can be estimated when a large sample of values of f at different positions are obtained, using

$$\hat{C}_f(\theta) = \frac{1}{N_A N_B} \sum_{a,b}^{N_A,N_B} f(\boldsymbol{\theta}_a) f(\boldsymbol{\theta}_b), \tag{1.31}$$

where $|\boldsymbol{\theta}_a - \boldsymbol{\theta}_b| = |\boldsymbol{\theta}| = \theta$. It can also be applied to the discrete case where the observed sets of points A, B describe measurable properties of each individual point, like the position or ellipticity of a single galaxy. The power spectrum of the function f is defined as the Fourier transform of the correlation function $C_f(\theta)$

$$P_f(k) = \int_{\mathbb{R}^n} d\boldsymbol{\theta}\, e^{i\mathbf{k}\cdot\boldsymbol{\theta}} C_f(\theta) \tag{1.32}$$

The matter power spectrum is then defined from the density contrast two-point correlation function

$$P_\delta(k) = \int d\mathbf{r}\, e^{i\mathbf{k}\cdot\mathbf{r}} \langle \delta(\mathbf{r})\delta(\mathbf{r}')\rangle. \tag{1.33}$$

Because of the linearity of the Fourier transform, and ideal linear growth of the primordial density fluctuations, we should expect the present matter power spectrum to be a scaled version of the primordial power spectrum, $P_0(k)$. A common assumption on the primordial power spectrum of inhomogeneities is that it follows a power law

$$P_\delta(k) = A_s k^{n_s - 1}, \tag{1.34}$$

where A_s and n_s are the **amplitude** and **spectral index of the primordial power spectrum**. This is called the Harrison-Zel'dovich-Peebles power spectrum, who first proposed it as the power spectrum of scale-invariant primordial fluctuations, characterized by $n_s \sim 1$

We see in reality that on smaller scales the Universe does not follow the shape of this primordial power spectrum This deviation is mainly caused by three effects

- We have assumed that dark matter is *cold*, meaning that it moves at non-relativistic speeds. If a fraction of dark matter is hot dark matter (HDM) , it may not be gravitationally bounded to the matter overdensities. This effect of *free streaming* prevents small scales structures of HDM to form.
- As we saw in 1.1.5, in the early Universe radiation dominates over matter and the expansion rate $a(t)$ is different from that in the matter-dominated phase.
- Causal physical interactions can only occur on scales smaller than the horizon radius $r_{H,\text{com}}$. For scales larger than this, the Newtonian approximation is no longer valid.

To account for these effects, the time and scale dependency of the matter power spectrum is assumed to take the form

$$P_\delta(k,t) = T^2(k)\frac{D_+^2(t)}{D_+^2(t_0)}P_0(k), \tag{1.35}$$

where $T(k)$ is called the *Transfer function* which accounts for different scales entering the horizon at different times. The assumptions to linear growth of structure include that the Universe is dominated by matter, but in reality right after recombination the Universe is dominated by radiation with an energy density which we saw decays as $\rho_\gamma \propto a^{-4}$. The resulting expansion prevents the perturbations to efficiently grow as they would if the Universe was dominated by matter. In addition to this, the local effect of radiation and pressure gradients smoothing out the inhomogeneities is only valid out to the horizon, d_{hor} which is the radius around the center of a perturbation which is in causal connection. This effectively means that super-horizon (small k values) structures grow unimpeded, while small under and over-densities are suppressed until matter starts dominating. This results in an unequal growth of structure below and above the scale defined by the horizon at the moment of matter-radiation equality, L_0 (With corresponding wavenumber k_0), typically of the shape

$$T(k) \approx \begin{cases} 1 & \text{for } k \ll 1/L_0 \\ (kL_0)^{-2} & \text{for } k \gg 1/L_0 \end{cases} \tag{1.36}$$

Since the primordial power spectrum is characterized by a spectral index close to unity, this results in a power spectrum which grows almost linearly up to k_0 and then decays sharply as $P(k) \propto k^{-3}$ for smaller scales, which is clearly seen in figure 1.2.

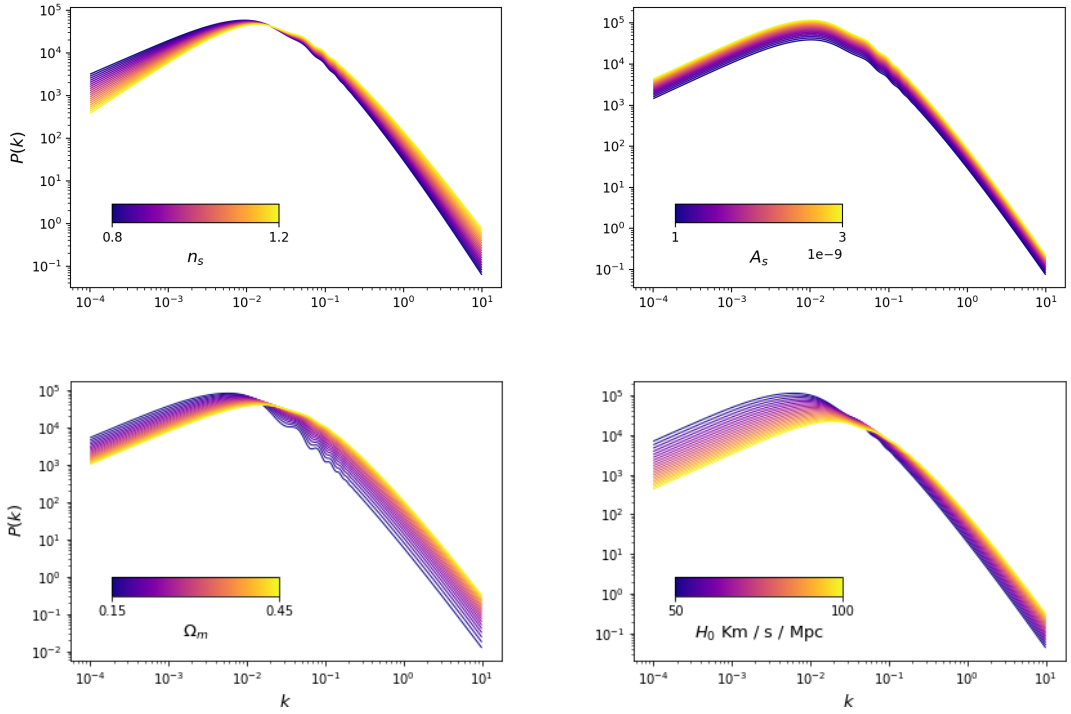

FIGURE 1.2: Matter power spectrum of linear density perturbations for different ranges of the spectral index n_s (top left), primordial amplitude A_s (top right), present day matter density Ω_m (bottom left) and Hubble constant H_0. Fiducial values when not being varied are $(n_s, A_s, \Omega_m, H_0) = (1,\ 2.2 \times 10^{-9},\ 0.3,\ 70\ \text{kms}^{-1}\ \text{Mpc}^{-1})$ In all cases the effect of the transfer function is visible, with a linear growth up until $k \sim k_0$ and a sharp decay proportional to k^{-3} afterwards.

From the definition of horizon distance in 1.15 (Schneider, 2006)

$$\begin{aligned} L_0 &= \int_0^{a_{\rm eq}} \frac{cda}{a^2 H(a)} = \frac{c}{H_0\sqrt{\Omega_m}} \int_0^{a_{\rm eq}} \frac{da}{\sqrt{a+a_{\rm eq}}} \\ &= \left(\sqrt{2}-1\right) \frac{2c}{H_0} \left(\frac{a_{\rm eq}}{\Omega_m}\right)^{1/2} \\ &\sim 16(\Omega_m h^2)^{-1}{\rm Mpc}, \end{aligned} \tag{1.37}$$

we can define the **shape parameter** $\Gamma = \Omega_m h$. As we saw above, the shape parameter determines the break point of the power spectrum and the start of the k^{-3} decay. While n_s and Γ_s, shape the current matter power spectrum, the complete normalization is defined by the characteristic amplitude of density fluctuations. The most common parameterisation for this amplitude is given by the fluctuations of the density inside spheres of radius R. The $R-$smoothed density field can be written as

$$\delta_R(\mathbf{x}) = \int d\mathbf{x}' \delta(\mathbf{x}') W_R(|\mathbf{x}-\mathbf{x}'|), \tag{1.38}$$

with $W_R(x)$ being top hat filter of radius R, which fulfills the conditions $\int W_R(x)dx = 1$. The power spectrum for the smoothed density field is then obtained as

$$P_R(k) = |W_R(k)|^2 P_\delta(k), \tag{1.39}$$

and the dispersion of fluctuations on the smoothed density field is then

$$\sigma_R^2 = \langle \delta_R^2(\mathbf{x}) \rangle = \frac{1}{(2\pi)^3} \int dk \left|\hat{W}(k)\right|^2 P_\delta(k). \tag{1.40}$$

where $\hat{W}_R(k)$ is the Fourier transform of the top hat filter. A commonly used value for R is $8h^{-1}$ Mpc, which at the density of matter in the Universe, corresponds to a sphere containing approximately the total matter of a galaxy cluster, so cluster counting directly measures σ_8. σ_8 is commonly referred to as the **amplitude of density fluctuations**. The abundance of structures is also dependent on the matter density Ω_m, as an increased matter component is expected to favor the growth of fluctuations. This yields a degeneracy of Ω_m and σ_8, which manifests as a 'banana shaped' confidence region in parameter constrains analysis of galaxy clustering and weak lensing analysis. This leads to joint constrains for the two parameters expressed by

$$S_8 = \sigma_8(\Omega_m/\Omega_m^*)^\alpha \tag{1.41}$$

where Ω_m^* is a fixed reference value, typically 0.3 and α usually observed to be 0.5.

Non linear growth of structure

The model for linear growth is restrictive in terms of the assumptions made about the components filling space-time and the regime over which they are applicable, with it

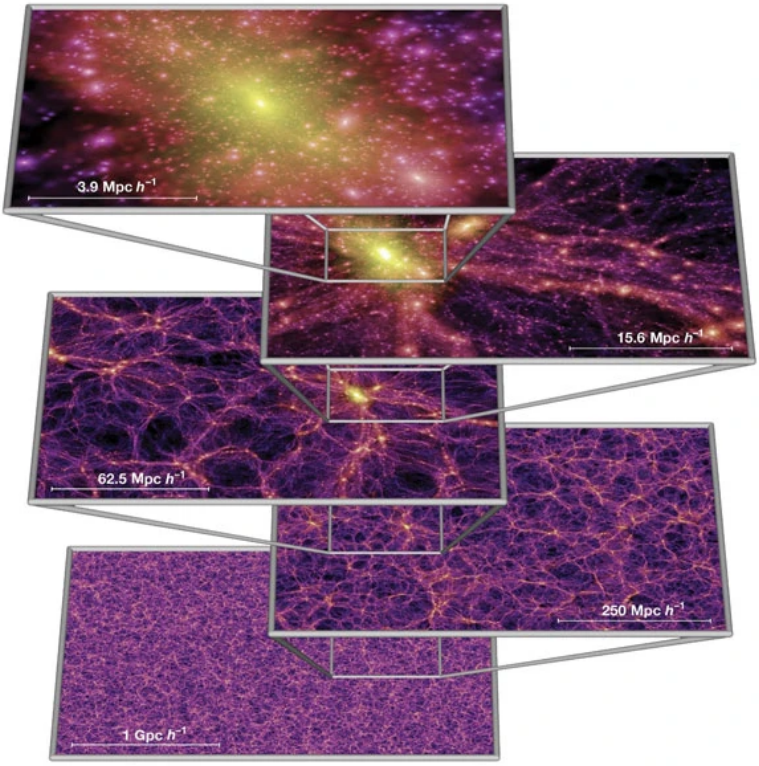

FIGURE 1.3: Zoom sequence centered around a single galaxy cluster halo formed in the Millennium simulation, showing color-coded density of dark matter particles from large linear scales to highly non-linear small scale structures. **Image credit:** (Springel et al., 2005)

only being an adequate approximation for scales where the density contrast δ is much smaller than unity. These assumptions don't hold at small scales (large k values) since the collapse of large structure into compact regions result in large values for the density contrast and the effects of radiation and pressure become more evident. The prediction of the behaviour of the matter power spectrum at small scales requires a numerical analysis in most cases, unless severe simplifications are made. N-body simulations (See fig 1.3; Springel et al., 2005; Alimi et al., 2012) can be used to track the evolution of dark matter beyond the scales of homogeneity, but their precision depends on how accurate are the models employed for density collapse and feedback, and typically come at the expense of large computation times.

The formation of galaxy clusters and smaller structures requires a more detailed description to include the effects observed at small scales, typically incompatible with the more broader picture presented by the linear density evolution model. The **spherical collapse model** assumes that instead of perpetually collapsing, dense concentrations of mass will result in stable concentrations as a consequence of virial equilibrium, with typical density contrasts of about $\delta \sim 200$. The **halo model** (Peacock & Smith, 2000; White, 2001) assumes that all the mass in the Universe is contained within one of such halos, and the statistical properties of the matter at those scales can be predicted from the clustering properties of the halos.

1.2.2 Inflation and origin of primordial fluctuations

The standard model used to describe the evolution of the structure in the Universe through the interaction of mass with space-time has been very successful at predicting a variety of observations, like the Helium to Hydrogen ratio of metal-poor gas or the existence of the CMB. In addition to the expansion caused by Dark Energy and the growth of structure by the gravitational interactions of the components, it characterizes the very early instants ($10^{-35} - 10^{-34}$s after the Big Bang) of the Universe with a fast, exponential expansion. This expansion is driven by the *inflaton scalar field*, possessing primordial quantum fluctuations which grow to become the primordial density fluctuations (Albrecht & Steinhardt, 1982; Linde, 1983; Vázquez et al., 2018).

The **inflation** model not only provides a natural explanation for the origin of primordial density fluctuations, but also explains some observational properties of the large scale structure. The *horizon problem* originates from the fact that several properties of the large scale structure are remarkably constant across the sky (CMB temperature, shape of the power spectrum), even at angular separations larger than the horizon distance at their respective redshifts. This contradicts the fact that at such separation there is no causal connection between them. The *flatness problem* refers to the fact that a flat universe today requires *very* precise initial conditions or fine-tuning. The *monopole problem* originates from the prediction that a large number of heavy, stable "magnetic monopoles" should have been produced in the early Universe. However, magnetic monopoles have never been observed, which can be explained by the rapid decline in number density during the inflationary period. (Guth, 1981; Linde, 1982) The inflationary proposes that shortly after the big bang follows a period of rapid exponential expansion due to the dominance of vacuum energy in the Hubble expansion. This expansion explains why two regions of the sky which would be causally disconnected under the standard expansion theory share so many similar properties. At the same time, this rapid expansions smooths out any initial curvature of the Universe, resulting on a flat geometry almost independent of the initial conditions; now any geometry different from flat requires fine-tuning of the initial conditions.

Inflation also provides an explanation to the origin of the primordial fluctuations which later evolve to become today's large scale structures. Heisenberg's uncertainty principle predicts that on quantum scales the matter distribution cannot be homogeneous. With inflation, these quantum fluctuations are amplified to macroscopic scales which by the end of the inflationary period are large enough to seed the growth of structure in the way we have described in previous sections.

1.3 Observational probes

Perhaps the two most important cosmological discoveries of twentieth century were the observation of the Cosmic Microwave Background (CMB), and the evidence of an accelerated expanding Universe revealed by the recession velocity and luminosity distances of type Ia supernovae. But several other observational tools have been also developed

and employed since then to characterize the cosmological model. In what follows we present a brief description of the historical and physical background and along with a description of the specific cosmological information they provide.

1.3.1 Cosmic Microwave Background

First identified by Penzias & Wilson (1965), the Cosmic Microwave Background (CMB) is the relic radiation remaining from the moment where the first atoms formed, leaving trapped radiation to stream free in all directions from the primordial hot plasma. Since the plasma was in thermodynamic equilibrium, the spectrum of this radiation is that of a black body. Using Sahaś equation, one can find the redshift for plasma temperature of 3,000 K corresponding to a redshift of approximately $z_{\rm rec} \sim 1100$. At this temperature the fraction of high energy photons ($h\nu > 13.6$ eV) is barely enough to keep the plasma ionized. Since then, this radiation has traveled across space and as the Universe expanded, its wavelength increased along with it, shifting the black body temperature to today's value of 2.725K.

While the measured isotropy of the temperature is one the strongest evidences for an inflationary period shortly after the Big-Bang, small fluctuations in the measured black body temperature can be observed at different positions in the sky. These fluctuations are believed to originate due to the interplay between the radiation, which tends to smooth out the density of primordial plasma, and baryons, which attract to each other via gravity. This interplay generated sound waves, or *Acoustic Oscillations* until radiation decoupled from the baryons at the epoch of recombination. Hence, the small anisotropy on the black body temperature at different positions in the sky is an indicator of the primordial structures that gave rise to the large scale structure of the Universe. While its existence was predicted early in the twentieth century, it wasn't until the early 90's when the first detection of anisotropy at the $\Delta T/T \sim 10^{-5}$ level was made by the **CO**smic **B**ackground **E**xplorer mission (COBE; Bennett et al., 1996). The power spectrum of the temperature fluctuations of the CMB is one of the main probes to characterize the ΛCDM model. The angular scale of the peaks of the CMB power spectrum constrain a degenerate relation between the curvature and matter content Ω_m of the Universe. The physical scale of the acoustic oscillations at the time of recombination depends on the matter content, and these physical scales can be related to an angular scale via the expansion history of the Universe, which in turns depends on the Hubble function $H(z)$. For different curvatures, the apparent angular sizes of these structures will change as a consequence of the photons traveling over the null geodesics of the curved space. The location of the first peak at $\ell \sim 220$, first characterized by the BOOMERANG (Crill et al., 2003) and MAXIMA (Jaffe et al., 2001) experiments, is consistent with the angular scales of primordial imprints in the primordial plasma for a universe with no curvature. The relative amplitude of the second peak with respect to the first (and in general, of even numbered with respect to odd numbered peaks) is an indicator of the baryonic density of the primordial plasma, which act by adding gravitational and inertial mass, making the over-densities associated

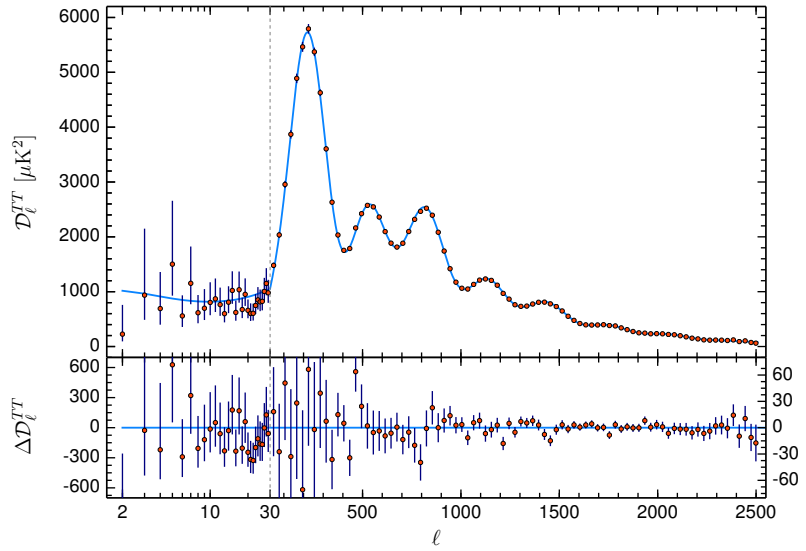

FIGURE 1.4: Angular power spectrum for temperature anisotropy of the CMB black-body radiation, observed by the Planck satellite. Red dots show the measurements with a log scale for $2 \leq \ell \leq 29$, and a linear scale for $\ell > 30$. Light blue solid line shows the predicted ΛCDM spectrum for the best fit values of the cosmological parameters obtained by the Planck collaboration.
Image credits: Planck Collaboration et al. (2020)

to odd peaks even more dense, while leaving the zones of under-density unchanged, which are in turn associated to even peaks.

In addition to the temperature fluctuations the CMB radiation is also polarized, and the spatial distribution of the different polarization measurements is a crucial measurement to constrain the cosmological model. This polarisation, which corresponds to the alignment of the electromagnetic field of photons with respect to the direction of propagation, is caused by inhomogeneities in the plasma at the *Surface of last scattering* which is the last scatter suffered by radiation just before streaming free at the time of recombination. This scattering, when caused by scalar perturbations, only generates a certain type of polarization orientation, called E-modes. Polarization can provide insight into the epoch of inflation as well, immediately after the Big Bang as inflation is believed to produce a background of gravitational waves, which can induce a second type of polarization orientation, called B-modes (Kamionkowski & Kovetz, 2016). While the polarization has already been detected, the observation of B-modes which could be related to primordial gravitational waves is still elusive, and only upper limits have been obtained (Nørgaard-Nielsen, 2018; de la Hoz et al., 2020).

Because of the interference of Earth's atmosphere at microwave wavelengths, most CMB experiments are either conducted with high altitude balloons, in the poles, or using satellite telescopes. The later experiments have provided the most detailed observations of the CMB on their respective eras. After the mentioned COBE was launched by NASA in 1989, followed the Wilkinson Microwave Anisotropy Probe (WMAP; Spergel et al., 2003), which allowed to map the fluctuation power spectrum past the first peak, constraining the curvature of the Universe and its age simultaneously. Currently, the

strongest cosmological parameter constraints from CMB measurements have been provided by the *Planck* satellite (Planck Collaboration et al., 2020), launched by the European Space Agency in 2009. They have provided the most stringent measurements of large scale geometry, combining *Planck* CMB data with low redshift galaxy clustering data from the Sloan Digital Sky Survey (SDSS; Alam et al., 2015). It is important to note that CMB measurements are not exempt from tension with other experiments. In recent years, a significant discrepancy in the measurement of the Hubble constant H_0 with low redshift measurements using strong gravitational lensing and type Ia supernovae has arisen. While *Planck* provides an estimate for today's rate of expansion, $H_0 = 67.4 \pm 0.5 \text{kms}^{-1}$ Mpc^{-1} (Planck Collaboration et al., 2020), the measurement of the redshift-velocity relation for standard candle supernovae yields a higher value of $H_0 = 70 - 75 \text{kms}^{-1}$ Mpc^{-1} (Depending on the particular experiment; see e.g: Phillips et al., 2019; Riess et al., 2019; de Jaeger et al., 2020). While it is remarkable that the estimation made based on the relic radiation emitted when the Universe was only 380,000 years old agrees so well with the estimations based on local structure, the 5-sigma tension between these values is still source of an intense debate and it will be interesting to see what will settle it.

1.3.2 Type Ia Supernovae

Type Ia supernovae are a particular type of supernova which provide valuable information for cosmology studies, based on the relative similarities of their light curves (Intensity as a function of time). They occur in binary systems where a carbon-oxygen burning white dwarf accretes material from a close companion, typically a diffuse red giant. Because of the fixed mass limit at which thermonuclear reactions are able to counteract the effect of gravity, the collapse and subsequent explosion of these stars result in a fairly stable luminosity output and decay, with some small variations coming from star rotation, metallicity, surrounding interstellar medium, and magnification effects by the gravitational lensing effect of the large scale structure. This transforms Type Ia supernovae in a *standard candle*, that is, an object with predictable intrinsic luminosity which is independent of its distance to any observer. Comparison between their observed and absolute magnitudes yields an estimate of its luminosity distance D_L, which can be used to derive the Hubble constant H_0 from equation 1.22. Calibration of their distances at low redshift are made using variable Cepheid stars observed in close galaxies such as Andromeda and the Magellanic clouds. This calibration has been the subject of recent debate on the determination of H_0, which we have already mentioned (Riess et al., 2019; Di Valentino et al., 2021; Efstathiou, 2020).

In addition to the determination of the Hubble parameter, Type Ia supernovae played a fundamental role in the discovery of the accelerated expansion of the Universe in the late 90s, with two seminal papers (Riess et al., 1998; Perlmutter et al., 1999) describing the apparently large luminosity distances for intermediate redshift Ia SN, under the assumption of an Einstein-de Sitter universe, described by $(\Omega_K, \Omega_m) = (1, 0)$. The interpretation of these observations was that the Universe is expanding more rapidly than predicted,

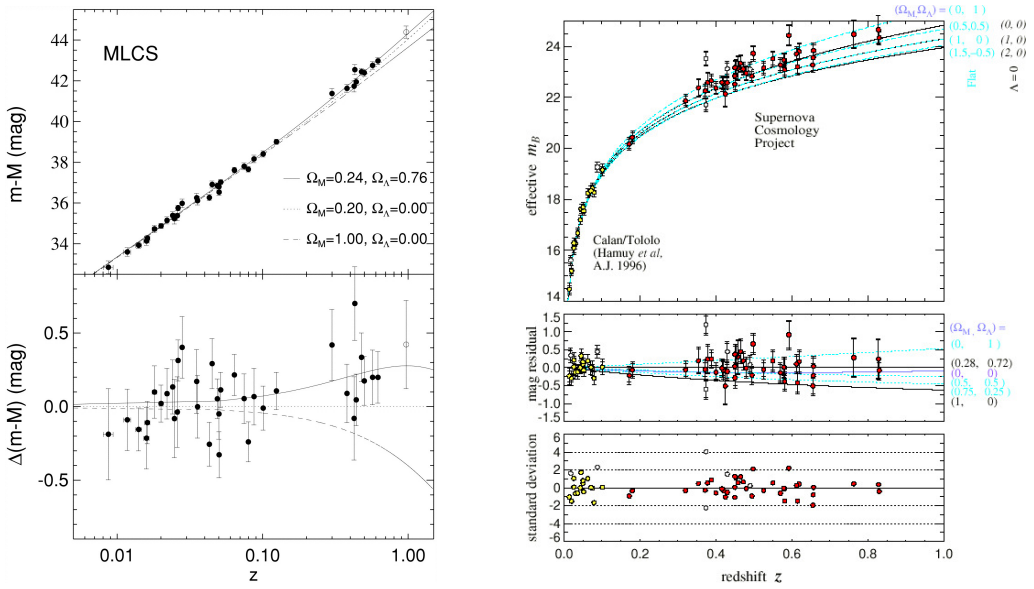

FIGURE 1.5: Left: Low and high-z Sn Ia Hubble diagram and distance modulus residuals with *low*, *high*, and Einstein-DeSitter cosmology from the MLCS method presented in Riess et al. (1998). **Right:** Hubble diagram and magnitude residuals for 42 high-redshift Type Ia SNe from the Supernova Cosmology Project, and 18 low-redshift Type Ia SNe from the Calán/Tololo Supernova Survey presented in Perlmutter et al. (1999)

which is consistent with the presence of a cosmological constant Λ with a relative density $\Omega_\Lambda \sim 0.7$ (See figure 1.5).

1.3.3 Gravitational lensing

General relativity predicts that the path of light emitted by an astrophysical source will follow a null geodesic, which from the perspective of Euclidean geometry, can appear as non-straight lines around massive sources distorting the space-time around them. This phenomena results in the apparent bending of light in the presence of a gravitational field, and can be used as a tool in various astrophysical fields. Here we focus on its application to cosmology, separating two regimes in which gravitational lensing can be detected: strong lensing and weak lensing. We will provide a brief qualitative description and general historical context for both here, and dive deeper into the formalism of weak lensing in chapter 2.

Strong gravitational lensing

Strong lensing refers to the regime where the distortion of space-time by a mass concentration is large enough to cause the light traveling from background sources to deviate significantly from a Euclidean straight line along its null geodesic. This can result in light coming from a source to arrive to an observer via multiple different paths and occurs when the projected surface mass density Σ is greater than a *critical mass* value

$$\Sigma_{\rm cr} = \frac{c^2}{4\pi G}\frac{D_{ds}}{D_d D_s} \tag{1.42}$$

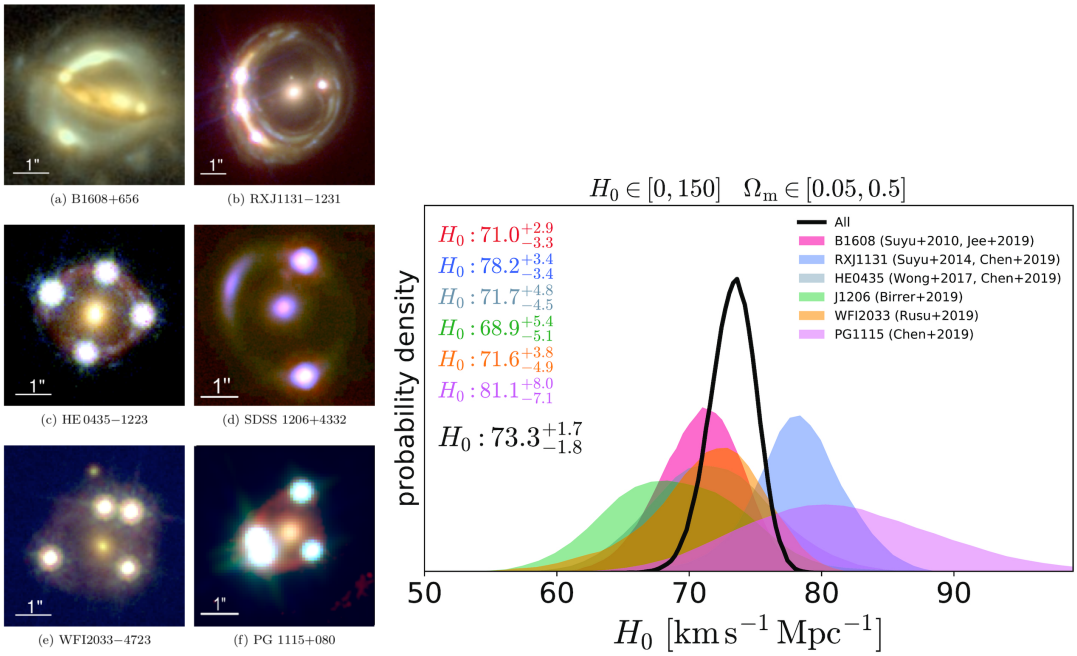

FIGURE 1.6: Left: Sample of lensed quasars employed for the determination of the Hubble constant H_0 by the H0liCOW experiment showing a variety of multiple images and arcs around lensing galaxies. **Right:** Marginalized 1D posterior distributions for the Hubble parameter H_0 obtained from time delay Δt estimations using the 6 strong gravitational lenses shown in the left panel.
Image credit: Wong et al. (2020)

where D_s and D_d are the distances from the observer to the source and lens respectively and $D_{ds} = D_s - D_d$. Depending on the extension of the source this can result in either multiples images for point sources or long elongated arcs for extended sources around the lensing structure. Although first conjectured by Einstein as a consequence of his theory of general relativity, the first detection of multiple images of a single source caused by strong gravitational lensing was made by Walsh et al. (1979), when two point sources at the same redshift and with similar spectrum were observed a massive galaxy inside a cluster. Since then, several other multi-object lens systems have been discovered, along with many arcs-shaped structures around galaxy clusters. A few examples of strongly gravitationally lensed sources and lens systems are shown in figure 1.6.

Strong gravitational lenses are of particular importance in cosmology as they provide a tool to independently measure the expansion rate of the Universe since the time delay Δt between the arrival of light in different images of the strongly lensed sources is proportional to H_0^{-1} (Refsdal, 1964). This time delay has two origins[2] (i) the trajectories followed by photons are geometrically different which make the length of their paths different for each image, causing a *geometrical time delay*, and (ii) light rays travel through a gravitational potential which retards them causing a *gravitational time delay*. The total time delay Δt is the sum of these two effects and is typically measured by monitoring variable sources such as QSOs and cross correlating their time curves.

[2] Actually, a third effect is associated to the increase of traveled distance as a consequence of the expansion of the Universe, but this effects cancels out for both trajectories when computing Δt

As mentioned in section 1.3.1, the estimation of the Hubble parameter has generated an intense debate in recent years as a consequence of the tension between estimations from Planck and the value obtained type Ia Sn. Recent independent estimations using 2 - 20 strong lensed systems have been made (Saha et al., 2006; Paraficz & Hjorth, 2010; Suyu et al., 2013; Wong et al., 2020) and while their results are broadly consistent with estimates of other probes, the uncertainties associated to the modelling of the lenses are still too large to draw conclusions with respect to the H_0 tension.

Weak gravitational lensing

When the projected surface mass density is smaller than the critical mass Σ_{cr} the effect of gravitational lensing becomes less pronounced, only slightly affecting the shapes of background objects without creating multiple images. The fact that it doesn't require a perfect alignment between lensing and background sources and that it is measurable over large fractions of the sky make it a very suitable tool to measure the large scale structure. Moreover, weak lensing constitutes a direct probe of the mass distribution regardless of whether it is dark or baryonic matter. The integrated effect of the large scale structure gravitational potential over the shapes of distant galaxies results in a magnification and shearing of the source shape. The cosmological principle ensures that no preferential direction exists on the Universe, from where it can be assumed that the average ellipticity of unsheared galaxies will average to zero, although in reality the effect of *intrinsic alignments* contributes a significant deviation from this assumption and must be accounted for. The same cannot be said about magnification as there is no first principle to constrain the sizes of galaxies making the shear the main observable of weak lensing experiments. The shear excerpted by the large scale structure is referred to as *cosmic shear*, and the main challenge of detecting this signal resides in the fact that the effect is typically small and difficult to differentiate from the intrinsic shapes and orientations of individual galaxies. In order to measure the effect of cosmic shear the angular correlation of shapes $\xi(\theta)$ is measured at the typical expected angular scales of the large scale structure, using surveys of galaxy shapes and positions with significant surface number densities. These angular correlations are directly related to the power spectrum of the projected 2D mass density, which in turn can be obtained from the 3D matter power spectrum $P_\delta(k)$ via the *limber integral*. This can be exploited to constrain the shape of the power spectrum, up to a degenerate combination of the matter density Ω_m and amplitude of perturbations σ_8 characterized by S_8 (1.41) We will present a detailed description of the weak lensing formalism in chapter 2.

An important aspect of weak lensing surveys is that the relative effect of the large scale structure on galaxy shapes depends on the total distance traveled by light from emission to observation. The determination of the distribution of sources along the line of sight then becomes a key ingredient in weak lensing as it allows to estimate the relative intensity of the cosmic shear on individual sources. A precise determination of the shapes of these distributions is key as biases can significantly alter the resultand inferred cosmological parameters (Ma et al., 2006; Samuroff et al., 2017). Another important use of

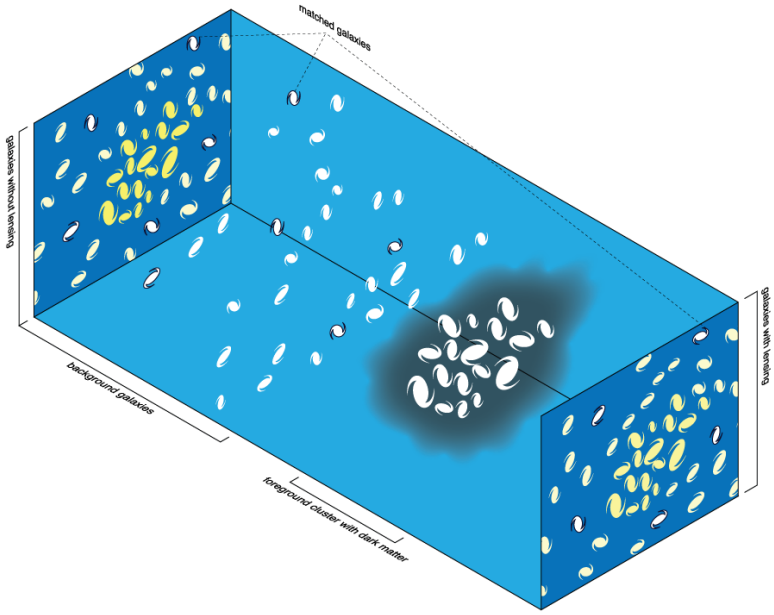

FIGURE 1.7: Graphical representation of the cosmic shear effect caused by the presence of a foreground mass concentration on the images of distant background galaxies. The upper left plane shows the unlensed images of galaxies projected on the plane, while the lower right plane shows the resultant effect of gravitational lensing with background galaxies aligning tangentially to the mass concentration.
Figure credits: Michael Sachs [CC BY-SA 3.0 or GFDL], via Wikimedia Commons

the redshift determination is the use of tomographic binning in which galaxy samples are split into a finite number of redshift slices. This makes it possible to study the 3D matter distribution instead of only a projection of it in 2D, and provides additional information about the temporal evolution of the matter distribution as demonstrated by Hu (1999). Because of the limitations of spectroscopy, weak lensing experiments typically employ multi-band photometry to estimate redshift using a family of techniques called *photometric redshifts* (See e.g.: Abbott et al., 2016; Troxel et al., 2018; Abbott et al., 2018b; Heymans et al., 2021, , etc.).

First detection of the cosmic shear signal was made simultaneously by four independent groups at the beginning of the 2000's (Kaiser et al., 2000; Bacon et al., 2000; Van Waerbeke et al., 2000; Wittman et al., 2000), based on observation of small datasets typically covering less than 1 square degree and a few tens of thousands of galaxies. While the limited sky coverage made it impossible to overcome the effect of cosmic variance, these observations set the foundations for a series of increasingly wider and deeper surveys. The first significant cosmological results were presented by CFHTLens (Heymans et al., 2013) and DES Science Verification (Abbott et al., 2016) (See figure 1.8) with both teams surveying around 150 square degrees and doing a tomographic analysis of the shear angular correlation functions.

Current weak lensing and galaxy clustering are already providing very competitive constraints on the expansion history of the Universe as shaped by Dark Energy, with HSC (Hamana et al., 2020), KiDS (KiDS1000; Heymans et al., 2021) and DES (DES Y1;

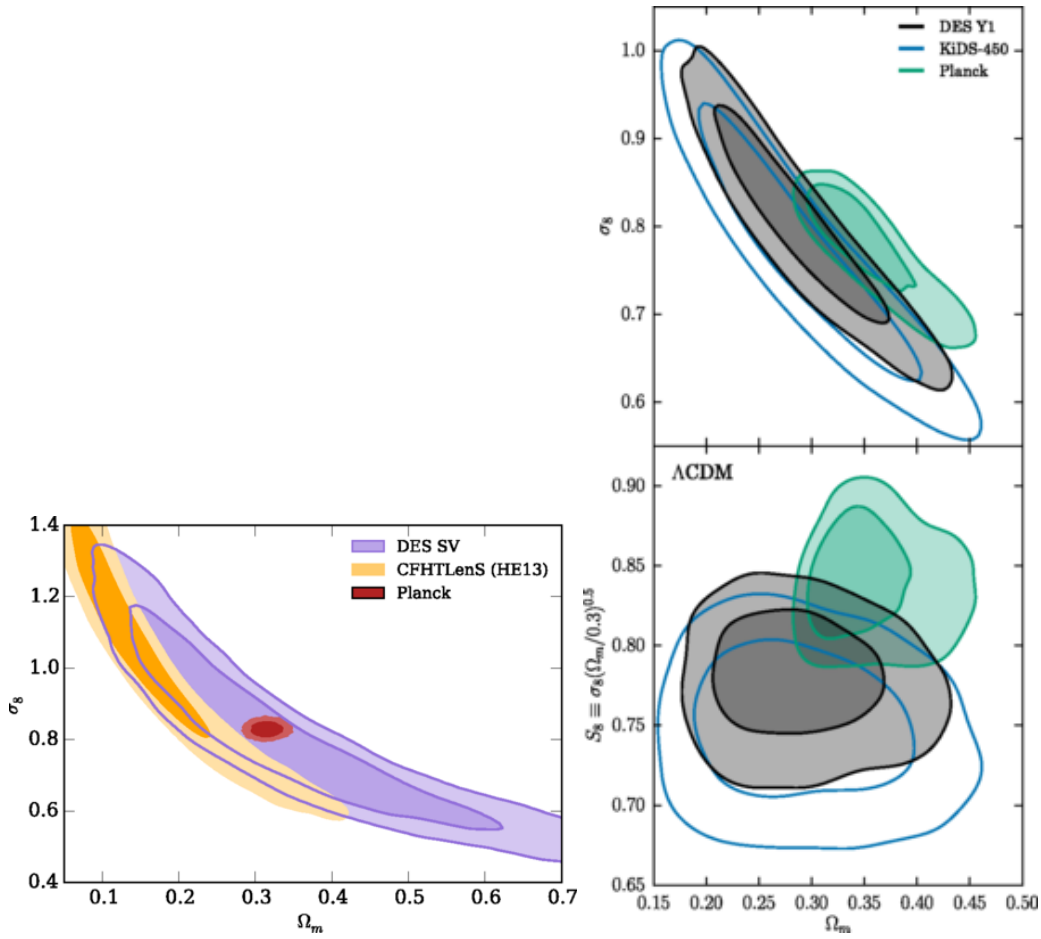

FIGURE 1.8: Left: Comparison of the 2D marginalized posterior of the (Ω_m, σ_8) cosmological parameters obtained from the weak lensing analysis of DES SV (Purple) and CFHTLens (Orange) and from CMB analysis of Planck (Red). The degeneracy direction $S_8 = \sigma_8(\Omega_m/0.3)^{0.5}$ can be clearly seen in the first two cases. **Right:** 2D marginalized posterior of the (Ω_m, S_8) cosmological parameters obtained from the weak lensing analysis of DES Y1 (Black) and KiDS-450 (Blue) and from CMB analysis of Planck (Green). Weak lensing experiments, which probe the late Universe large scale structure yield consistently low estimates of S_8 compared to CMB predictions.
Image credits: Abbott et al. (2016); Troxel et al. (2018)

Troxel et al., 2018) recently providing results based on analysis of 100s to 1000s of square degrees of data and combining it with large spectroscopic surveys to better characterize the color-redshift relations in photometric redshifts. These results not only have sparked significant discussion given the apparent tensions in the inferred cosmological parameters between them (for an in-depth discussion see 3.2.1), but also have provided estimates of the clustering parameter S_8 consistently lower (2-3σ) than those of cosmic microwave experiments like Planck.

Multiple cross-reanalysis of the different datasets reach no consensus on the source of these discrepancies, with some pointing to flat-sky approximations of the Limber integral (Kitching et al., 2016), inaccurate intrinsic alignment models (Kitching et al., 2017) and photometric redshift calibration errors (Choi et al., 2016). Next generation surveys, some of which are described in the next section, will provide orders of magnitude more data and constraining power which will allow weak lensing to achieve estimations of cosmological parameters competitive with those of CMB and perhaps will shed light on the origin of the observed tensions.

1.3.4 Galaxy clustering, galaxy-galaxy lensing and Baryon Acoustic oscillations

The evolution of the dark matter density field leads galaxies to form inside halos by its gravitational interaction with baryonic matter. From this scenario galaxies would be expected to cluster in the same way as dark matter, but early wide galaxy surveys quickly revealed that is not the case. In general it is assumed that the number density of galaxies is different from that of dark matter, but they can be related via a **galaxy bias parameter** b_g such that

$$P_b(k,z) = b_g(k,z)P_\delta(k,z), \tag{1.43}$$

where P_b is the galaxy over-density power spectrum, defined similarly to the density contrast power spectrum. The measurement of galaxy positions and their large scale arrangement can be used to infer properties of the underlying matter density and its evolution as long as we are able to provide an accurate model for galaxy bias. Above the scales assumed for the validity of the cosmological principle ($\gtrsim$ 100 Mpc) there doesn't seem to be a significant scale or redshift dependency of b_g (Desjacques et al., 2018).

Galaxy clustering and galaxy-galaxy lensing

Large optical surveys constrain the clustering signal by computing the two-point correlation function of galaxy positions, $\omega(\theta)$, typically using a sample of large scale structure tracers. DES Y1 used the REDMAGIC algorithm (Rozo et al., 2016) to select bright red sequence galaxies from massive clusters as tracers, while HSC used a photo-z quality metric to select galaxies. Alternatives include the use of precise spectroscopic samples of bright galaxies which are less affected by the limitations of spectroscopy observed in

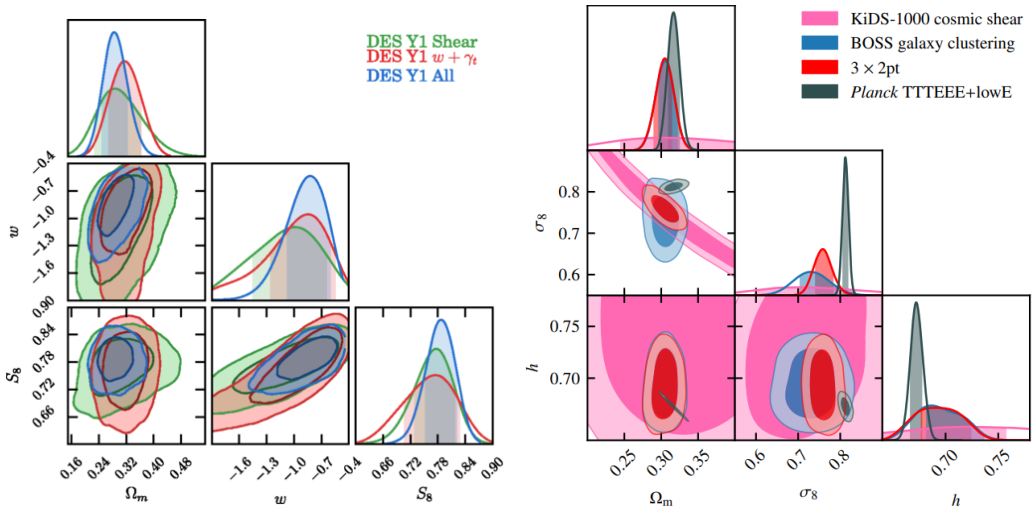

FIGURE 1.9: Cosmological parameter constrains from 3x2pt analysis combining weak lensing, galaxy clustering and galaxy-galaxy lensing correlation functions from DES Y1 (Abbott et al., 2018b) and KiDS-1000 (Heymans et al., 2021), clearly showing the increase in constraining power over pure weak lensing inference.

the faint end. An example of this is the clustering analysis made by the KiDS collaboration (Heymans et al., 2021), which employs spectroscopy from the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al., 2013)

A third two-point correlation can be employed to further constrain the matter distribution, taking advantage of the connection between the position of lensing structures and the cosmic shear observed on distant galaxies. The cross correlation between cosmic shear and galaxy position, γ_t, provides additional information about the connection between the underlying matter density distribution and the baryonic distribution alone. The assumption follows a similar argument to that of cosmic shear in the sense that the effect of a foreground sample of galaxies will exert a deformation on the shapes of background galaxies with the magnitude of the tangential shear being determined by the mass of the foreground distribution.

The three two-point correlations described so far commonly called 3x2pt, $\xi(\theta)$, $\omega(\theta)$, γ_t can be used simultaneously to better constrain the properties of the matter distribution and its evolution or at the very least can serve as a mutual consistency check. Both DES (Abbott et al., 2018b) and KiDS (Heymans et al., 2021) have performed 3x2pt analysis using weak lensing, galaxy clustering and galaxy-galaxy lensing correlation functions, with noticeable increases in constraining power over weak lensing alone (See figure 1.9).

Baryon Acoustic Oscillations

The acoustic oscillations caused by the interplay between radiation pressure and gravitational attraction in the primordial plasma not only left their mark on the temperature distribution of radiation escaping the last scattering surface, but also generated a series

of wave patterns in the primordial baryonic matter distribution, or baryon acoustic oscillations (BAO). By the time matter and radiation decoupled, the imprints of these oscillations left on the matter density field grew to become today structures. The typical size of structures at that time is called the sound horizon, which then becomes the typical distance between today's galaxy clusters and can be used as a standard ruler to constrain the expansion history of the Universe. The role of dark matter and Dark Energy can be revealed by comparing the observed typical distances between the largest structures in the Universe and the predictions based on the size of the sound horizon.

To detect the BAO signal, the positions of galaxies tracing the large scale structure are measured and with it the overabundance of galaxies separated at physical distances. By comparing against a completely random distribution of galaxies in the same area of the sky, the BAO signal should appear as an excess probability of finding galaxy pairs at an angular distance consistent with that of the sound horizon.

First detection of the BAO was made simultaneously by the 2dF Galaxy Redshift Survey (Eisenstein et al., 2005) and SDSS (Cole et al., 2005) where a small excess in number of pairs of galaxies separated by ~ 150 Mpc was observed. While these pioneer observations required the use of spectra to accurately measure the three dimensional positions of galaxies, modern experiments have been able to detect this signal in purely photometric surveys using photometric redshifts (Padmanabhan et al., 2007; Abbott et al., 2019b), or resorting to alternative astrophysical sources such as quasars (Ata et al., 2018) or the Lyman-α forest (Slosar et al., 2013).

1.4 Future cosmological surveys

Time has given us the perspective to identify different milestones achieved by generations of scientists and collaborations aiming at characterising the ΛCDM model. As Albrecht et al. (2006) proposed

> *... we describe dark-energy research in Stages: Stage I represents dark-energy projects that have been completed; Stage II represents ongoing projects relevant to dark-energy; Stage III comprises near-term, medium-cost, currently proposed projects; Stage IV comprises a Large Survey Telescope (LST), and/or the Square Kilometer Array (SKA), and/or a Joint Dark Energy (Space) Mission (JDEM).*

we are currently approaching the end of what, at the time, corresponded Stage III experiments. The Dark Energy Survey finalized its observing runs in January 2019, after six years and 5000 square degrees of sky covered, and the KiDS collaboration has already published preliminary results of cosmic shear measurements of over 1000 square degrees. While there is still plenty of data to be analysed, with DES just starting to publish fully reduced data of observation made up to the third year and three more years in the queue and KiDS set to provide an analysis of around 1500 square degrees when completed, new surveys spectroscopic and photometric surveys with an order of magnitude increase in coverage, data throughput, and expected lifetime, are just around the corner.

Vera C. Rubin Telescope and the Legacy Survey of Space and Time

Perhaps one of most heavily publicised new generation surveys is the Large Synoptic Survey Telescope (LSST; LSST Science Collaboration et al., 2009; Ivezić et al., 2019) which is expected to begin full operations atop the Vera C. Rubin telescope in Cerro Pachón, northern Chile, at the end of 2022. One of the key aspects, aside from the large increase in coverage with respect to Stage III surveys, is that the telescope instruments and survey design are specifically tuned for studies of cosmology (among other few scientific goals). Covering nearly 20,000 square degrees of southern sky in 6 optical bands and mapping all available sky every few nights, not only the survey will reach unprecedented deep magnitudes ($m_r \lesssim 27$) but it will also provide enough cadence to reduce systematics on study of transient events, including Supernovae and time delays in strong gravitational lenses. Similarly to current optical surveys, LSST will aim at constraining the Dark Energy equation of state and characterize the evolution of large scale structure from observations of type Ia supernovae, BAO, weak lensing and galaxy clustering.

The Square Kilometer Array

Optical surveys are limited by their ability to constrain the instrumental and atmospheric response and are typically limited to detect galaxies with $z \lesssim 1$ at competitive signal-to-noise ratios. Radio surveys on the other hand are able to detect sources at higher redshift via their synchrotron emission, and does not have to deal with stochastic PSF effects as the beam size is deterministic and based on the quantifiable limitations of the interferometric setup.

The Square Kilometer Array (SKA; Square Kilometre Array Cosmology Science Working Group et al., 2020) will become the largest interferometric survey of its generation, by mapping the very early Universe in a frequency range from 50 Mhz to 14 Ghz and divided into two stages, SKA1 (Low to mid frequency range, start of operations in 2023) and SKA2 (Full frequency range, 2030). Detectors will form baselines of up to 3,000 km, and will soon be construction in Australia and South Africa. The main cosmological goals of SKA are to provide high redshift galaxy surveys by detecting their HI emission as well as providing maps of extended HI intensity from the intergalactic medium. The possibility to perform cosmological analysis from these galaxy surveys, such as weak lensing and BAO detections have been proposed in Harrison et al. (2016) and Yahya et al. (2015); Bull (2016) respectively. An important aspect that makes such analysis important is the fact that several systematic associated to the determination of shapes and positions of galaxies are uncorrelated across the wavelength spectrum. This allows the use of cross-correlation of the measurements to characterize and alleviate the uncertainties associated.

Euclid & the Nancy Grace Roman space telescope

An important limitation in optical and radio surveys originates from the effect of Earth's atmosphere on the observed properties of astronomical sources which adds up on top of the instrumental response and accounts for a large fraction of the uncertainty associated

to its systematic effects. A way to circumvent this limitation is to carry out surveys using space-based instruments. *Euclid* and the Roman space telescope are two future near-infrared satellite missions currently under construction which will observe $\sim 15,000$ and $\sim 2,000$ square degrees each, in addition to both having spectroscopic capabilities.

Chapter 2

Weak gravitational lensing

The first observational evidence of the effect of gravity on light was obtained on 1919 when Arthur Eddington detected the slight shift in the position of background stars when appearing close to the surface of the Sun, during a total solar eclipse (Dyson et al., 1920). The deflection angle was precisely predicted earlier by Albert Einstein on his General Relativity formalism. Following the discovery of the first extra-galactic gravitational lens, observed as multiple images of quasar QSO 0957+561, lensed by the YGKOW G1 galaxy (Walsh et al., 1979), arcs around dense, massive galaxy clusters have been found all across the sky. While the above examples can be classified as strong lensing effects, weak lensing in turn deals with the regime where the distortions of distant sources, caused by the large scale mass distribution of the Universe, are *very* small, typically a magnitude smaller than the apparent size of the galaxy. These effects cannot be identified by looking at just one galaxy, since the intrinsic shape of the galaxy is not known *a-priori*. One relies on the observation of multiple galaxies and a series of assumptions regarding the intrinsic distributions of shapes and alignments.

Weak lensing has steadily become a very competitive probe of the expansion history of the Universe and its constituents, as it can directly characterise massive structures by measuring the coherent distortions of the observed shapes of distant background galaxies caused by gravitational fields along the line of sight. The statistical significance of the detection of the weak lensing signal needs to be improved by either observing large regions of the sky, increasing the depth of the optical surveys to lower surface brightness, or both simultaneously.

In chapter 1 we learned about the many observational probes to study the large scale structure of the Universe and how each of the main cosmological parameters of the standard model can be constrained from independent observations. We identified the following cosmological parameters:

- **Hubble parameter** H_0
- **Matter density** Ω_m
- **Baryon density** Ω_b
- **Dark Energy density** Ω_Λ

- **Amplitude and spectral index of primordial power spectrum** A_s, n_s
- **RMS amplitude of density fluctuations** σ_8

In section 2.1 we describe the formalism to describe the deflection angles caused by the lensing of distant galaxy images caused by the large scale structure gravitational influence. We then derive the relation between the statistical description of the shear signal and the matter power spectrum, and the concept of tomography. In section 2.2 we describe what observables can be used to generate this statistical description: the observed shapes of distant galaxies, the covariance matrix describing the uncertainty on these measurements, and the distribution of sources along the line of sight. Finally, in section 2.3 we describe the main systematic effects that can affect the estimation of the cosmic shear signal.

2.1 Weak gravitational lensing formalism

2.1.1 The lens equation

We start by deriving a relation to describe the deflection angle from a light ray coming from a source, being lensed by the large scale structure and then received by an observer. A common simplified model is to assume that the lensing structure has dimensions along the line of sight much smaller than the typical source-lens and lens-observer distances. Under this picture the mass distribution at each relative position on the sky $\boldsymbol{\theta}$ is interpreted as a thin sheet characterized by a surface mass density $\Sigma(\boldsymbol{\theta})$ (the *thin lens* approximation), and the light traveling from the distant source is only *lensed* once, when passing through the plane of the lensing mass distribution. Before and after the light is assumed to travel in straight lines (the *Born approximation*).

In the steps shown here we assume a more complete picture in which the lensing structures can span the full line-of-sight range, but employing simplifications which have the same spirit as the above approximations. We start by assuming the shear signal we observe on Earth is a consequence of the distorted light path a photon follows due to the presence of an inhomogeneous gravitational field on small scales which spans the whole space. This space is characterized by the metric (Bartelmann & Schneider, 2001)

$$ds^2 = dt^2 \left(1 + \frac{2\Phi}{c^2}\right) c^2 - a^2(t) \left(1 - \frac{2\Phi}{c^2}\right) d\chi^2. \tag{2.1}$$

~~where $\Phi(t, \chi)$ is the Newtonian potential of small inhomogeneities, expected to be much smaller than c^2.~~ This metric is obtained by solving the Einstein Field equations 1.1 for a stress energy Tensor $T_{\mu\nu}$ described by a small overdensity ρ which varies slowly with time, and satisfies Poisson equation 1.26. The metric described a slight variation of the Minkowski metric of flat spacetime. Light travels along null geodesics, with $ds = 0$, hence

$$\left(1 + \frac{2\Phi}{c^2}\right) c^2 dt^2 = a^2(t) \left(1 - \frac{2\Phi}{c^2}\right) d\chi^2 \tag{2.2}$$

We define the *refraction index* $\eta = c/c'$ which characterizes the relative speed of the light ray moving in this space, with

$$\begin{aligned} c' &= \frac{d\boldsymbol{\chi}}{dt} = ca(t)\sqrt{\frac{1+2\Phi/c^2}{1-2\Phi/c^2}} \\ &\sim ca(t)\left(1+\frac{2\Phi}{c^2}\right) \end{aligned} \tag{2.3}$$

We can obtain the total time for a light ray to travel from point A to point B as

$$\begin{aligned} T &= \int_A^B dt = \frac{1}{c}\int_A^B \eta(\chi)d\chi \\ &= \frac{1}{c}\int_A^B \left(1-\frac{2\Phi}{c^2}\right)a(t)d\chi. \end{aligned} \tag{2.4}$$

where χ is the comoving distance along the light path and we have used the approximation

$$\frac{1}{1+\frac{2\Phi}{c^2}} \sim 1-\frac{2\Phi}{c^2} \tag{2.5}$$

Fermat's principle of least time dictates that the above expression is a minimum, hence $\delta T = 0$. Applying Euler-Lagrange equations (See appendix A) we obtain the total angle deflection by a mass concentration

$$\hat{\boldsymbol{\alpha}} = -\frac{2}{c^2}\int \nabla_\perp^p \Phi d\chi, \tag{2.6}$$

where $\nabla_\perp^p \Phi$ is the gradient of the potential perpendicular to the photon path. Figure 2.1 shows a representation of the deflection of a light ray with respect to a reference light path. The distance between two rays, expressed by $\mathbf{x}$, subtends an angle $\boldsymbol{\theta}$ given by:

$$\mathbf{x}(\chi) = f_K(\chi)\boldsymbol{\theta}, \tag{2.7}$$

where f_K, defined in 1.5, depends on the curvature of the Universe. At the distance χ', a mass concentration with a transverse gradient $\nabla_\perp^p \Phi$ deflects the light ray by an angle $\hat{\boldsymbol{\alpha}}$. In the absence of a deflector, the source at comoving distance χ would have subtended an apparent angle $\boldsymbol{\beta} = \mathbf{x}(\chi)/f_k(\chi)$. The change in separation is then given by:

$$d\mathbf{x} = f_K(\chi-\chi')d\hat{\boldsymbol{\alpha}}. \tag{2.8}$$

Replacing this expression in 2.6 yields the separation between the two rays at any given distance χ, both being subject to deflections along their paths:

$$f_K(\chi)\boldsymbol{\theta} = \frac{2}{c^2}\int_0^\chi d\chi' f_K(\chi-\chi')\left[\nabla_\perp\Phi(\chi') - \nabla_\perp^0\Phi(\chi')\right], \tag{2.9}$$

where the term $\nabla_\perp^0\Phi(\boldsymbol{\chi})$ refers to the perpendicular gravitational potential at the path of the reference ray. If we introduce an approximation by substituting the unperturbed

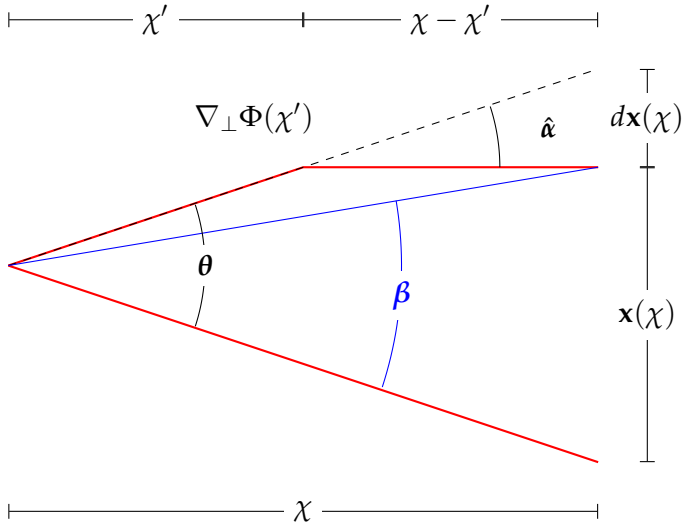

FIGURE 2.1: Two light rays coming from a single source, indicated by the red lines. The top ray is deflected by the presence of a Newtonian potential at comoving distance χ'. The deflection angle $d\hat{\boldsymbol{\alpha}}$ is proportional to the perpendicular gradient of the potential at comoving distance χ'. The angular position $\vec{\theta}$ indicated by the dashed line is then observed instead of the original angular position $\vec{\beta}$ indicated by the blue line.

trajectory $\mathbf{x}_0(\chi) = f_K(\chi)\boldsymbol{\theta}$ on the integral, divide both sides of this equation by $f_k(\chi)$, and write the second term on the right hand side as the deflection angle $\boldsymbol{\alpha}$, then we can write:

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}, \tag{2.10}$$

which is called the lens equation and describes the transformation between the original position of a source and the observed position due to the deflection caused by a gravitational potential. We can now map each point of the source image to the observed image using the Jacobian of the transformation between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, $A = \partial\boldsymbol{\beta}/\partial\boldsymbol{\theta}$, where

$$\begin{aligned} A_{ij} &= \frac{\partial\beta_i}{\partial\theta_j} = \delta_{ij} - \frac{\partial\alpha_i}{\partial\theta_j} \\ &= \delta_{ij} - \frac{2}{c^2}\int_0^\chi d\chi' \frac{f_K(\chi-\chi')f_K(\chi')}{f_K(\chi)} \frac{\partial^2}{\partial x_i \partial x_j}\Phi(f_K(\chi')\boldsymbol{\theta}, \chi') \end{aligned} \tag{2.11}$$

Using

$$\frac{\partial}{\partial\theta_i} = f_K(\chi)\frac{\partial}{\partial x_i},$$

we can define the *lensing potential* (Bartelmann & Schneider, 2001; Scheneider P., 2006):

$$\Psi(\boldsymbol{\theta}, \chi) = \frac{2}{c^2}\int_0^\chi d\chi' \frac{f_K(\chi-\chi')}{f_K(\chi')f_K(\chi)}\Phi(f_K(\chi')\boldsymbol{\theta}, \chi'), \tag{2.12}$$

and the Jacobi matrix A can be written as

$$A_{ij} = \delta_{ij} - \partial_i\partial_j\Psi, \tag{2.13}$$

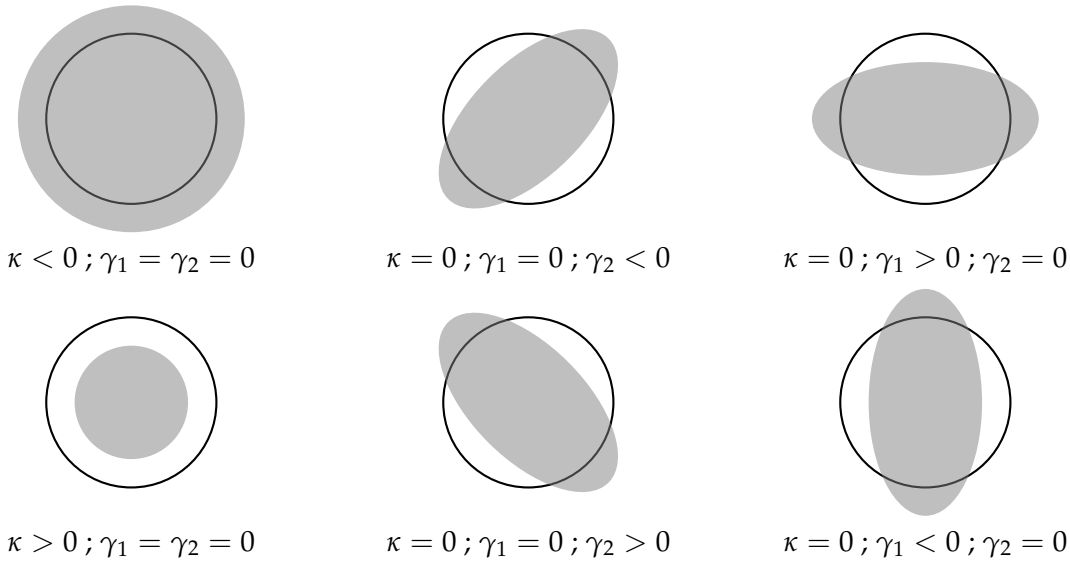

FIGURE 2.2: Illustration of the effects of the distortion matrix A as a function of its components, the convergence κ, and shear (γ_1, γ_2). The value κ indicates the magnification or demagnification of the sources, with magnified sources having a negative convergence value κ. Because of the symmetry of the ellipticity definition, the shear is a *spin-2* field: a $\pi/2$ rotation of the galaxy results in the same shear value.

which can be further parameterized by introducing two new quantities, the convergence κ and the complex shear $\gamma = \gamma_1 + i\gamma_2$, which are obtained as the second derivatives of the lensing potential:

$$\begin{aligned} \kappa &= \frac{1}{2}\left(\partial_1\partial_1 + \partial_2\partial_2\right)\Psi = \frac{1}{2}\nabla^2\Psi \\ \gamma_1 &= \frac{1}{2}\left(\partial_1\partial_1 - \partial_2\partial_2\right)\Psi \\ \gamma_2 &= \partial_1\partial_2\Psi. \end{aligned} \tag{2.14}$$

where δ_i corresponds to the partial derivative with respect to the $i-$th coordinate of the position $\boldsymbol{\theta}$. In this parameterisation, the lensing of the 3D matter distribution can be treated as an equivalent lens plane with deflection potential Ψ, effective surface mass density κ and shear γ. The Jacobian matrix can be rewritten as:

$$A = \begin{bmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{bmatrix} \tag{2.15}$$

Under this parameterization, the images of circular galaxies appear as ellipses whose semi-axes values with respect to the original circular radius depend on the eigenvalues of the matrix, and the orientation is given by the phase φ of the complex shear $\gamma = |\gamma|e^{2i\varphi}$ (See figure 2.2). We can further rewrite the Jacobi matrix as

$$A = (1-\kappa)\begin{bmatrix} 1-g_1 & -g_2 \\ -g_2 & 1+g_1 \end{bmatrix}, \tag{2.16}$$

where the *reduced shear* g is defined as

$$g = \frac{\gamma}{1-\kappa} = g_1 + ig_2 = |g|e^{2i\varphi}$$

The interpretation of this parameterization is that images of distant galaxies are magnified isotropically according to the convergence value $1-\kappa$, and sheared according to the complex value γ. Since, in principle it is not possible to directly infer what is the level of magnification from the galaxy images without any knowledge on the intrinsic size of the source objects, the reduced shear g becomes the main observable of a galaxy distribution, since the orientation of galaxies is expected to be completely random under the cosmological principle of no preferential directions. Nevertheless, magnification can be estimated indirectly by cross-correlating the number density of observed distant objects with the positions of foreground lensing galaxies (See e.g. Garcia-Fernandez et al., 2018), as the magnification effect typically results in slightly increased galaxy detection numbers because of the increase in magnitude of magnified objects. The reduced shear can be estimated from the apparent observed shapes of galaxies, and the spatial correlation of these measurements is the main piece of information used in weak lensing experiments to probe the large scale structure.

2.1.2 Limber's equation

So far we have described the effect of a particular distribution of mass along the line of sight on the observed shapes of distant objects projected around it. Typically we do not know what the mass distribution is and a few measured shapes tell us little about the underlying properties of the Universe, such as the total mass density or the shape of the power spectrum. We are in a middle ground that connects the general observables of weak lensing with the higher level descriptions of the Universe. In what follows we describe how the formalism described so far can be related to the matter power spectrum. Then in the next section we describe how to build up from the fundamental observations to constrain the complete statistical description of the Universe.

We defined the convergence κ as

$$\kappa = \frac{1}{2}\nabla^2\Psi \tag{2.17}$$

where Ψ is the lensing potential defined in equation 2.12. We mentioned above that the convergence is the two-dimensional projection of the mass distribution along the different lines of sight $\boldsymbol{\theta}$

$$\kappa(\boldsymbol{\theta},\chi) = \frac{1}{c^2}\int_0^{\chi} d\chi' \frac{f_K(\chi-\chi')}{f_K(\chi)} f_K(\chi')\nabla^2\Phi. \tag{2.18}$$

In order to find the connection between the convergence field and the matter power spectrum, we can relate κ to the density contrast parameter δ introduced in 1.23. Using Poisson equation 1.26, plus the definition of the density parameter we obtain a linear relation

between the convergence and the density contrast (See appendix B):

$$\kappa(\boldsymbol{\theta},\chi) = \frac{3H_0^2\Omega_m}{2c^2}\int_0^{\chi}\frac{d\chi'}{a(\chi')}W(\chi,\chi')\delta(f_K(\chi')\boldsymbol{\theta},\chi) \tag{2.19}$$

$$W(\chi,\chi') = \frac{f_K(\chi-\chi')}{f_K(\chi)}f_K(\chi') \tag{2.20}$$

where δ is the density contrast parameter which describes the relative matter over-density with respect to the mean matter density, and W is a term to describe the effect of spatial curvature. If we assume that the distribution of mass along the line of sight is given by a probability distribution $n(\chi)$, then we can obtain an effective surface mass density κ_{eff}

$$\begin{aligned}\kappa_{\text{eff}}(\boldsymbol{\theta}) &= \int_0^{\chi_{\max}} n(\chi)\kappa(\boldsymbol{\theta},\chi)d\chi \\ &= \frac{3H_0^2\Omega_m}{2c^2}\int_0^{\chi}\frac{d\chi}{a(\chi)}g(\chi)f_K(\chi)\delta(f_K(\chi')\boldsymbol{\theta},\chi)\end{aligned} \tag{2.21}$$

where

$$g(\chi) = \int_{\chi}^{\chi_{\max}} d\chi' n(\chi)\frac{f_K(\chi'-\chi)}{f_K(\chi')} \tag{2.22}$$

is called the *lensing efficiency function*.

As we mentioned in chapter 1, we can only predict the statistical behaviour of the matter density, encoded in the matter power spectrum $P_\delta(k)$. What we have so far is a collapsed 2D distribution from the 3D density field through an efficiency function kernel $g(\chi)$ which can only be expected to be estimated from averaging over many lines of sight, so we can't reconstruct the 3D matter density field. What we can do is predict its statistical properties from the statistical properties of the collapsed effective convergence, both encoded by their two-point correlation functions. Following the If $\delta(f_K(\chi')\boldsymbol{\theta},\chi)$ is a homogeneous and isotropic Gaussian random field, then any projection of the form

$$\kappa(\boldsymbol{\theta}) = \int d\chi q(\chi)\delta(f_K(\chi)\boldsymbol{\theta},\chi) \tag{2.23}$$

is also an homogeneous and isotropic Gaussian field, this time only in 2D. The angular auto-correlation of $\kappa(\boldsymbol{\theta})$ (see 1.30) will only be

$$C_\kappa = \langle\kappa(\boldsymbol{\varphi}_1)\kappa(\boldsymbol{\varphi}_2)\rangle = C_\kappa(|\boldsymbol{\varphi}_2-\boldsymbol{\varphi}_1|) \tag{2.24}$$

and its power spectrum will be the Fourier transform of the correlation function

$$P_\kappa(\ell) = \int\int d\varphi_1 d\varphi_2 e^{i\varphi_1\ell}e^{i\varphi_2\ell}\langle\kappa(\varphi_1)\kappa(\varphi_2)\rangle \tag{2.25}$$

where we use ℓ as the wavenumber instead of k to avoid confusion with κ. If we expand this using 2.23 (Bartelmann & Schneider, 2001; Scheneider P., 2006) we obtain

$$P_\kappa(\ell) = \int\int d\varphi_1 d\varphi_2 e^{i\varphi_1\ell}e^{i\varphi_2\ell}\int d\chi q^2(\chi)\langle\delta(f_K(\chi)\varphi_1)\delta(f_K(\chi)\varphi_2)\rangle \tag{2.26}$$

Taking the change of variables $\theta_i = f_K(\chi)\varphi_i$ and re-arranging the integrands

$$P_\kappa(\ell) = \int d\chi \frac{q^2(\chi)}{f_K^2(\chi)} \int\int d\theta_1 d\theta_2 e^{i\theta_1 \frac{\ell}{f_K(\chi)}} e^{i\theta_2 \frac{\ell}{f_K(\chi)}} \langle \delta(\theta_1, \chi)\delta(\theta_2, \chi)\rangle \tag{2.27}$$

Here we note that the second integral on the right hand side is the definition of the Fourier transform of the density contrast correlation function, but for a scaled wavenumber $\ell/f_K(\chi)$. Replacing $q(\chi)$ with the corresponding values from 2.21 we find

$$P_\kappa(\ell) = \left(\frac{3H_0^2\Omega_m}{2c^2}\right)^2 \int d\chi \frac{g^2(\chi)}{a^2(\chi)} P_\delta\left(\frac{\ell}{f_K(\chi)}, \chi\right) \tag{2.28}$$

where we have found that the convergence power spectrum is directly related to the density matter power spectrum $P_\delta(k)$. This result, which can be derived from the approximations and formulation presented in Limber (1953), is typically called the *Limber integral* and tells us that we can completely describe the 3D matter power spectrum from only knowing the statistical properties of the effective convergence, encoded on its power spectrum $P_\kappa(k)$. In the next section we will see that the convergence power spectrum can be directly obtained from the shear field.

2.1.3 Shear correlation functions

We can write the effective convergence $\kappa(\boldsymbol{\theta})$ as an inverse of its Fourier transform

$$\kappa(\boldsymbol{\theta}) = \frac{1}{(2\pi)^2} \int d\boldsymbol{\ell} \hat{\kappa}(\boldsymbol{\ell}) e^{-i\boldsymbol{\ell}\boldsymbol{\theta}} \tag{2.29}$$

and in a similar way define the Fourier transform of the lensing potential, $\hat{\Psi}(\boldsymbol{\ell})$. In general, the correlation function of the Fourier transform $\hat{\kappa}(\boldsymbol{\ell})$ can be written in terms of its power spectrum $P_\kappa(\ell)$ as

$$\langle \hat{\kappa}(\boldsymbol{\ell})\hat{\kappa}^*(\boldsymbol{\ell}')\rangle = (2\pi)^2 \delta_D(\boldsymbol{\ell} - \boldsymbol{\ell}') P_\kappa(\ell) \tag{2.30}$$

where δ_D is the Dirac delta function. From the definition of the convergence in terms of the lensing potential (2.14) we can write:

$$-|\boldsymbol{\ell}|^2 \hat{\Psi}(\ell) = 2\hat{\kappa}(\boldsymbol{\ell}) \tag{2.31}$$

where we have used the fact that differentiation in real space can be replaced by a $-i\ell_i$ multiplication in harmonic space. Using the definition of the shear components γ_i in 2.14 and replacing in 2.31 we find that the Fourier transform of the shear field $\gamma(\boldsymbol{\theta})$ can be written as a function of the transform of the convergence κ

$$\begin{aligned} \hat{\gamma}(\boldsymbol{\ell}) &= \left(\frac{\ell_1^2 - \ell_2^2 + 2i\ell_1\ell_2}{|\boldsymbol{\ell}|^2}\right)\hat{\kappa}(\boldsymbol{\ell}) \\ &= e^{2i\beta}\hat{\kappa}(\boldsymbol{\ell}) \end{aligned} \tag{2.32}$$

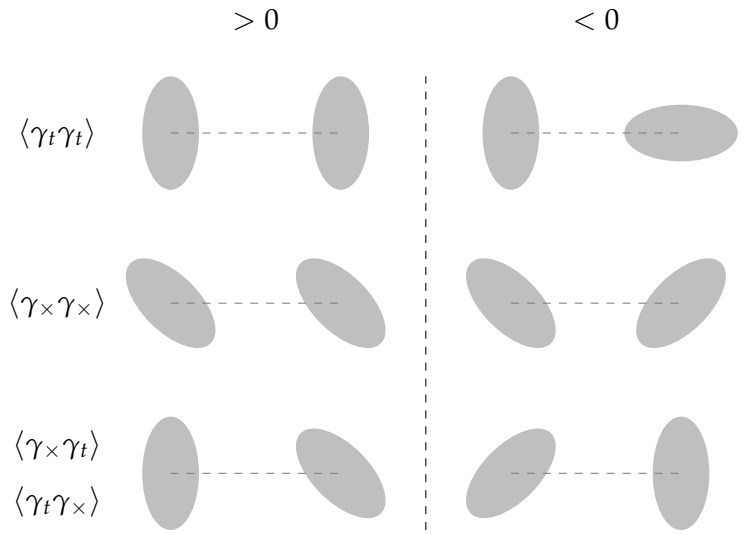

FIGURE 2.3: Shear pairs with different orientations and the resulting correlation functions $\langle\gamma_t\gamma_t\rangle$, $\langle\gamma_\times\gamma_\times\rangle$ and $\langle\gamma_t\gamma_\times\rangle$. While the individual $\gamma = \gamma_1 + i\gamma_2$ values for each object change depending on the orientation of the reference frame, the resulting cross- and tangential-shears $\gamma_\times, \gamma_t$ change accordingly because of the orientation of the pair angle φ ($\varphi = 0$ in this figure). Both changes result in an invariant set of correlation functions $\xi_\pm$.

where β is the polar angle of the wavenumber vector $\boldsymbol{\ell}$. Using equation 2.30 and 2.32 we find that

$$\langle\hat{\gamma}(\boldsymbol{\ell})\hat{\gamma}^*(\boldsymbol{\ell}')\rangle = (2\pi)^2\delta_D(\boldsymbol{\ell} - \boldsymbol{\ell}')P_\kappa(\ell) \tag{2.33}$$

This tell us that the power spectrum of the shear field $\gamma(\boldsymbol{\theta})$ is exactly equivalent to that of the convergence field $\kappa(\boldsymbol{\theta})$. Instead to trying to estimate the convergence field power spectrum, we can obtain it from the shear which can be estimated easily from the shapes of galaxies.

Two particular correlation functions of the shear can be used to estimate P_κ, which are easily measurable from a large catalog of observed galaxy shapes and their positions. If we consider two points characterized by $\gamma^i = \gamma_1^i + i\gamma_2^i$ each, and separated by an angular distance $\boldsymbol{\theta}$ along the direction angle φ, we can define the tangential and cross-component of the shear at these positions as

$$\gamma_t = -\mathcal{R}e(\gamma e^{2i\varphi}) \qquad \gamma_\times = -\mathcal{I}m(\gamma e^{2i\varphi}) \tag{2.34}$$

from where we can define two correlation functions[1]

$$\xi_\pm = \langle\gamma_t(\theta)\gamma_t(\theta)\rangle \pm \langle\gamma_\times(\theta)\gamma_\times(\theta)\rangle \tag{2.35}$$

[1] A third cross-correlation $\xi_\times(\theta) = \langle\gamma_t\gamma_\times\rangle$ vanishes due to the symmetry of the shear field under a mirror transformation

If we replace $\gamma(\boldsymbol{\theta})$ in terms of its Fourier transform $\hat{\gamma}(\boldsymbol{\theta})$ and use 2.33 we find (Bartelmann & Schneider, 2001; Scheneider P., 2006)

$$\xi_+(\theta) = \int \frac{d\ell\ell}{2\pi} J_0(\ell\theta) P_\kappa(\ell) \tag{2.36}$$

$$\xi_-(\theta) = \int \frac{d\ell\ell}{2\pi} J_4(\ell\theta) P_\kappa(\ell) \tag{2.37}$$

where the first kind Bessel functions of order n are used

$$J_n(\ell\theta) = \frac{1}{i^n \pi} \int_0^\pi e^{i\ell\theta\cos\varphi} \cos(n\varphi) d\varphi \tag{2.38}$$

Owing to the orthonormality of the Bessel functions these two equations can be inverted to obtain

$$P_\kappa(\ell) = 2\pi \int d\theta \xi_+(\theta) J_0(\ell\theta) \tag{2.39}$$

$$P_\kappa(\ell) = 2\pi \int d\theta \xi_-(\theta) J_4(\ell\theta) \tag{2.40}$$

At last, we have found a set of quantities, usually obtainable directly from data, which relate directly to the convergence power spectrum $P_\kappa(\ell)$, which in turn can help us find the matter power spectrum $P_\delta(\ell)$. In the next sections we describe how the shear correlation functions can be estimated from data, and what are the observational products necessary in a weak lensing experiment to perform a full cosmological analysis.

2.1.4 E- and B- modes

We saw that the lensing potential is related to the lensing potential via the Poisson equation

$$\kappa = \frac{1}{2}\nabla^2\Psi \tag{2.41}$$

We can then define the field vector $\mathbf{u}$ such that

$$\mathbf{u} = \nabla\kappa \tag{2.42}$$

which, using the relations 2.14, can be written as

$$\mathbf{u} = \begin{bmatrix} \partial_1\gamma_1 + \partial_2\gamma_2 \\ \partial_1\gamma_2 + \partial_2\gamma_1 \end{bmatrix} \tag{2.43}$$

By definition, the curl $\nabla \times \mathbf{u}$ vanishes which introduces second order derivatives constrains for γ_1 and γ_2. The constrains imposed on the shear and the fact that both components of the complex shear γ are can be obtained from a single lensing potential, implies that the two components are not completely independent from each other. This leads to only certain combinations of (γ_1, γ_2) being possible in the presence of a lensing structure (See figure 2.4). If the field fulfills this conditions completely, then it is called and E-mode field.

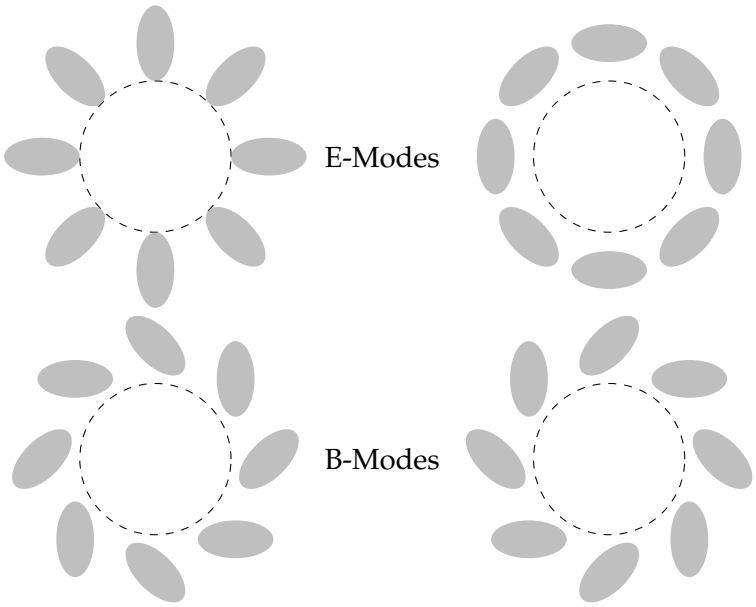

FIGURE 2.4: E- and B- modes of the shear. Upper row shows the E-mode pattern coming from either a spherical over density (left) or under density (right), where tangential and perpendicular alignments with respect to the position of the density distribution. B-modes are 45 degree rotation of E-modes which cannot occur due to the presence of a lensing structure.

In practice, the observed shear field can have non-vanishing curl, $\nabla \times \mathbf{u}$, or B-mode which can be associated to several effects:

- Higher order terms neglected from the Born approximation.
- Higher order terms arising from the shear to reduced shear transformation ($\gamma \rightarrow g$).
- Selection biases in galaxy samples.
- Shape-shape and shear-shape correlations.
- Image quality and and errors in the shape analysis.

B-modes have been consistently used to test systematics on weak lensing surveys, as they are expected to vanish because of the parity of the shear field under the Born approximation. With the advent of larger surveys, the non detection of this signal must be treated with care as it is known that physical effects can produce a non systematic B-mode.

2.1.5 Tomography

We recall from equation 2.21 that the effective convergence $\kappa_{\text{eff}}(\boldsymbol{\theta})$ can be interpreted as the projected surface mass density of the underlying density field, weighted by the lens efficiency from equation 2.22. The lensing efficiency determines the relative contribution to the total shear field of the different lensed sources at different comoving coordinates, given by the distribution $n(\chi)d\chi$ along with the effect of geometry via the term $f_K(\chi)$. We can obtain the distribution of lensed sources by observing a large sample of galaxies and

obtaining their redshifts, which we can relate easily to their comoving coordinates. Considering that the effective convergence is a 2D projection of the mass distribution along the line of sight, most of the three dimensional effects, and therefore the time evolution of the underlying fields parameters is lost due to this projection. The introduction of tomographic binning allows more precise constraining of the cosmological parameters and their evolution with cosmic time. If we define a binning of the comoving distance range $[0, \chi_{\max}]$ composed of N_b bins, we can compute the lensing efficiency for each bin as

$$W_i(\chi) = \int_0^{\chi_{lim}} d\chi' n_i(\chi') \frac{f_K(\chi' - \chi)}{f_K(\chi')} \tag{2.44}$$

where $n_i(\chi)$ is the distribution of sources along the line of sight for that particular bin. Replacing in 2.19 we obtain the projected convergence for that redshift bin, and using 2.28 we can obtain $N_z(N_z - 1)/2$ different convergence power spectra,

$$P_{\kappa,ij} = \frac{9H_0^4\Omega_m^2}{4c^2} \int_0^{\chi_{lim}} d\chi' \frac{W_i(\chi')W_j(\chi')}{a^2(\chi')} P_\delta \left(\frac{\ell}{f_K(\chi')}, \chi' \right)$$

The lensing efficiency is a broad function of the redshift, and in addition, photometric redshift error broadens it even more, thus the different power spectra are not independent from each other. Also, large scale structure contributes to the overlapping and correlation of the spectra, since the structure can extend over more than a redshift bin and low redshift structure affects the observed shear in larger redshift galaxies. While in principle, one could be tempted to choose a large number of redshift bins, uncertainties in photometric redshifts wash away any possible fine details in the cosmological parameter evolution that could possible arise of such a binning. Fine tuning is necessary to find the appropriate number of bins in terms of the expected uncertainties in the obtained parameters of the survey, but it has been shown (Hu, 1999; Simon et al., 2004) that a small number of bins $N_b \lesssim 10$ is sufficient to improve the cosmological constrains over analysis with no binning at all.

2.2 Weak lensing observables

It is worth mentioning that alternative approaches exist to estimate the convergence and matter density power spectrum by directly measuring the shear power spectrum (Bond et al., 1998; Seljak, 1998; Hu & White, 2001; Köhlinger et al., 2017). These techniques measure the average shear inside sky 'pixels' and compute the Fourier transform of the shear field to estimate the convergence power spectrum (which we saw in the last section is equivalent to the shear power spectrum). Since surveys typically don't cover the full sky and mask or weight regions differently to account for several observational conditions, these methods require a careful description of the footprint window function.

Nevertheless, the two-point shear correlation functions in real-space (equations 2.36) constitute the main observables of many modern weak lensing experiments because of

the simplicity of their estimation. Assuming a catalog of observed shears at several positions $\boldsymbol{\theta}^i$, an estimator for the shear correlation functions can be written as:

$$\hat{\xi}_{\pm}(\theta) = \frac{\sum_i \sum_j w_i w_j \left[\gamma_t(\boldsymbol{\theta}^i)\gamma_t(\boldsymbol{\theta}^j) \pm \gamma_\times(\boldsymbol{\theta}^i)\gamma_\times(\boldsymbol{\theta}^j)\right]}{\sum_i \sum_j w_i w_j}$$

where w represent the relative weights assigned to each individual shear measurement, and the sum is done over all points separated by an angular distance $\theta = |\boldsymbol{\theta}^i - \boldsymbol{\theta}^j|$. In reality it is virtually impossible to employ this estimator since the angular distance between pairs of points is a continuous value which will never repeat for any two pair of points, rendering the average useless. The usual approach is to bin the complete angular range and compute the correlation functions for all points separated by an angular distance $\theta \in [\theta_k, \theta_k + \Delta\theta)$. Usually this estimator is very computationally intensive as the number of pairs and cross- and tangential-shears to be computed is very large. A commonly used approximation is to perform a partition of the observed positions and their associated shears using clustering algorithms such as *k*d-trees (Schneider et al., 2002; Jarvis, 2015), where the mean shear inside compact groups of points are used in the estimator instead of individual points. The fundamental required measurements are the positions and shapes of distant galaxies from where we estimate the cosmic shear based on the cosmological principle.

2.2.1 Shear

Assuming the intensity of a galaxy as a function of position $\boldsymbol{\theta}$, can be described by a surface density profile $I(\boldsymbol{\theta})$, we can find the coordinates of the centroid (θ_1^c, θ_2^c) of the intensity distribution as the first moment of the intensity

$$\theta_i^c = \frac{\int d\boldsymbol{\theta} I(\boldsymbol{\theta})\theta_i}{\int d\boldsymbol{\theta} I(\boldsymbol{\theta})} \tag{2.45}$$

Using this centroid we can define the second order moments

$$Q_{ij} = \frac{\int I(\boldsymbol{\theta})(\theta_i^c - \theta_i)(\theta_j^c - \theta_j)}{\int d\boldsymbol{\theta} I(\boldsymbol{\theta})} \tag{2.46}$$

where i and j are the two coordinates of the position vector $\boldsymbol{\theta}$. The most widely used definition of *ellipticity* ε is given by

$$\varepsilon = \frac{Q_{11} - Q_{22} + 2iQ_{12}}{Q_{11} + Q_{22} + \sqrt{Q_{11}Q_{22} - 2Q_{12}}} \tag{2.47}$$

It can be shown (Schneider & Seitz, 1995) that the Jacobi matrix defined in 2.16 transforms the unlensed profile $I(\boldsymbol{\theta})$ of a source such that the lensed ellipticity $\varepsilon_{\text{lensed}}$ can be written as

$$\varepsilon_{\text{obs}} = \frac{\varepsilon_{\text{source}} + g}{1 + \varepsilon_{\text{source}} g^*} \tag{2.48}$$

where $\varepsilon_{\text{source}}$ is the ellipticity of the unlensed profile, and g, g^* are the reduced shear and its complex conjugate, respectively. If we average over multiple ellipticity measurements, the cosmological principle which tells that there is no preferential direction results in $\langle \varepsilon_{\text{source}} \rangle = 0$. Then, the average observed ellipticity will be $\langle \varepsilon_{\text{obs}} \rangle = g$, an unbiased estimator of the reduced shear g.

Two families of methods can be identified from the literature: estimation of *image moments*, and *parametric fitting*. Image moment methods directly compute the image quadrupoles of equation 2.46 in real (e.g. KSB+; Kaiser et al., 1995; Luppino & Kaiser, 1997; Hoekstra et al., 1998) or Fourier space (Hosseini & Bethge, 2009), under the assumption that the observed shape is a combination of the effects described in figure 2.5. This allows the observed ellipticity to be decomposed into individual components, including those of the PSF and the shear.

Parametric model fitting techniques attempt to reproduce the observed intensity distribution $I(\boldsymbol{\theta})$ using sums of elliptical parametric models, such as the Sérsic profile. Instead of deconvolving a post-PSF galaxy image, the model is generated *on top of the atmosphere*, and then convolved by the previously estimated PSF to be compared to the noisy image. The main advantage of model fitting methods is that the ellipticity of these parametric models is well known. Other similar approaches such as *Shapelets* (Refregier & Bacon, 2003) decompose the galaxy image into a finite sum of polynomial linearly independent basis 2D functions. Last generation surveys methods mostly employ model fitting algorithms (NGMIX, IM3SHAPE, LENSFIT; Sheldon, 2015; Zuntz et al., 2013; Miller et al., 2007, respectively), with some more sophisticated implementations (e.g. METACALIBRATION Huff & Mandelbaum, 2017; Sheldon et al., 2020) being able to perform a full photometric analysis plus estimation and calibration of the shear.

2.2.2 Covariance matrix of the angular correlation functions

Weak lensing observables, either two-point shear correlation functions at different angle bins θ_k or the lensing power spectrum on each tomographic bin are correlated due both physical and systematic effects: the distribution of large scale structure and the relative density contrast in terms of galaxies per unit volume can impact the detectability of structure due to large Fourier modes, correlating low and high angular scales. Also, the extent of structure in the radial direction can impact the correlation of low redshift structures with the shear of high redshift sources (See 2.3.1). Finally, the assumption of a Gaussian shear field for the Limber integral is not perfect, owing to the non linear evolution of the density field at small scales. Photometric redshift errors and the highly degenerate color-redshift relation can also lead to galaxies being placed into incorrect bins or result in with wide lensing efficiency functions with large overlaps between tomographic bins. This can introduce spurious correlations making the observed two-point correlations functions not completely independent from each other. The usual assumption made in the process of parameter inference is that the distribution of measurement errors of these observables is Gaussian, and along with its correlation, it can be described by a covariance

matrix $\mathbf{C}$

$$\mathbf{C}_{ij} = \mathrm{cov}(x_i, x_j) \tag{2.49}$$

where x_i are the elements of the vector of observables, or datavector $\mathbf{x}$, typically two-point shear correlation functions estimated at angular bins θ_k. The importance of this covariance matrix is that it allows us to compare the observed datavector to a theoretical value predicted by a set of cosmological parameters ϑ via the likelihood function $\mathcal{L}$

$$\mathcal{L}(\mathbf{x}|\vartheta, M) = \frac{1}{(2\pi)^{m/2}|\mathbf{C}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_t)^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{x}_t)\right], \tag{2.50}$$

In chapter 4 we will formally introduce the concept of Bayesian inference and how this likelihood function is employed to estimate the confidence regions of parameter space for our models, given the observations made in the cosmic shear pipeline. Several approaches can be taken to estimate the covariance matrix both from data or simulations. Analytical models (Joachimi et al., 2008; Schneider et al., 2002) estimate the covariance matrix directly from the definition of second order statistics, under the assumption of the shear being a Gaussian field. Depending on the complexity of the model, these methods can provide an inverse covariance matrix suitable for cosmological inference, as long as they are able to account for the non-Gaussian effects such as the correlation of modes of the density field inside the survey footprint which could be correlated with super-survey modes, cosmic variance and shot noise.

A widely used family of methods for covariance matrix estimation rely on the realization of multiple N-body simulations where $\xi_\pm$ are obtained from projected particle densities in the redshift range of the simulations. Hartlap et al. (2007) showed that the main requirement for this procedure is to have a number of simulations, p, larger than the vector of correlations $\xi_\pm$, otherwise the estimated matrix $\mathbf{C}$ cannot be inverted. But even if the matrix is not singular, the use of an estimator for equation 2.49, while being unbiased, could lead to a biased estimation of the inverse matrix $\mathbf{C}^{-1}$. Anderson (2003) showed that an inverted estimator $\hat{\mathbf{C}}^{-1}$ can be biased by a factor $(p-1)/(n-p-2)$ where p is the number of realizations and n is the length of the data vector $\xi_\pm$ we defined before; correcting factor for this bias is known as the Anderson-Hartlap correction.

Jackknife methods are a simple estimation of $\mathbf{C}$ where in a similar fashion to the estimation using N-body simulations, the terms of the covariance matrix are estimated taking the average over many realizations of $\xi_\pm$ over 'incomplete survey samples'. A simple way to implement this is to measure the correlation function in sub-samples of the total survey footprint. In addition to the fact that the estimate of C^{-1} is also biased and must be corrected using the Anderson-Hartlap correction factor, this method will be also biased by any general bias associated to the estimation of the two-point correlation functions.

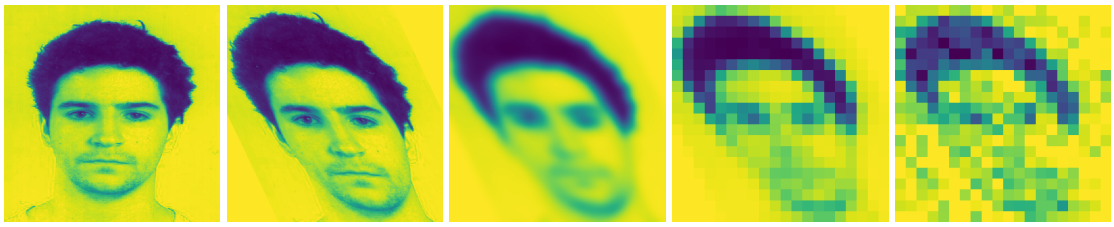

FIGURE 2.5: Illustration of the forward problem. From left to right the panels show how the original image is sheared by the cosmic shear, blurred by the point spread function of the atmosphere and instrument response, pixelised by the detectors and noised by multiple sources including thermal effects on the telescope and readout noise.

2.2.3 Photometry and Redshifts

Estimation of the matter density power spectrum using Limber integral requires precise knowledge of the lensing efficiency kernel defined in equation 2.22, which in turns depends on the line of sight distribution of lensed sources, $n(\chi)$ along the comoving distance range $[0, \chi_{\max}]$. We typically estimate $n(\chi)$ by computing the line of sight distribution as a function of redshift, which as we saw in 1 is an indicator of distance. Because of the limitations of spectroscopy, both in terms of budget and time constraints and completeness, redshifts are estimated using a family of techniques called photometric redshifts, which often provide less accurate point estimates of the distributions, but are more complete down to fainter magnitudes, and reasonably accurate for large ensembles of sources. These estimate the redshift using measurements of the incident galaxy flux through few narrow and/or filter bands

$$f_b = \int d\lambda W_b(\lambda) F_b(\lambda) I(\lambda) \tag{2.51}$$

where W_b and F_b represent the instrument response and filter passband, and I is the spectral energy distribution of the source, all as a function of wavelength λ. We will discuss the formalism and literature on this particular measurement in chapter 3.

2.3 Systematics and their effects on cosmology

The *forward process* described in figure 2.5 and the photometric redshift approach describe in a very simplified way the main challenge faced by cosmic shear experiments: detecting the effect of cosmic shear from small, noisy, pixelised images of distant galaxies and estimating the redshift from only a handful of photometric measurements. In addition to that, large cosmological surveys are tasked with estimating these quantities for several million objects, which is only accomplished by developing complex computational tools to perform these measurements accurately and efficiently. A significant fraction of the time and effort spent on these large experiments is devoted to the testing, validation and characterisation of these methods and their associated biases or uncertainties.

Through this thesis we will frequently use the concepts of **systematic bias** and **uncertainty**. A **bias** is defined as a the deviation of a measured statistic from its true value,

typically caused by errors associated to deviations from optimal measurement conditions, simplification of complex behaviour of these measured systems, or assumptions made over the theoretical models to interpret the data. Systematic errors are constant deviations from these optimal conditions and result in biases which are not averaged out over multiple repetition of the experiment; statistical errors on the other hand, typically average out over the ensemble of the experiment. While random and systematic errors are difficult to minimize, they can be characterized by their effect on the observables, and propagated into the model being constrained. This characterization, usually presented as a function describing the range and associated confidence each point of this range is the true value, is called the **uncertainty**, and plays a fundamental role in the inference of cosmological parameters: large uncertainties in the measurements necessarily translate to large uncertainties in the parameters we are interested in measuring.

In this section, we will briefly describe the main sources of uncertainty in shear estimates and how biases in the measured fundamental quantities of weak lensing affect the cosmological parameters. In the next chapter we will do a similar description of photometric redshift biases.

2.3.1 Shear biases

The response of an optical system to an ideal point source is referred as the *point spread function*, or PSF, which indicates how light deviates and with what intensity at different locations around the expected position of the source, meaning that true intensity distribution of sources are convolved by this PSF resulting in the observed images. This evidently complicates the task of measuring the ellipticity of sources as it deforms the light distribution, and typically make small extended sources appear rounder that they otherwise would appear in the absence of a PSF distortion. In order to account for this effect, a *deconvolution* of the PSF must be done using a model that can be either estimated theoretically or obtained empirically from the data. But the PSF is usually dependent on the position of the source and orientation of the telescope and also can vary over time, which complicates the task even further. A model of a PSF can be built by acquiring many images of stars at different positions of the detector, which can be considered as point sources. Globular clusters offer an excellent target for obtaining stars on small fields of view at the expense of having to make specific observations at different times where the conditions of the optics may have changed. By interpolating the properties of the PSF to all positions in the CCD a general model can be used for specific positions of galaxies (for examples of modern sophisticated methods, see Liaudat et al., 2020; Jarvis et al., 2021, and references therein). Another option is to obtain a PSF using a physical model of the telescope and its optics, such as Tiny Tim for the HST (Krist et al., 2011), where several parameters can be fine adjusted to match the desired conditions of the observations.

Another important systematic effect called *blending* comes from the non-zero probability that two extended sources appear close enough in the projection of the sky to be mistakenly identified as a single object (for a more in depth description, see e.g. Dawson et al., 2016; Gaztanaga et al., 2021). This typically results in the detection of apparently

single objects with a larger ellipticity orientated along the vector that connects the pair. Because of the complex survey footprints shapes masking, it is usually more convenient to estimate the total effects of blending using suites of image simulations rather than trying to predict the probability and spatial distribution of pairs analytically (see for example MacCrann et al., 2020b).

Some other astrophysical and observational effects can also have an effect on the observed distributions of galaxies from a sample. For instance, dense outskirt dust structures can significantly obscure or redden the intensity from edge-on observed galaxies (Padilla & Strauss, 2008). This can result in them falling outside of the selection function of the survey, which can potentially lead to biases in the two-point correlation functions. Incorrect galaxy-star identification can also lead to an underestimation of the shear

The most common parameterisation for the biases of observed galaxy shear is of the form

$$\gamma_i^{\rm obs} = (1+m_i)\gamma_i^{\rm true} + c_i \tag{2.52}$$

where m_i, c_i are called the multiplicative and additive shear biases, respectively. Since the contributors to the additive biases and to the differences between m_1 and m_2 usually originate from the effect of the PSF in the shear field, it is common to assume that any residual bias after correction for it is independent from the measured component $m_1 = m_2 = m$, and $c_i = 0$. In a tomographic analysis, each bin is typically characterized by its own multiplicative shear bias parameter. Assuming that the tomographic shear biases are position independent, then the biased convergence power spectra from equation 2.25 (which we showed are equivalent to the convergence power spectrum) can be written as

$$\begin{aligned} P_{\kappa,ij}(\ell)^{\rm bias} &= \int\int d\varphi_1 d\varphi_2 e^{i\varphi_1\ell} e^{i\varphi_2\ell} \langle (1+m_j)\gamma(\varphi_1)(1+m_j)\gamma(\varphi_2)\rangle \\ &= (1+m_i)(1+m_j)\int\int d\varphi_1 d\varphi_2 e^{i\varphi_1\ell} e^{i\varphi_2\ell} \langle \gamma(\varphi_1)\gamma(\varphi_2)\rangle \\ &= (1+m_i)(1+m_j) P_{\kappa,ij}(\ell) \end{aligned} \tag{2.53}$$

This results in an over- or underestimation of the convergence power spectrum by a factor $(1+m_i)(1+m_j)$, and from the Limber equation 2.28, this carries on as an under- or overestimation of the matter density power spectrum amplitude, which from equation 1.40 can directly impact the inferred value of σ_8

Intrinsic alignments

The cosmological principle of no preferential directions in the Universe means that at large, galaxy orientations must be uncorrelated and that on average their orientations must average to zero. We have used this to justify the use of ellipticity measurements as an unbiased estimator of cosmic shear by taking the average of equation 2.48 over many galaxies. This assumption, however, neglects the fact that individual galaxies form and evolve on scales where the cosmological principle does not hold true. The processes involved in this evolution, including tidal torquing, gas accretion and merging are typically highly in-homogeneous in the evolution timescales of galaxies, meaning that their

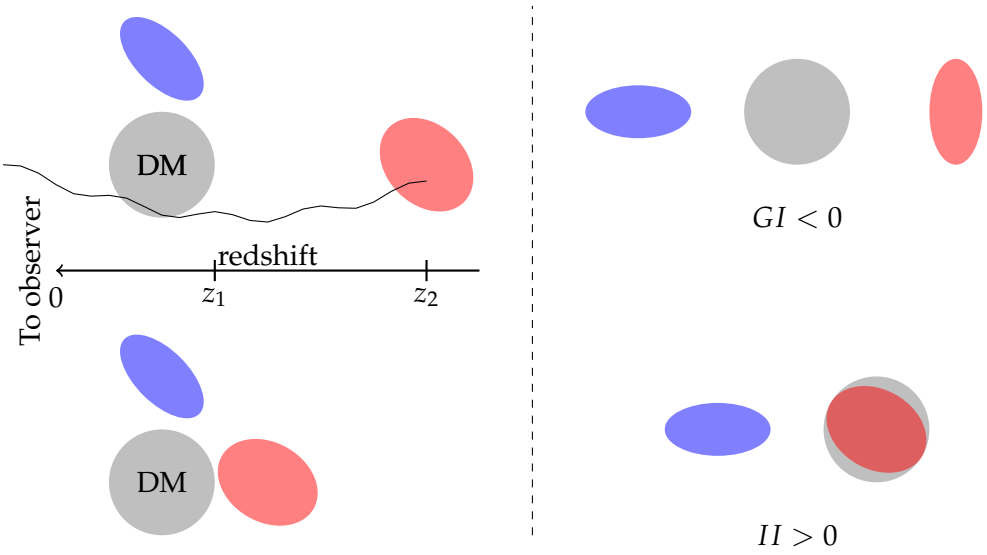

FIGURE 2.6: Illustration of the intrinsic alignment correlations. The top panel depicts the GI correlation caused by shapes of foreground objects being correlated with the shapes of background lensed objects. The blue foreground object tidally interacts with a matter structure at a similar redshift z_i pointing towards it. Background objects lensed by the same structure (red) appear to orientate tangentially, thus generating a negative correlation. Bottom panel shows the direct alignment between galaxies at similar redshift, typically resulting in a positive correlation II as objects point radially toward mass overdensities

effects on orientation cannot be expected to average out. As an example, brightest cluster galaxies typically show an orientation comparable to that of the cluster to which they belong, which in turn is heavily influenced by the large scale filamentary structure. (Jõeveer et al., 1978; Foëx et al., 2017). In the presence of a non-zero intrinsic alignment ellipticity correlation, the shear correlation of equation 2.48 can be written as

$$\begin{aligned} \langle e_{\rm obs}(\boldsymbol{\theta}) e_{\rm obs}(\boldsymbol{\theta}')\rangle &\sim \langle \gamma(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}')\rangle \\ &+ \left(\langle e_{\rm src}(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}')\rangle + \langle \gamma(\boldsymbol{\theta}) e_{\rm src}\rangle(\boldsymbol{\theta}')\right) \\ &+ \langle e_{\rm src}(\boldsymbol{\theta}) e_{\rm src}(\boldsymbol{\theta}')\rangle \\ &= GG + GI + II \end{aligned} \tag{2.54}$$

where GG refers to the cosmic shear correlation components and GI, II each describes the correlation of the intrinsic ellipticity with the shear field and itself, respectively. A basic interpretation of the origin of these alignment terms is presented in figure 2.6. The general idea is that galaxies correlate both with the structure nearby because of the dynamical and gravitational effects such as tidal torquing, merging, gas accretion (II correlation), and correlate with the shapes of background objects which are sheared by the same structures they are correlated with in the first place (GI correlation). Each of these effects can be modelled as an additive bias term to the shear power spectrum, and has been shown to be a significant contributor to the total bias in cosmic shear inferences (Kirk et al., 2012; Krause et al., 2016).

2.3.2 Photometric redshift biases

Until now we have focused primarily on establishing the weak lensing formalism and explained the general challenges and associated biases of measuring shapes, which are affected by the cosmic shear of the large scale structure. While this is an important step to reconstruct the statistical properties of the matter density, the total effect of the large scale structure on the images of background images cannot be fully described by the shear field. As we saw in equation 2.22, the lensing efficiency depends directly on the distribution of mass along the line of sight, $n(\chi)$. This dependency determines what is the relative effect of mass formations at different distances and is fundamental to reconstruct the 3D nature of the matter power spectrum.

The number density distribution of sources as a function of distance is typically estimated by measuring the redshift of sources, which for cosmological distances is an accurate measure of distance. Because of the limitations of spectroscopy and because most of the time we are more interested in the shape of this distribution $n(z)$ rather than individual galaxy redshifts, weak lensing surveys employ a family of techniques called *photometric redshifts*. In chapter 3 we will explain in detail the formalism, observables and associated biases of the n(z) estimation process and will describe the process employed in the analysis of the Dark Energy Survey first three years of data.

Chapter 3

Photometric redshifts in the Dark Energy Survey Y3 analysis

As we described in chapter 2, the measurement of the cosmic shear signal imprinted by the large scale structure on the shapes of distant galaxies is a powerful tool to directly characterize the mass distribution and expansion history of the Universe. In addition to the precise measurement of shapes and positions of galaxies, weak lensing and galaxy clustering surveys require a precise knowledge of the line-of-sight distribution of sources and a detailed description of the potential sources of uncertainty to correctly determine the uncertainty of the parameters describing the cosmological model.

In this chapter, we will focus on the observational challenge of obtaining these distributions and characterizing the uncertainties coming from different systematic sources. In section 3.1 we will introduce the concept of photometric redshift, and its importance in wide cosmology surveys, with an emphasis on its use for weak lensing and galaxy clustering experiments. Then we will describe the general aspects of the Dark Energy Survey and the DES Y3 analysis, paying special attention to the validation of the redshift distribution estimation pipeline presented in Myles et al. (2020). Here we will describe the characterisation of the different systematic effects and the steps employed to quantify their associated uncertainty, showing the results for the contribution to the uncertainty from a number of systematics.

We will focus on two particular sources of uncertainty in which we directly contributed to their estimation: sample variance in the deep photometric estimations, and uncertainty associated to the random nature of the Self-organizing map training.

The author of this contributed by designing and performing the test to quantify the source of uncertainty associated to the Self-organizing training consisting on running several repetitions with different random seeds. The author also aided in the setup to quantify cosmic variance by running the SOMPZ scheme on the BUZZARD simulated data to generate a set of $n(z)$ realisations which were also used in the tests in chapter 5. This also included methods to select the randomized BUZZARD catalogues and analysis of the intermediate results, as well as help in developing the SOMPZ code.

3.1 Photometric redshifts

Weak lensing and galaxy clustering experiments rely on the auto- and cross-correlation of shapes of a sample of galaxies, typically referred to as the *source* sample and the positions of set of large scale structure tracer galaxies, called the *lens* sample. The angular correlation functions describe the scales at which these quantities are related on the projected sky. Along with the description of the expansion history of the Universe they link these angular scales to physical distances to characterize the large scale structure. As shown in equation 2.22 to extract the relative weights of these signals detected in these correlation functions, encoded in the lensing efficiency functions, it is imperative to know the distances to the sources being measured. Moreover, separation of these samples into slices or *tomographic bins* along the line-of-sight provides additional insight on the evolution of these correlations (Hu, 1999).

3.1.1 Biases in the line of sight distributions of sources

Biases in the characterisation of the line-of-sight distribution of galaxies and how they are assigned to each bin have a large impact on the derived cosmological parameters. If unaccounted for or undetected, these can lead to a wrong interpretation of the history of the Universe. It has been shown that properties of the redshift distributions above their mean have a subdominant effect on cosmology (Huterer et al., 2006), specially considering the smoothing they are subject to in the lensing efficiency functions (Tessore & Harrison, 2020). Nonetheless, small biases in the mean of the distributions can have noticeable effects on the inferred cosmology. For example, if distances are systematically underestimated then we will be observing a younger Universe than we realise, leading to an underestimation of the growth factor. This can lead, for instance, to an overestimation of the matter density Ω_m or the amplitude of density fluctuations σ_8. An example of how the inference of those two parameters change from a simple Monte Carlo analysis given biased estimates of the distribution of sources along the line of sight is shown in figure 3.1. Furthermore, a given observed angular scale will be assigned an underestimated angular diameter distance, leading to an overestimation of its physical size. The exact bias in the cosmological parameters will depend on the combination of parameters being inferred.

At cosmological scales, the very definition of distance is a complex process: the Universe is constantly expanding and the physical distances change as a consequence of this expansion. Because of the homogeneous expansion at large scales expected from the cosmological principle, redshift becomes a convenient quantity to describe the distances to objects as it is directly related to their distances via Hubble's Law and is a relatively straightforward quantity to obtain. Modern optical surveys map ever-increasing fractions of the sky, at deeper limiting magnitudes resulting in catalogs containing tens to hundreds of millions of sources. While spectroscopy provides accurate measurements of redshift, it is prohibitively expensive even for the largest fiber-fed spectrographs being

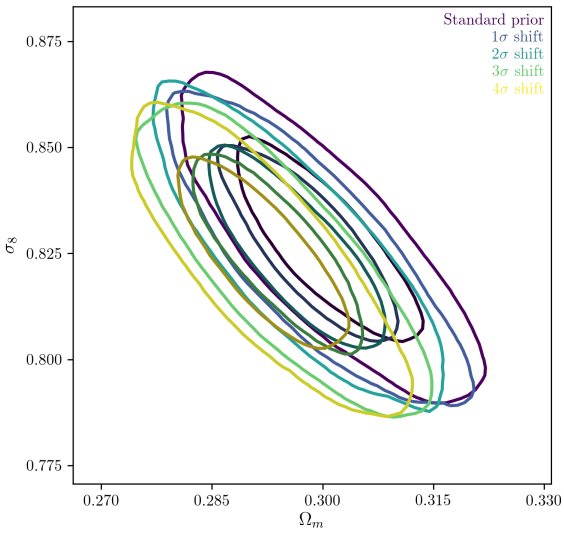

FIGURE 3.1: Evolution of the (Ω_m, σ_8) confidence contours as a function of bias on the line-of-sight distribution function. Each contour represents a MCMC chain obtained with a distribution biased by a factor proportional to $\sigma \sim 0.005$ in the negative direction for all tomographic bins, effectively underestimating the distances to objects, making the Universe appear to have formed the same structures at a lower redshift (later time), which is consistent with a Universe with a lower matter density Ω_m and/or a lower amplitude of density fluctuations σ_8.

able to obtain only a couple hundred objects at a time under very tight constraints of location and relative brightness. Another disadvantage of spectroscopy is the fact that it is typically only accurate up to relatively bright magnitudes, with high incompleteness being an issue for $I > 24$ (DESI Collaboration et al., 2016). A family of techniques that overcome these limitations, called **photometric redshifts** (PZ), utilizes information contained in wide-band photometry to constrain the redshift of galaxies using different sources of ancillary information to constrain the color-redshift relation of sources. The main limitations of these techniques is the characterization of the degeneracies in the color-redshift relation: two sources with different measured fluxes in a set of photometric bands can be equally consistent with more than one unique redshift. Figure 3.2 depicts this degeneracy by showing how different galaxy types can be observed to have the same measured colors when observed at different redshifts. In order to break these degeneracies, either additional or more precise photometric measurements and ancillary data are required.

Depending on the procedure used to extract information from photometry and the available ancillary data, PZ techniques can be broadly classified into three main categories:

- **Template Fitting**: This is perhaps the most widely utilized type of PZ technique. It involves fitting the observed band photometry to a series of Spectral Energy Distribution (SED) templates that broadly describes the expected types of galaxies in a survey. These templates, integrated over a band transfer function $W_b(\lambda)$ and a instrument specific optical response curve provide a redshift-dependent expected set of colors which can be compared to the input values inside a likelihood function. (Benítez, 2000, BPZ) first introduced the use of galaxy type priors to accommodate

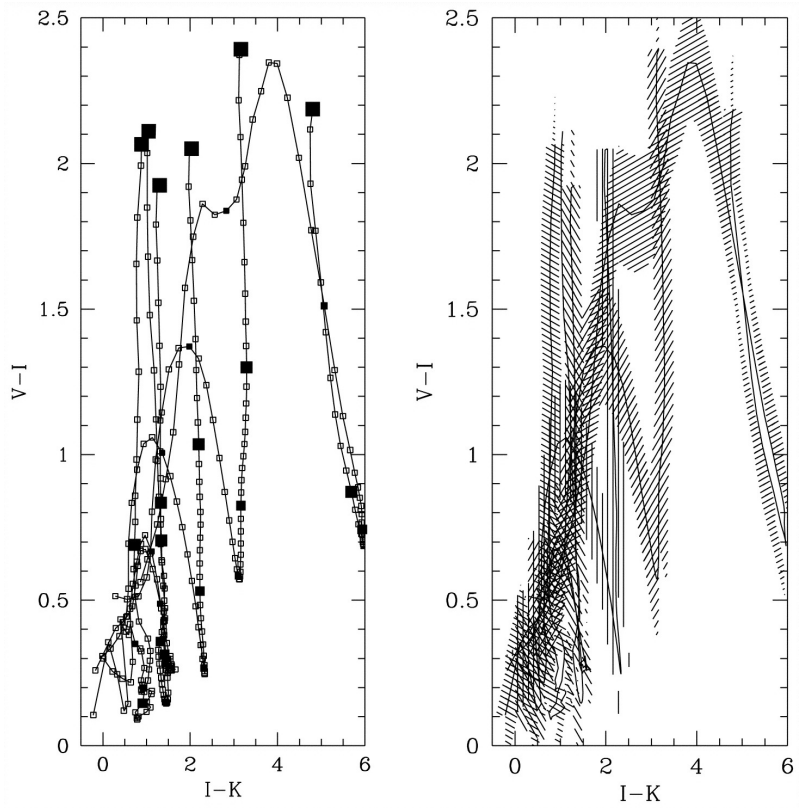

FIGURE 3.2: *Left:* panel shows the evolution of the color-color position of 6 spectral energy distribution templates as they get redshifted. Each line intersection corresponds to a degeneracy where two templates at different redshift appear to have the same colors. Right panel shows the same degeneracy, this time with a 0.2 magnitude error, showing how the degeneracies increase.
Figure credit: Benítez (2000)

for the fact that certain SED templates are not expected with the same probability in the full redshift range. This approach was used in DES SV (Bonnett et al., 2016) and Y1 (Hoyle et al., 2018) weak lensing and galaxy clustering cosmology results, as well as in CFHTLens (Erben et al., 2013) and KiDS (Kuijken et al., 2015). Bayesian approaches for template fitting return a full probability distribution for the photometric redshift of a source given its photometric measurements and similarities with a library of SEDs. $n(z)$ distributions are then obtained by *stacking* of these distributions, which typically means drawing a random sample from each redshift probability distribution obtained and constructing a histogram of sampled values. A recurring problem with this a approach is that the degeneracies in the color-redshift relation lead to multiple peaked probability distributions which when stacked can lead to high rates of catastrophic failures, where the true redshift and estimated photometric redshift differ significantly.

- **Machine Learning**: This category encompasses a series of techniques which use different training sets to characterize the color-redshift space and its mapping to the photometric measurements. While the specific set of ancillary data used can spawn its own sub-classification, these methods typically use a sub-set of data with precise redshift information and photometry measured in conditions similar to those of the bulk of the data to be analysed. Using this data a model of the color-redshift relation is constructed by a process called *training*, in which a mapping between the two, characteristic of the particular training set, is generated. Several different approaches have been used to estimate photometric redshifts, including the use of Neural Networks (Gerdes et al., 2010), decision trees and random forests (Carrasco Kind & Brunner, 2013), Gaussian processes (Way et al., 2009) among others. DES SV and Y1 provided alternative PZ estimations using two of such methods: ANNZ (Collister & Lahav, 2004), a method based on Neural Networks and DNF (De Vicente et al., 2016) which builds a metric in color space to assign best-fit galaxies to training set elements. HSC employed and compared several machine learning algorithms for its first data release (Tanaka et al., 2018) and KiDS (Wright et al., 2020) recently employed a Self-organized map algorithm.

- **Angular correlations**: This family of PZ techniques relies on constraining the redshift distribution of sources by using their angular correlation with a tracer population with well constrained redshifts. These methods are becoming increasingly popular as the number of surveyed sources increases, allowing the statistical significance of the methods to increase. They also provide independent constraints to conventional PZ methods and are subject to uncorrelated systematics which further help breaking the color-redshift degeneracies. These methods typically rely on the cross correlation of source samples with the positions of a physically close tracer sample (Newman, 2008; Ménard et al., 2013), and with the measured shear of a non-physically related tracer sample (Sánchez et al., 2020a). The former approach has been used in DES (Davis et al., 2017) and KiDS (Hildebrandt et al., 2021).

Current generation surveys aim at estimating the positions, shears and photometric redshifts of tens to hundreds million galaxies, which requires efficient, yet accurate estimations of their properties. The Dark Energy Survey, in particular, has generated a large catalog of shapes for over 100 million objects which will be used to estimate the two-point shear correlation functions and an accurate estimation of their redshift distribution is fundamental to fully take advantage of its statistical power. In the next section we describe the main characteristics of the survey and provide some detail in the process to obtain to estimate the redshift distributions, calibrate the methods employed and characterize their uncertainty.

3.2 DES Y3: Redshift Calibration of the Weak Lensing Source Galaxies

3.2.1 The Dark Energy Survey

The Dark Energy Survey[1] (DES) is an optical and near infrared survey of the southern sky, aimed at characterizing Dark Energy by constraining the evolution and geometry of large scale structure through obtaining multi-band imaging of nearly 5000 square degrees, up to a limiting magnitude for extended sources of 23 in the i band using a purposely built CCD camera, DECam, mounted on the 4m Blanco telescope at CTIO, Chile. While the survey benefits from multiple ancillary data to help perform its analysis, including overlap with spectroscopic and many-band photometric redshift surveys, the main data component are observations carried over a period of over 6 years with first light starting in September 2012 and last science images obtained in January 2019. The survey aims to characterize more than 300 million objects with 4 primary probes of the large scale structure of the Universe in mind. Since we already presented a detailed description of these probes in 1.3, here we only present a few highlights related to the latest published results obtained up until the release of the Y1 analysis of data, with Y3 currently underway.

- **Supernovae:** DES presented its first SNe cosmological parameter constrains using a sample of 207 spectroscopically confirmed SNe Ia in the redshift range $0.07 < z < 0.85$ (DES-SN sample) in combination with a low-z sample ($z < 0.1$) from the literature. The SNe detection was made by comparing high cadence observations of ten 2.7 deg^2 fields over three periods of five months each between Aug. 2013 and Feb. 2016. Results presented in (Abbott et al., 2019c) report a value for $\Omega_m = 0.331 \pm 0.038$ based purely on the SN observations, while combination with CMB yield a dark energy equation of state $\omega = -0.978 \pm 0.059$ and $\Omega_m = 0.321 \pm 0.018$, consistent with a cosmological constant. A combination with BAO to constrain absolute magnitudes of the SN is also used in Macaulay et al. (2019) to provide constrains of the Hubble parameter, finding $H_0 = 67.8 \pm 1.3$ kms^{-1} Mpc^{-1}, consistent with estimates derived from CMB in a ΛCDM Universe.

[1] https://www.darkenergysurvey.org

- **Baryon Acoustic Oscillations**: Abbott et al. (2019b) presented the measurements of the BAO scale on a sample of 1.3 million galaxies in the redshift range $0.6 < z < 1.0$, providing constraints on the ratio of the angular diameter distance to the sound horizon at the drag epoch, D_A/r_d. The measurement was obtained at the effective redshift of the sample, $z = 0.81$ and parameter likelihoods were obtained using a set of 1800 DES mocks which simulate the redshift distributions of sources of DES data. Reported values for the ratio, $D_A/r_d = 10.75 \pm 0.43$, obtained using three separate estimators for the clustering signal (Angular correlation function $\omega(\theta)$, angular power spectrum measured through the spherical harmonics C_ℓ, and projected comoving separation correlation function $\xi(s_\perp, s_\parallel)$) are both self consistent and consistent with estimates from different probes in the context of a flat ΛCDM model.

- **Galaxy Clustering and Weak Lensing** Cosmological parameters from joint analysis of galaxy clustering and weak lensing analysis have been presented in Abbott et al. (2018a), where shapes and photometric redshifts of 26 million galaxies in the redshift range $0.2 < z < 1.3$, separated into 4 tomographic bins have been used to obtain the two-point shear correlation functions $\xi_\pm(\theta)$. Positions and photometric redshifts of a sample of $\sim 650,000$ bright red sequence galaxies selected using the REDMAGIC algorithm (Rozo et al., 2016), with redshifts between $0.15 < z < 0.9$, separated into 5 tomographic bins were also used to compute the galaxy clustering angular correlation $\omega(\theta)$. Shapes have been measured using the METACALIBRATION (Sheldon et al., 2020) method and both samples employ the BPZ photometric redshift code, described in section 3.1. In addition to the two correlation functions above, the cross correlation between shapes and positions of both samples of galaxies, $\gamma_t(\theta)$, have been used to constrain the matter density Ω_m and clustering parameter $S_8 \equiv \sigma_8 \left(\Omega_m/0.3\right)^{0.5}$ both in the ΛCDM ($S_8 = 0.773^{+0.029}_{-0.020}$, $\Omega_m = 0.267^{+0.030}_{-0.017}$) and ωCDM models ($S_8 = 0.782^{+0.036}_{-0.024}$, $\Omega_m = 0.284^{+0.033}_{-0.030}$, $\omega = -0.82^{+0.021}_{-0.020}$), which are consistent across the three correlation functions. Both results appeared to be slightly in tension with CMB results from Planck (Planck Collaboration et al., 2016)

Abbott et al. (2019a) presents cosmological constraints from the four probes described above, independent from any other external experiment and consistent with a spatial flatness and ruling out a Universe with no dark energy (See figure 3.3). These are the most constraining results to date for cosmological parameters obtained purely from an optical survey.

In addition to the tension with CMB experiments that most weak lensing experiments show, tension between DES and the results reported in the analysis of the first release of 450 square degrees of multi-band data of KiDS+VIKING (KV450; Hildebrandt et al., 2017) sparked an interesting discussion regarding the origin of these inconsistencies, citing calibration of photometric redshifts as a probable source of inconsistencies. KV450 utilized a *Direct* calibration approach where the redshift distribution of a spectroscopic sample is weighted using a k-nearest-neighbor matching method to the source sample in color

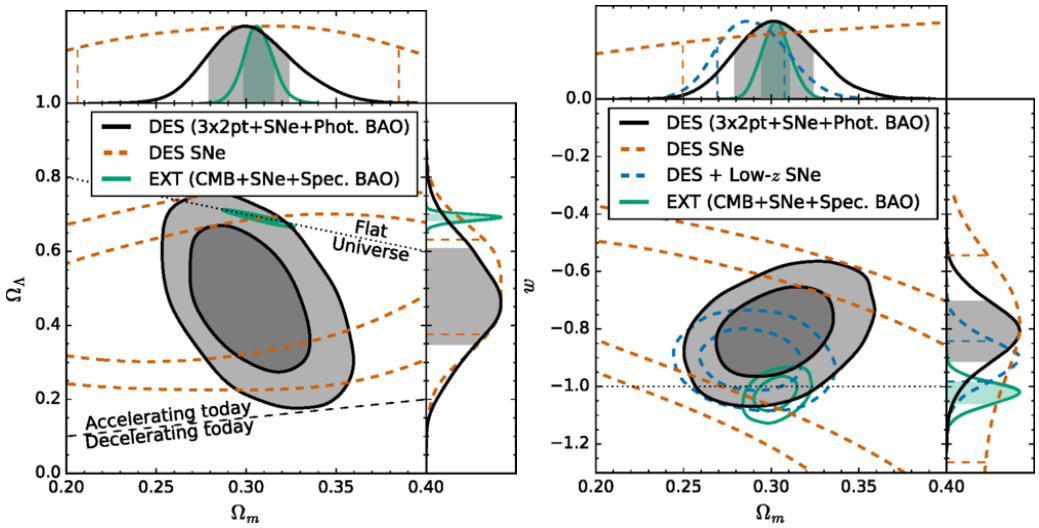

FIGURE 3.3: *Left:* 68% and 95% contour levels of the joint marginalized posterior for the dark energy density Ω_λ and matter density Ω_m comparing the best external datasets to the date of publication (May 2019, green contours), DES-SNe alone (Dashed orange) and four combined DES probes (Black). Dotted and dashed black lines identify a flat universe ($\Omega_k = 0$) and the interface between a decelaring and accelerating Universe, respectively. *Right:* Constraints on the dark energy equation of state ω and matter density Ω_m. Color scheme is the same as in right, with the added green dashed contours showing the impact of adding a low-z SNe dataset.
Figure credit: Abbott et al. (2019a)

space, while DES used a set of precise many-band photometric redshifts (COSMOS30; Laigle et al., 2016) to calibrate the photometric redshift estimates obtained with BPZ and to quantify their uncertainties. Joudaki et al. (2020) argues that the high fraction ($\sim$ 6%) of catastrophic outliers in the COSMOS30 photometric redshifts is enough to bias the mean S_8 estimates by $\sim 0.8\sigma$. They also comment on the use of uncorrelated nuisance parameters as a potential source of discrepancies. Gruen & Brimioulle (2017); Hartley et al. (2020b) show that the use of spectroscopic samples for redshift calibration is subject to large incompleteness because of selection biases, which can result in biases in the mean redshifts of weak lensing samples of the order of $\Delta_z \sim 0.04$. In any case, it is extremely difficult to arrive to any conclusions on the true source of discrepancy since both surveys construct their datasets based on observations which are not equivalent. Regardless, the approach devised for the Y3 analysis of DES data is conscious of these sources of uncertainty, and hence aimed at including a scheme which alleviates the effect of catastrophic outliers and includes information coming both from precise spectroscopic redshift samples and precise photometric redshifts which do not suffer from incompleteness at lower magnitudes.

3.2.2 DES Y3 source redshift calibration

The redshift calibration of DES Y3 weak lensing source galaxies combines a machine learning photometric redshift algorithm SOMPZ, and two angular correlation techniques to further constrain the fiducial distribution obtained using independent measurements

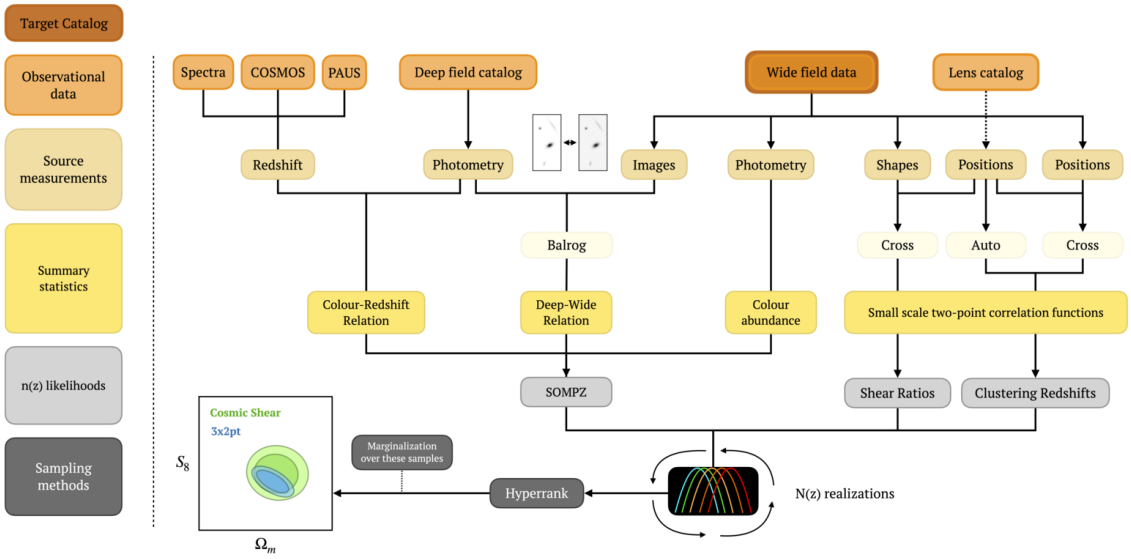

FIGURE 3.4: Flowchart illustrating the weak lensing redshift distributions calibration scheme for the DES Y3 weak lensing and galaxy cluster analysis. The three main redshift distribution likelihood functions of the analysis, shown in gray, are SOMPZ, clustering redshifts, and shear ratios.
Figure credit: Myles et al. (2020)

of the angular correlation with galaxy positions and shears from complementary galaxy samples:

- Self-Organizing Map $p(z)$ (SOMPZ; Buchs et al., 2019; Myles et al., 2020) leverages the Y3 DES Deep Fields to accurately determine the number density of galaxies in deep $ugrizJHK_s$ color space. Since redshifts are well-constrained at a given $ugrizJHK_s$ color, this number density can be used to properly weigh galaxies within a sample of credible redshifts in a way that is not subject to selection biases. In brief, this method relies on determining the $p(z)$ at a given cell in 8-band color space from galaxies with deep 8-band coverage, the probability of each cell in 8-band color space contributing to the galaxies in a given cell in noisy 3-band color-magnitude space, and the abundance of galaxies in 3-band color-magnitude space, to compute the overall redshift distribution of the Year 3 lensing source galaxy sample. We describe this method in detail in section 3.2.3

- Clustering redshifts (WZ; Gatti et al., 2020a; Cawthon et al., 2020) constrains the distances to source galaxies from their angular correlation with the positions of a reference sample of galaxies with secure redshifts estimates. This method is based on the fact that the amplitude of this correlation function is proportional to the fraction of source galaxies in physical proximity to those reference galaxies. Clustering redshifts validate and refine photometric $n(z)$ with the key benefit of avoiding any reliance on the statistical color-redshift relation and bypassing the completeness issues associated with spectroscopic survey coverage.

- Shear ratios (SR; Sánchez et al., 2020a) provide additional constraining power by measuring the position-shear correlation signal from the source sample with respect to a sample of lens galaxies, in contiguous redshift bins and at small scales. The ratio of this signal from two source bins reflects the ratio of mean lensing efficiencies of objects in those source bins with respect to the lens bin redshift. This, in turn, depends on the redshift distribution of the sources. Because this methodology utilizes lensing signals, it is virtually independent from SOMPZ and clustering redshifts. Both the clustering and shear ratio redshift constraints are derived from data on small angular scales, which allows the redshift constraints to remain largely statistically independent of cosmological constraints based on larger-scale signals.

Each of these three methods provides a likelihood function for the redshift distribution of sources, which can be combined into a *three-step Dirichlet* sampling (3SDIR) process to provide samples of the $n(z)$ posterior. Unfortunately this sampling procedure is slow as many samples are rejected by the clustering likelihood. By contrast, a Hamiltonian Monte-Carlo (HMC) sampler has the ability to draw from the joint combination of likelihoods and, although drawing individual samples is slower, sampling the joint space becomes much more efficient and fast. A combined 3SDIR + HMC approach, described in Alarcon et al. (2020); Bernstein (2020) is used to generate samples of the $n(z)$ posterior distribution.

In the following section we focus on describing the SOMPZ method and the characterisation of the redshift uncertainty associated to the observables for the DES Y3. We give special attention to two sources of uncertainty: Sample variance and shot noise, and stochastic uncertainty from the SOM training.

3.2.3 The SOMPZ scheme

The SOMPZ scheme is based on the formalism proposed in Buchs et al. (2019) which takes advantage of how precise *deep* photometry can help break the degeneracy between noisy *wide* photometric measurements and a training set of secure redshifts z. We start by assuming the *wide* sample consists of a large set of galaxies with photometry $\hat{\mathbf{x}}$ described by a photometric error $\hat{\mathbf{\Sigma}}$, for which we want to infer its redshift distribution. The *deep* sample refers to a set of galaxies for which its observed photometric properties, $\mathbf{x}$, are obtained with greater precision and/or in a larger number of bands which helps breaking the color-redshift degeneracy, typically a sub-set of the wide sample. The condition of overlapping is not a requirement, as long as both samples are complete and map the same galaxy populations. Secure redshifts z can be obtained by alternative methods like spectroscopy or many-band photometric redshifts, and must be obtained for a sub-set of galaxies of the deep sample so the later can be used to leverage the redshift information to characterize the wide sample (See figure 3.5).

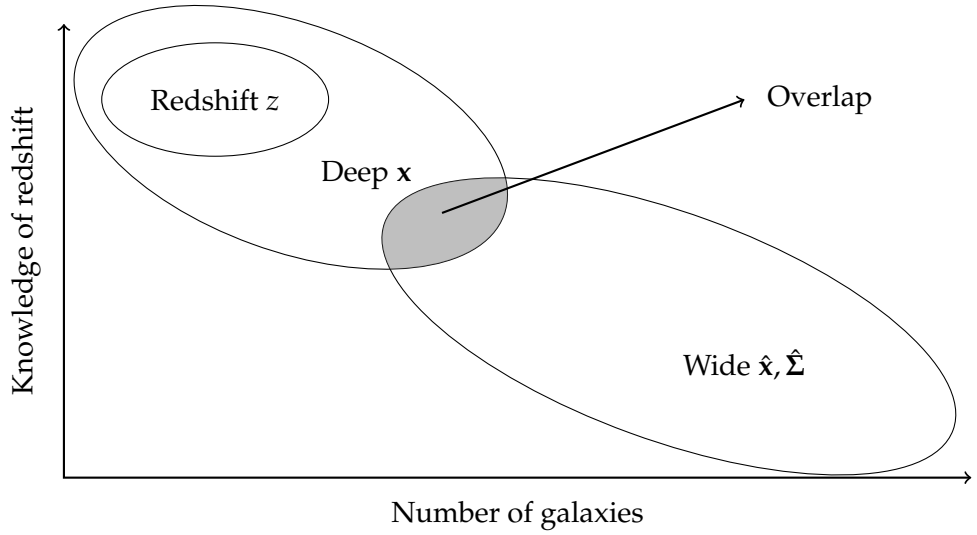

FIGURE 3.5: Cartoon representation of the samples employed in the estimation of the source redshift distributions using the SOMPZ scheme. The $n(z)$ distribution of the target *wide* sample is estimated by leveraging its overlap with a smaller sample of galaxies, the *deep* sample, for which its photometry is acquired either with higher precision or in a larger number of bands. A smaller set of very precise redshift estimates, the *redshift* sample, is then used to anchor the transition matrix between the *wide* and *deep* sample discretizations.

For any single galaxy, an estimator of its redshift probability distribution given a set of photometric measurements $\hat{\mathbf{x}}$ and associated photometric errors $\hat{\mathbf{\Sigma}}$ is given by:

$$p(z|\hat{\mathbf{x}}, \hat{\mathbf{\Sigma}}) = \int d\mathbf{x}\, p(z|\mathbf{x})\, p(\mathbf{x}|\hat{\mathbf{x}}, \hat{\mathbf{\Sigma}}), \tag{3.1}$$

where it is assumed that the photometric errors do not add additional information to the relation between redshift and deep photometry, (i.e. $p(z|\mathbf{x}, \hat{\mathbf{x}}, \hat{\mathbf{\Sigma}}) = p(z|\mathbf{x})$). In reality, we are interested in learning what is the distribution for an ensemble of galaxies,

$$p(z) = \frac{1}{N} \sum_{i=1}^{N} \int d\mathbf{x}\, p(z|\mathbf{x})\, p(\mathbf{x}|\hat{\mathbf{x}}_i, \hat{\mathbf{\Sigma}}_i) \tag{3.2}$$

but in any case, learning the mapping between the deep and wide photometric spaces is infeasible given the high dimensionality and non-linearity of the color-redshift degeneracy. The SOMPZ scheme overcomes this difficulty by discretizing the color space into cells called *phenotypes*. This term refers to the fact that observable quantities $\mathbf{x}$ and $\hat{\mathbf{x}}$ are a result of underlying intrinsic properties of these galaxies. The main assumption is that galaxies assigned to the same cells of this discretization share the same underlying properties, as long as the cells are small with respect to the errors $\hat{\mathbf{\Sigma}}$. Given a discretization of the space of wide colors, c and deep colors $\hat{c}$ the discretized version of equation 3.2 can be written as

$$p(z) = \sum_{c,\hat{c}} p(z|c)\, p(c|\hat{c})\, p(\hat{c}) \tag{3.3}$$

where the three terms can be determined from the specific partition $c, \hat{c}$ and the assignment of galaxies to each cell. $p(z|c)$ is the distribution of secure redshifts for galaxies assigned to the partition of the deep photometry, c. In order to determine this distribution there must exist a sub-sample of galaxies for which both deep photometry and secure redshifts are provided. Not all galaxies in the deep sample have to have a secure redshift measurement, but is expected that the redshift sample z is *complete*, meaning that all cells c contain at least one member with a redshift measurement. The galaxies assigned to that cell will be assumed to have a distribution equal to that of the galaxies which do have redshift measurements. $p(\hat{c})$ is the fractional assignment of galaxies in the wide partition $\hat{c}$, and $p(c|\hat{c})$ is called the transfer matrix, which indicates the probability of galaxies with noisy photometry assigned to discretization $\hat{c}$ of having precise photometry corresponding to a cell c in the deep partition. This last term is computed using a set of galaxies with observations made both in the deep subset and the wide sample, called the *overlap* sample, or alternatively simulated injection of galaxies with known properties into the wide imaging.

Once galaxies have been categorized into phenotypes based on their photometric observations, tomographic bins are constructed and assign each phenotype $\hat{c}$ to a bin according to the following procedure:

1. To construct a set of n tomographic bins $\hat{b}$, begin with an arbitrary set of $n+1$ bin edge values e_j.

2. Assign each galaxy in the redshift sample to the tomographic bin $\hat{b}$ in which the best-estimate median redshift value of its $p(z)$ (or its secure redshift z) falls. This yields an integral number of galaxies $N_{\mathrm{spec},(\hat{c},\hat{b})}$ satisfying the dual condition of membership in a wide discretisation cell $\hat{c}$ and a tomographic bin $\hat{b}$.

3. Assign each wide cell $\hat{c}$ to the bin $\hat{b}$ to which the majority of its constituent Redshift Sample galaxies are assigned

4. Adjust the edge values e_j post hoc such that the numbers of galaxies in each tomographic bin $\hat{b}$ are approximately equal and repeat the procedure from step (ii) with the final edges e_j.

The approach taken by the SOMPZ is to use a smart discretization of the space of colors employing Self-organized maps, both in the deep and wide samples, which allows for a relatively uniform number of samples assigned to each discretization element, and an easy two dimensional representation of data.

Self-Organized Maps

A self-organized map (SOM; Kohonen, 1982, 2001) is a data structure generated via an iterative machine learning algorithm which generates an adaptive discretization of the parameter space containing the input samples, or *features*. This discretization is mapped

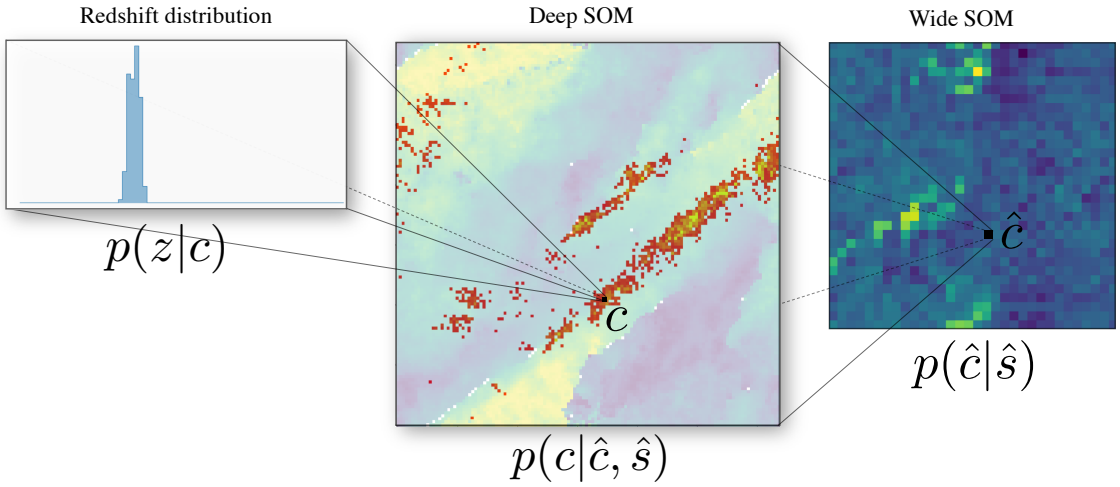

FIGURE 3.6: Graphical representation of the SOMPZ scheme. Not all galaxies in a cell c of the deep SOM have a measured redshift, so all of them are assumed to follow the same distribution as the ones who do have a measurement, shown in the left panel. Galaxies from several cells in the deep SOM would be assigned to wide SOM cells $\hat{c}$ due to its noisy photometry. These cells are highlighted in the central panel and contribute their histogram, weighted by the transition matrix computed previously either using simulated data or the overlap sample.
Figure credit: Buchs et al. (2019)

to a lower l-dimensional representation of this multidimensional input space. The process is unsupervised, meaning that the features of data aimed at being identified and classified are not used as input information in the process. Depending on the correlation between the features and the unsupervised properties, the map will generate a smooth mapping along the dimensions of the low-dimensional representation, preserving the notion of locality. This applies very well to the problem of estimating redshifts (the unsupervised property) from a set of colors or fluxes (the features). The process of generating self-organized maps is divided into two stages for a given set of data.

- The *training* stage is a competitive process in which a series of randomly initialized points in the input feature space, called *nodes*, are iteratively rearranged to resemble the distribution of the input sample. These nodes are represented by *weight vectors*: points on the input space defining the boundaries of the discretization around them based on a pre-defined distance metric. The training starts by defining a set of C weight vectors $\omega_k \in \mathcal{R}^m$ where m is the number of features describing each sample. These can be either assigned randomly, from a pre-defined prior, or by randomly choosing C elements from the input sample.

  Once the training stage is completed, either after a fixed number of iterations or based on a metric to quantify the similarity between the distribution of weight vectors and the input sample, during the *assignment* phase, each element of input sample is matched to the node whose weight vector best resembles it, according to the metric defined above.

While the relative positions of these weight vectors will change in input space during training, their definition of neighborhood remains constant in the low dimensional representation space. Their topology can be defined *a priori* as well, and in the case of SOMPZ, the cell grid is assumed to be square with a periodic boundary condition, thus resembling a three dimensional torus.

On each step t of the training stage a sample from the training set with values $\mathbf{x}(t)$ is chosen at random, and the closest weight vector is identified, according to

$$d^2(\mathbf{x}, \omega_k) = (\mathbf{x} - \omega_k)^\top \Sigma^{-1} (\mathbf{x} - \omega_k) \tag{3.4}$$

where Σ is a covariance matrix which describes the relative importance of the elements of the input vector $\mathbf{x}$ in the distance metric. Once the closest node is identified, its position and the positions of a subset of nodes around it are updated by moving them in the direction of the sampled point, inversely proportionally to its distance according to

$$\omega_k(t+1) = \omega_k(t) + a(t) H_{b,k}(t) \left[\mathbf{x}(t) - \omega_k(t)\right]. \tag{3.5}$$

It is worth noting that the closeness of nodes is defined on the fixed low-dimensional grid rather than the input space. $a(t)$ is the learning rate function, and is in charge of modulating the relative effect of new samples as the distribution of nodes starts resembling the input sample. Without it, the effect of the first samples would be completely erased by the samples taken at later stages of the training. It takes the form $a(t) = a_0^{t/t_{\max}}$ with $a_0 \in [0, 1)$. Similarly, the subset of neighbouring nodes which are updated in each successive sample also decreases according to the function $H_{b,k}(t)$, which is defined by a Gaussian kernel centered around the closest node to the sample at iteration t:

$$H_{b,k}(t) = \exp\left[-D_{b,k}^2 / \sigma^2(t)\right] \tag{3.6}$$

with $D_{b,k}$ being the Euclidean distance between nodes in the grid:

$$D_{b,k} = \sum_{i=1}^{l} (c_{b,i} - c_{k,i})^2 \tag{3.7}$$

The width of the Gaussian kernel is parameterized by $\sigma(t) = \sigma_s^{1-t/t_{\max}}$ and similarly to the learning rate function, prevents the map to be over-trained at later iterations by reducing the volume of nodes around the central node which are affected in equation 3.5. A graphical representation of one step of the training stage is shown in figure 3.7.

3.2.4 Applying the SOMPZ scheme to DES Y3 data

The application of the SOMPZ scheme in the DES Y3 analysis employs data products obtained from different stages of the weak lensing pipeline. The *wide* sample for which the $n(z)$ distribution is to be obtained consists of over 100 million galaxies with measured r, i, and z METACALIBRATION photometry and shapes. While photometry in the g band

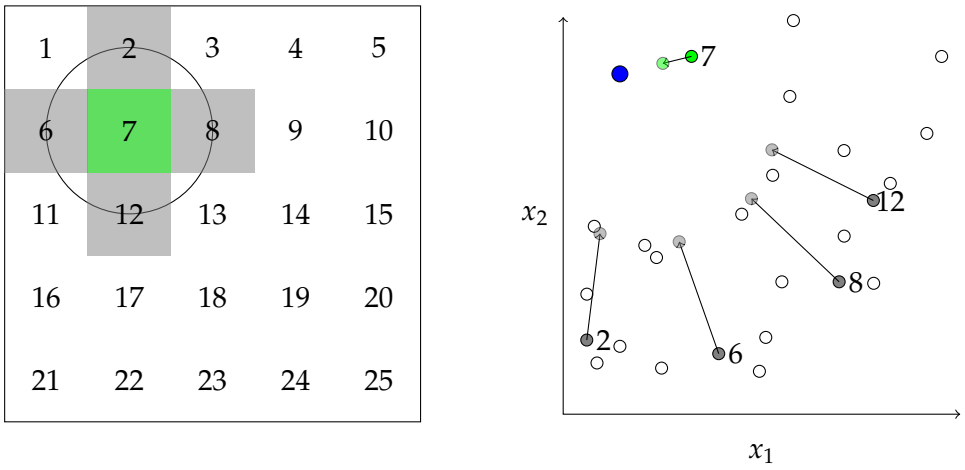

FIGURE 3.7: Graphical representation of the self-organizing map training stage. A sample from the training set is chosen randomly (green circle in right panel) and the closest weight vector in feature space $\mathbf{x} = (x_1, x_2)$ is identified, both in the feature space (blue circle in right panel) and in the low dimensional map (7th grid point in left panel). The neighbouring grid points in the low-dimensional map (Gray squares on left panel and gray circles in right panel) are identified according to $H_{b,k}(t)$ (Black circle on left) and their positions adjusted proportionally to the distance and in the direction of the sampled point. The process is repeated $t_{\texttt{max}}$ times with each set of neighbouring points changing depending on the sampled point.

exists, the determination of the PSF model is particularly challenging and estimates of the shear and photometry using METACALIBRATION have shown to be unreliable. Details about the catalog and the PSF issues are described in detail in Gatti et al. (2020b) and references therein. The *deep* sample corresponds to approximate 2.8 million objects observed in 4 regions of the DES Y3 footprint overlapping with the UltraVista (McCracken et al., 2012) and VIDEO (Jarvis et al., 2013) near-infrared surveys, providing photometry in the *ugrizJHK* bands. The *redshift* sample is a compilation of spectroscopic and many-band secure redshifts from surveys with overlap with the deep sample. Spectroscopic surveys include zCOSMOS (Lilly et al., 2009), C3R2 (Masters et al., 2017, 2019), VVDS (Le Fèvre et al., 2013) and VIPERS (Scodeggio et al., 2018), and secure photometric redshifts are obtained from COSMOS2015 30-band PZ catalogue (Laigle et al., 2016) and PAUS+COSMOS which combines photometric data from the PAU Survey (Padilla et al., 2019; Eriksen et al., 2019) and COSMOS2015. In order to characterize the transition matrix $P(\hat{c}|c)$ DES uses nearly 2.5 million injection-redetection BALROG galaxy pairs. BALROG (Everett et al., 2020) is a procedure in which artificial galaxies with photometric properties drawn from the deep sample are injected into real images and then recovered by the main photometry pipeline.

The training of the deep and wide SOM employs galaxy magnitudes μ based on the definition given in (Lupton et al., 1999), called *luptitudes*:

$$\mu \equiv -a \left[\sinh^{-1}\left(\frac{x}{2b}\right) + \ln b \right] \tag{3.8}$$

where a, b are constants, with b a softening parameters setting the flux x at which the new magnitude systems starts resembling the traditional one using logarithms. This

definition ensures that a value is defined even in the case of negative fluxes, which can occur when detections of sources are based on a single band. The feature vector $\mathbf{x}$ for the deep SOM is a set of lupticolors $\mu_x - \mu_i$ defined from the observed luptitudes μ_i, μ_r, and μ_z, all with respect to luptitude in the i band. For the wide SOM the feature vector is the same as the deep SOM, plus the luptitude in the i band, μ_i.

$$\mathbf{x} = \{\mu_x - \mu_i\} \tag{3.9}$$

$$\hat{\mathbf{x}} = \{\mu_i, \mu_{x_m} - \mu_i\}. \tag{3.10}$$

The choice of number of cells and dimensions of the 2D SOM maps must take into consideration two competing effects. On one hand, a large SOM map allows for a more precise discretization of the feature space $\mathbf{x}$, as long as the associated errors $\mathbf{\Sigma}$ are small compared to the typical resultant distances between weight vectors for each node. On the other hand, a large number of cells can result in a noisy fractional assignment if the number of galaxies is low or even comparable to the number of cells. The *wide* and *deep* maps employed in DES Y3 use 32×32 and 64×64 cells respectively. While empty cells after assignment is not an impediment to compute the tomographic redshift distributions, it can be a symptom of the SOM training failing to correctly distribute the weight vectors across the input space.

Figure 3.8 shows the resulting ensemble of redshift distributions obtained by sampling from the 3sDIR + HMC combining the estimates of the SOMPZ, clustering redshifts WZ. The 3sDIR + HMC also combines the uncertainty estimates directly into the sampling, hence these $n(z)$ realisations all incorporate the total uncertainty identified in the pipeline. In the next section we describe what are the identified systematic effects in the pipeline, and how their associated uncertainties impact the $n(z)$ estimates.

3.3 Uncertainty characterisation of the DES Y3 source redshift distributions

Identifying the sources of uncertainty across the entire weak lensing analysis pipeline is a fundamental step to construct a reliable set of estimations on the cosmological parameters. The uncertainties associated to these systematic effects must be clearly identified and characterized to be propagated into the model parameters. Six main contributor effects to the uncertainty on the DES Y3 pipeline are identified in Myles et al. (2020):

(i) **Sample variance**: Fluctuations in the underlying matter density field determine the abundance of observed deep field galaxies of a given 8-band color and at a given redshift in the footprint of the DES survey.

(ii) **Shot Noise**: shot noise in the counts of deep field galaxies of a given 8-band color and at a given redshift

(iii) **Redshift Sample Uncertainty**: biases in the redshifts values and incompleteness of the secure redshift galaxy sample

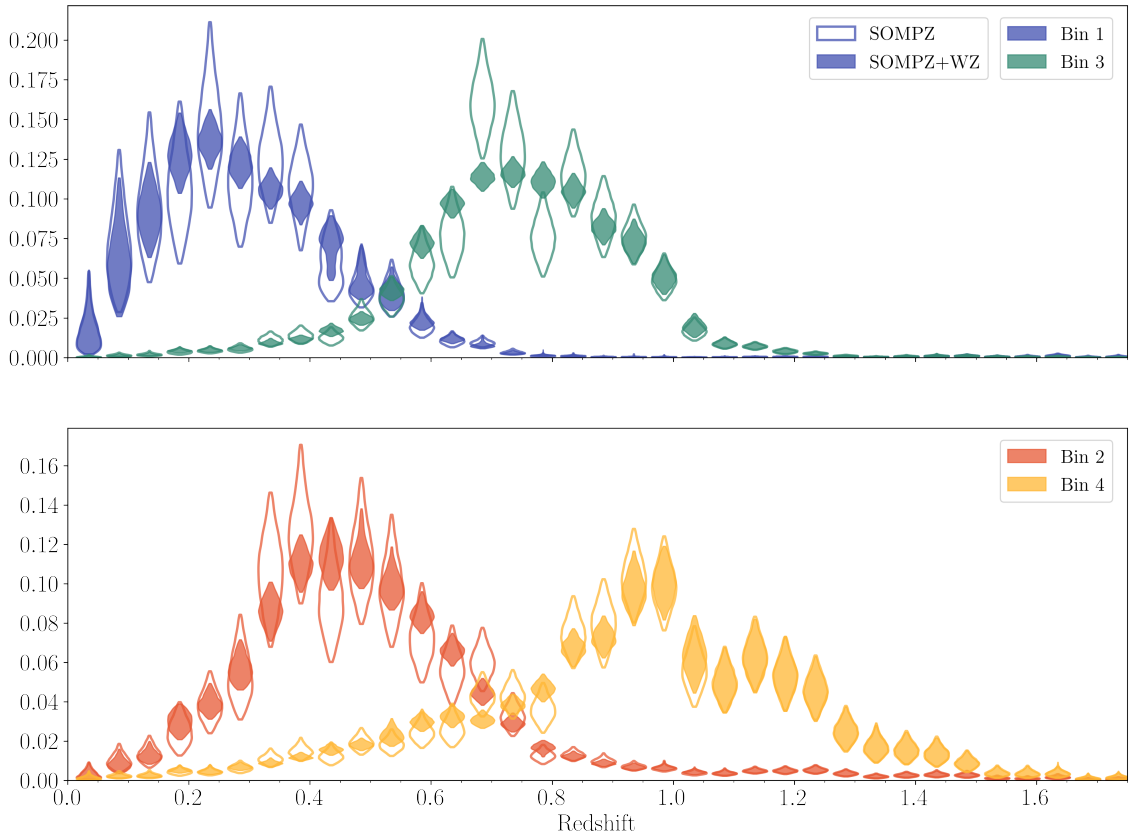

FIGURE 3.8: Visualization of the ensemble of redshift distributions in four tomographic bins, as inferred from SOMPZ only (open), and from SOMPZ combined with WZ (filled), obtained using the 3SDIR + HMC formalism.
Figure credit: (Myles et al., 2020)

$n(z)$ mean values		Bin 1 0.0—0.358	Bin 2 0.358—0.631	Bin 3 0.631—0.872	Bin 4 0.872—2.0
$\langle z \rangle$ SOMPZ		0.332	0.520	0.750	0.944
$\langle z \rangle$ SOMPZ + WZ		0.339	0.528	0.752	0.952
Effective $\langle z \rangle$ SOMPZ + WZ + Blending		0.336	0.521	0.741	0.935
Effective $\langle z \rangle$ SOMPZ + WZ + SR + Blending		0.343	0.521	0.742	0.964
Uncertainty	**Method**				
Shot Noise & Sample Variance	3SDIR	0.006	0.005	0.004	0.006
Redshift Sample Uncertainty	Sampling	0.003	0.004	0.006	0.006
BALROG Uncertainty	None	<0.001	<0.001	<0.001	<0.001
Photometric Calibration Uncertainty	PIT	0.010	0.005	0.002	0.002
Inherent SOMPZ Method Uncertainty	PIT	0.003	0.003	0.003	0.003
Combined Uncertainty: SOMPZ (from 3SDIR)	-	0.012	0.008	0.006	0.009
Shot Noise & Sample Variance	3SDIR MFWZ	0.011	0.007	0.005	0.010
Combined Uncertainty: SOMPZ (from 3SDIR-MFWZ)	-	0.015	0.010	0.007	0.012
Combined Uncertainty: SOMPZ + WZ	-	0.016	0.012	0.006	0.015
Effective Combined Uncertainty: SOMPZ + WZ + Blending	-	0.018	0.015	0.011	0.017
Effective Combined Uncertainty: SOMPZ + WZ + SR + Blending	-	0.015	0.011	0.008	0.015

TABLE 3.1: Values of and approximate error contributions to the mean redshift of each tomographic bin at each stage of the analysis. We find that Sample Variance in the deep fields is the greatest contributor to our overall uncertainty for our fiducial result. The Shot Noise & Sample Variance term here is computed with the SPC sample. At low redshifts, the photometric calibration uncertainty is also significant, motivating improved work on the deep field photometric calibration. As expected, the uncertainty due to choice in Redshift Sample is a leading source of uncertainty for the third and fourth bins, motivating follow-up spectroscopic and narrow-band photometric observations. Note, the uncertainties combine non-linearly, so the combined uncertainties are not necessarily the quadrature sum of the contributing factors. Note, we label all results that incorporate blending as 'Effective' because we expect non-zero shifts on the mean redshift due to blending, but we do not expect non-zero shifts on the mean redshift between SOMPZ and WZ.

(iv) **Photometric Calibration Uncertainty**: uncertainty in the 8-band color of deep field galaxies, and the estimated photometric zero-points

(v) **BALROG uncertainty**: imperfections in the procedure of simulating the wide field photometry of deep field galaxies in BALROG for the computation of the transfer matrix

(vi) **SOMPZ Method Uncertainty**: uncertainty associated to the stochastic nature of the SOM training and bin assignment

Methods for propagating these uncertainties are presented in chapter 4, here we identify and quantify these uncertainties, with an emphasis on the estimation of sample variance and the procedure to quantify the SOMPZ uncertainty.

3.3.1 Redshift sample, photometric calibration and BALROG uncertainties

We briefly describe the other three sources of uncertainty and the methods to estimate them. More detail about these estimations are presented in section 5 of Myles et al. (2020), and the estimated error contributions are summarized in table 3.1.

Redshift sample uncertainty

The redshift sample employed to characterize the color-redshift discretization of the SOMPZ method comes from three independent sources:

- Spectroscopic surveys (S): zCOSMOS (Lilly et al., 2009), C3R2 (Masters et al., 2017, 2019), VVDS (Le Fèvre et al., 2013) and VIPERS (Scodeggio et al., 2018)
- COSMOS2015 (C): 30-band PZ catalogue (Laigle et al., 2016)
- PAUS+COSMOS (P): Combination of PAU Survey (Padilla et al., 2019; Eriksen et al., 2019) and COSMOS2015.

Each of these sources is characterized by its own selection criteria, z uncertainties, completeness. As a consequence, the obtained distribution can vary when using each sample independently in the SOMPZ scheme.

In order to quantify these variations, three combinations of sources for the redshift samples R are defined: SPC = all combined, PC = PAU+COSMOS, SC = Spectra + COSMOS. Each combination is assigned a probability $p(R)$ of being the correct combination, and the 3SDIR + HMC sampling is done accepting samples obtained from each combination R with that probability. The uncertainty is then computed by measuring the typical deviation in mean redshift from the $n(z)$ ensembles with respect to the global average, after assigning $p(R) = 1/3$ for all source combinations. The mean observed deviation from the average is $\Delta z_{\rm Redshift} \sim 0.003 - 0.006$ across the 4 tomographic bins. This uncertainty is naturally transferred into the final $n(z)$ ensemble.

Photometric calibration uncertainty

The deep field photometric observations are performed over 4 separate fields of the survey footprint, each with an independent solution of the photometric zero-point (See sections 3.2.1). The instability of the photometric solution from field to field can result in small variations of the fluxes used to compute the luptitudes and lupticolors, which result in incorrect assignment of galaxies to SOM cells.

To quantify the effect of these uncertainties, zero-points from each field are perturbed by a small magnitude shift $\Delta m_{\rm field}$ drawn from a Gaussian distribution. A $n(z)$ ensemble is then generated by applying the SOMPZ scheme to each realisation of the zero-point perturbed photometric catalogs. The observed deviation in mean redshift from the average of the ensemble are of the order of $\Delta z_{\rm ZP} \sim 0.002 - 0.010$ across the 4 tomographic bins. In order to transfer this measured uncertainty to the 3SDIR + HMC realisations, each realisations is shifted using a Probability Integral Transform (PIT) based on the ensemble of realisations generated by shifting the zero-points. This results in a recomputed set of 3SDIR + HMC realisations with the uncertainty already encoded on them.

BALROG uncertainty

The injection and re-detection of galaxies in the wide imaging using BALROG is used to estimate the transfer matrix $P(c|\hat{c})$ between cells of the deep and wide SOM (Everett et al., 2020). The determination of this transfer matrix is crucial to use the deep fields to connect the wide photometry to the redshift sample. Since the observing conditions across the entire footprint change significantly because of differences in depth, PSF estimation and

zero-points, it is expected that the transfer matrix of a sub-set of the footprint could vary from the average. In order to estimate effect of this variability, the transfer matrix is computed for subsets of the data covering contiguous patches of the footprint, using a bootstrap approach. Since the observed variability in the $n(z)$ ensemble obtained from the bootstrap transfer matrices $\Delta z_{\mathrm{BALROG}}$ is observed to be smaller than 10^{-3}, the effect of the BALROG uncertainty is neglected.

3.3.2 Sample variance and shot noise

A large uncertainty contributor in the DES Y3 $n(z)$ pipeline is expected to come from the sample variance associated with the limited coverage and completeness of the different samples employed in the SOMPZ scheme. Because of the small size of the redshift and deep samples, variation of the large scale matter distribution can result in the observed sample not being fully representative of the true underlying galaxy populations Nevertheless, these two samples provide a comparable amount of information to break the color-redshift degeneracy and the effect of sample variance can have a large impact on the $n(z)$ estimations. Common approaches to estimate these effects usually rely on bootstrapping (Repeating the $n(z)$ estimation on random subsamples of the data) or on scaling the uncertainty measured in multiple simulated datasets. The first approach can be inaccurate at measuring the effect of cosmic variance (Friedrich et al., 2016), while the second requires a careful fine-tuning of the simulation-to-data scaling of uncertainty.

The method to estimate sample variance in DES Y3 follows the formalism described in (Sánchez et al., 2020b), where an analytical model of sample variance is constructed using a *three-step Dirichlet* sampling procedure (3SDIR). This procedure discretizes the redshift range into bins z_i and estimates the probability $f_{z,c}$ that a galaxy in the deep samples has a redshift z and is assigned to SOM cell c such that $\sum f_{z,c} = 1$ and $f_{z,c} \in [0, 1)$. This means that the distribution of $f_{z,c}$ values follows a Dirichlet distribution

$$\mathrm{Dir}(\{f_{z,c}\}; \{\alpha_{z,c}\}) \propto \prod f_{z,c}^{\alpha_{z,c}-1} \tag{3.11}$$

If the redshift sample used to inform the color-redshift relation of deep sample was completely representative of the deep sample then $f_{z,c}$ would be proportional to $N_{z,c}$, the number of galaxies from the redshift sample assigned to a two dimensional histogram in z, c space. Given that the redshift sample used to describe the color-redshift relation is only finite, and potentially correlated with large scale structure because of cosmic variance, the $\alpha_{z,c}$ associated to each bin are transformed in a way that they match the observed variance estimated from large N-body simulations.

To validate this method and to simulate the effect of sample variance on the determination of the redshift distribution of the wide sample, 300 SOMPZ $n(z)$ realisations of a fixed simulated catalog were generated, using randomized mock deep fields and redshift catalogs from the BUZZARD simulations (Buchs et al., 2019; DeRose et al., 2019). This estimate of sample variance, typically quantified by the width of the distribution of moments of the $n(z)$, can be then compared to the obtained values using the 3SDIR approach. The

simulated catalogues used for these realisations, which we will describe in detail in chapter 5, are built to match the properties of the DES survey and provide true redshifts and magnitudes and their respective realistic errors in the same bands of the deep and wide fields.

To generate the mock deep fields, 3 square regions of 3.32, 3.29 and 1.94 square degrees respectively are cut from the simulation footprints and used as deep data. 100,000 true redshifts are drawn from a 1.38 $\deg^2$ square region from the simulations. The conditions for generating these 4 regions is that they are not overlapping between them or with the survey edge. In order to simulate the BALROG catalogs, photometric noise is added to the deep galaxies in order to match the expected photometry errors of the wide sample and then the relative assignment to the deep and wide SOM is recorded. This procedure is repeated with multiple realisations of the noise model to better constrain the transition matrix $P(c|\hat{c})$. In all of these realisations, the training and assignment of galaxies from the wide sample is kept fixed to minimize the effects associated to the random nature of the training stage. Figure 3.9 shows the obtained ensemble of realisations, where it can be seen the large uncertainty associated to the histogram values of the realisations coming exclusively from cosmic variance. These peculiarities can potentially result in differences in the inferred cosmology if they are not accounted for. We discuss this in detail in the remaining chapters, where we use this ensemble of realisations to test the effect of peculiarities and other high order uncertainties in the inferred cosmological parameters. Regardless of this apparent variability, the average distribution (dashed color lines) very closely resemble the truth distributions from the BUZZARD samples (solid color lines) which serves as a powerful re-validation of the SOMPZ scheme, previously tested in Buchs et al. (2019).

The equivalent $\Delta z_{\rm sv+sn}$ error, which is the standard deviation of the difference in mean redshift from each tomographic bin with respect to the average from the generated $n(z)$ ensemble, is around 0.06 for each tomographic bin, and in most cases is the main contributor to the total uncertainty.

3.3.3 SOMPZ method uncertainty

The SOMPZ scheme described above uses an unsupervised machine learning technique to obtain a direct mapping from input space to a low-dimensional representation which can be used to easily visualize the arrangement of the unsupervised quantities. Because of the number of moving parts in this scheme the final mapping obtained can be expected to change as a consequence of the stochastic nature of the training stage. The weight vectors are initialized at random, either from a sample of points from the training set, or at random positions on a predefined region. Even if the sequence of samples on the iterative process described in equation 3.5 was exactly the same, the randomized initial positions plus the decaying learning rate $a(t)$ would result in different final positions for the weight vectors. This difference is amplified once the randomized sampling comes into play.

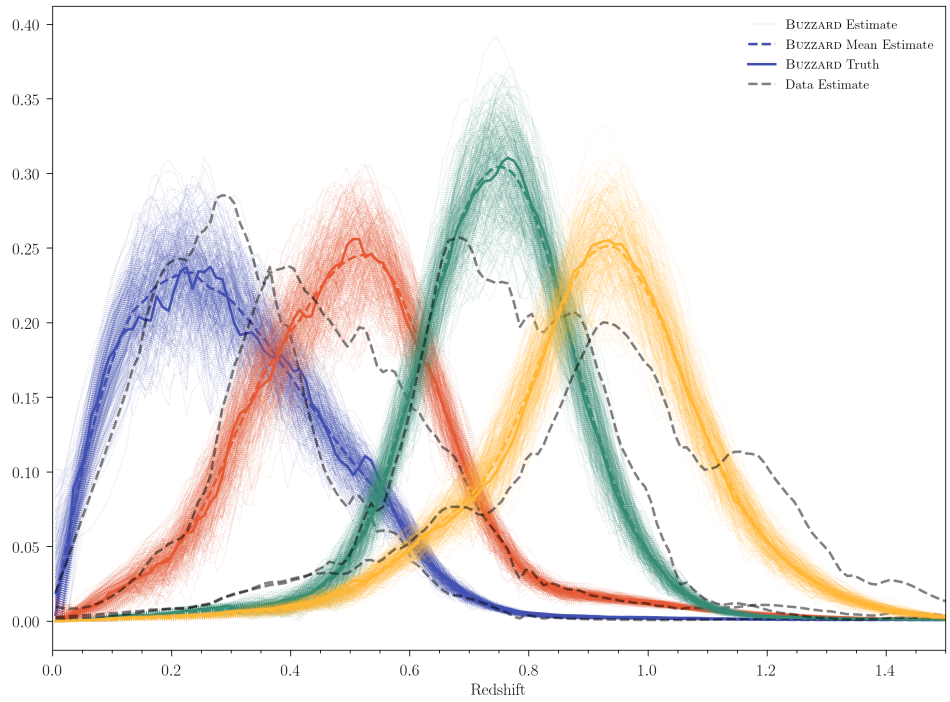

FIGURE 3.9: Estimated $n(z)$ in four tomographic bins from the BUZZARD simulations using an ensemble of 300 different sets of deep fields on the BUZZARD sky (colorful fine dashed lines). The similarity of the mean of the estimated $n(z)$ (colorful broad dashed lines) relative to the truth (color broad solid lines) is a basic illustrative validation of the method. The Redshift Sample used here has 100000 galaxies drawn from 1.38 deg^2, the Deep Sample in each realisation is drawn from three fields of size 3.32, 3.29, and 1.94 deg^2, respectively from the BUZZARD simulated sky catalogue. The variation in estimated $n(z)$ reflects the uncertainty of the SOMPZ method primarily due to sample variance in the deep fields.

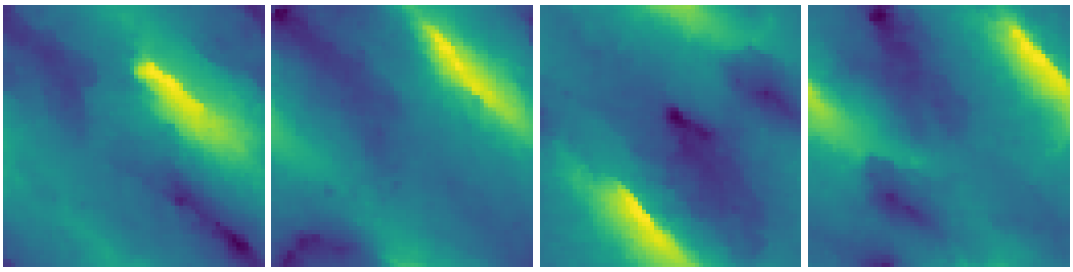

FIGURE 3.10: Four different self-organized map arrangements showing the distribution of the first component of the weight vectors of each node. The arrangements are built using the same input dataset and the same dimensionality for the SOM map, only changing the random seed used for the iterative training process. While the structure show some resemblance with each other, they are not identical, which can lead to a different assignment of the input sample and potentially different $n(z)$ distributions.

The effects of different weight vector positions becomes evident in the assignment phase where objects from the target sample are assigned to the closest node in input space according to equation 3.4. Two samples with similar photometric properties, once assigned to the same cell c can end in different cells depending on the relative positions of their closest weight vectors. Following the bin assignment procedure described in 3.2.3, this in turn can result in objects being assigned to different tomographic bins in extreme cases. An example of this effect can be seen in figure 3.10 where different arrangements of the first component values of the weight vectors assigned to each node are shown for four different SOM training of the same dataset.

We perform a simple test to estimate the effects of this random sampling can be constructed by fixing the wide, deep, BALROG and redshift samples and training the deep and wide SOM maps with different random seeds each time. This will result in a different map and assignment each time and potentially different redshift distributions. The width of the distribution of mean redshifts from an ensemble of $n(z)$ samples can be used as a quantifier of the total uncertainty, similarly to the characterisation of sample variance. We run the SOM training 100 times and compute the $n(z)$ for a fixed sample of BUZZARD galaxies using the SOMPZ scheme described above, only changing the initial conditions for the positions of the weight vectors and the random seed for the training sampling. This results from an ensemble of 100 $n(z)$ realisations shown in figure 3.11 from where the uncertainty contribution is estimated to be approximately $\Delta z_{\mathrm{SOMPZ}} \sim 0.003$ in all tomographic bins. In figure 3.12 it can be seen this is a significant source of uncertainty, comparable to the other sources mentioned in this section.

3.3.4 Summary

In this chapter we have described the fundamentals of photometric redshift estimations and their importance in weak lensing experiments. We also showed the effect on cosmological parameter inference of biases in the determination of the line-of-sight distribution of sources.

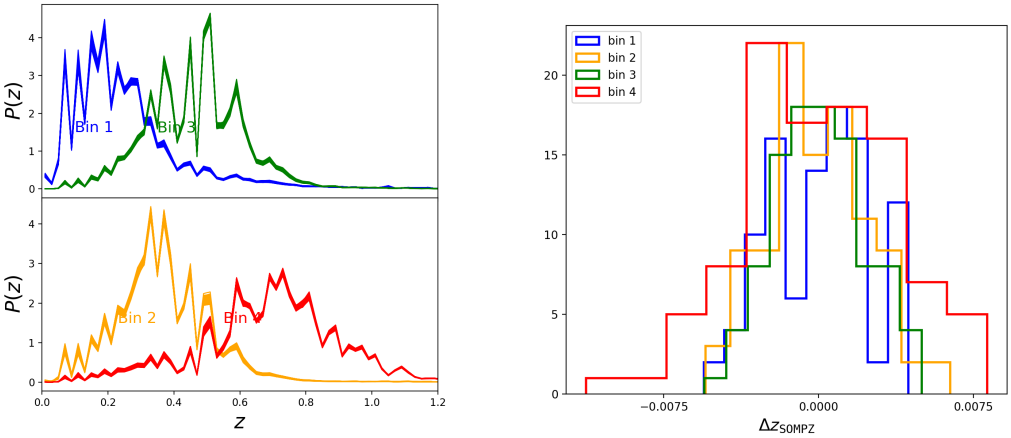

FIGURE 3.11: Left: Ensemble of redshift distributions obtained using the SOMPZ scheme, where each realisation is obtained by training the deep SOM using a different initial random seed.
Right: Distribution of $\Delta z_{\rm SOMPZ}$ values for the ensemble of realisations shown on the left panel.

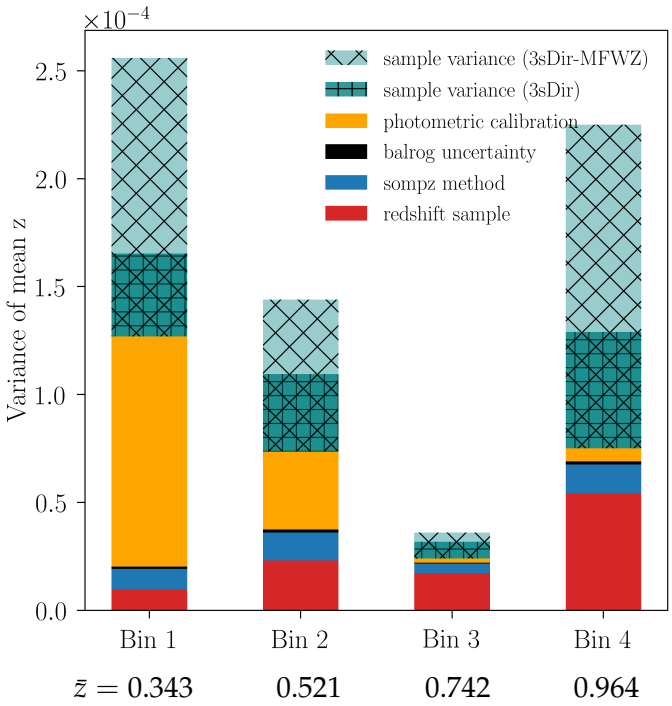

FIGURE 3.12: Variance of each source of uncertainty in each tomographic bin. As shown here the redshift sample uncertainty becomes a larger contributor to the uncertainty for higher redshift tomographic bins. Note, the contributing sources of uncertainty combine non-linearly. As a result, to illustrate the relative magnitude of each source of uncertainty in each bin, and the relative importance of each contributing source of uncertainty as a function of redshift, total variance in this figure is rescaled to match the combined uncertainty shown in Table 3.1). The small uncertainty associated to tomographic bin 3 can be attributed to the fact that galaxies assigned to that bin span the narrowest range ($0.631 \leq z \leq 0.872$), see Table 3.1. Since the bin widths are set so all bins have roughly the same number of galaxies, this redshift range has the highest density of sources per redshift range.
Figure credit: Myles et al. (2020)

Within this context, we introduced the Dark Energy Survey Y3 analysis pipeline to estimate the distribution of galaxies from measurements of its photometry in multiple bands. This pipeline employs a novel machine learning technique based on Self-organizing maps, SOMPZ, which takes advantage of deep photometry obtained for a subset of the data to build a transition matrix which helps breaking the color-redshift degeneracy. In addition to SOMPZ, the DES Y3 analysis includes the use of angular correlations with positions and shears of a secure redshift sample to further constrain the distribution of sources. These constrains are combined into a three-step Dirichlet and Hamiltonian Monte Carlo sampling scheme to provide samples from the redshift distribution posterior.

Finally, we described the identified sources of uncertainty caused by systematic effects associated to the SOMPZ scheme and the different measurements employed in the process: sample variance and shot noise, non-uniformity of the redshift sample, photometric calibration errors, imperfections in the BALROG simulations and stochastic uncertainty of the SOMPZ method. We also described the procedures to quantify each of these uncertainties, and presented the results for sample variance and the SOMPZ stochastic uncertainty, in which we directly contributed.

The 3SDIR + HMC samples obtained in this process encode the full uncertainty associated to the redshift distribution estimation pipeline. In chapter 4 we present HYPERRANK, a scheme to directly propagate the uncertainties mentioned above into the cosmological parameters.

Chapter 4

The HYPERRANK algorithm

In chapter 2 we described the formalism for weak lensing experiments and the potential source of systematic errors that can bias cosmology, both associated with the observables of weak lensing and the processing of data in the pipeline. We also described in detail the process to obtain the line of sight distribution of sources required for this analysis using a novel technique, called SOMPZ. In this chapter we introduce HYPERRANK, a novel technique to propagate the uncertainties arising from photometric redshift estimation into the cosmological parameters using the products obtained from the SOMPZ scheme. To put this technique in context, we first describe the process of parameter estimation using Bayesian Statistics and the use of stochastic sampling techniques, specially nested sampling algorithms. The main motivation behind HYPERRANK is to provide a scheme to propagate higher order descriptions of uncertainty into the inferred cosmological parameters. We explain what aspects of the scheme can have an impact on sampling efficiency and how these can be mitigated.

The methodology and validation tests presented here have been submitted for publication in Cordero et al. (2020) as part of the DES Y3 analysis, and the description of tests and results from applying this methodology to simulations and data are presented in chapters 5 and 6 respectively. We describe the Bayesian parameter inference formalism and the concept of uncertainty propagation in 4.1 and then present the HYPERRANK method in 4.2

The author of this thesis designed the HYPERRANK formalism described in this chapter, and implemented it as a module in CosmoSIS (Zuntz et al., 2015), including the one- and multi-dimensional approaches shown in sections 4.2.1 and 4.2.2, and the alternative ranking schemes using the solution to the linear sum assignment and the uneven grid approach.

4.1 Bayesian Parameter Inference

The scientific method provides us with a powerful tool to develop knowledge by testing hypothesis based on observations of different phenomena in physics. By the constant repetition and accumulation of results from multiple experiments we are able to predict

the behaviour of certain physical systems by summarizing their properties using models. These models typically describe these behaviours in a mathematical language which not only makes it easy to concatenate to other models for other physical phenomena, but also provides a numerical context to understand the relative values and limits of these models relative to already established ones. Determination of these parameters requires direct observation of the physical behaviour of the systems and measurements of *observables*, by means of an *experiment*. These observables are manifestations of the properties of these systems and must be carefully defined for them to usefully contribute to the determination of the model parameters. But given the practical and physical limitations in the setup of an experiment and observation, our interpretation of the results is never that of certainty, but reliant on the concept of probability.

There are two main interpretations of what a probability is: The *frequentist* interpretation assumes that a probability is a quantification of the frequency a result will occur after an infinitely large number of repetitions of an experiment or observation. The assumption is that each repetition is independent from all others and it is performed under similar conditions. The Bayesian interpretation, in contrast, defines probability as a quantification of the degree of certainty of the results of an experiment, constructed based on fundamental properties of the experiment:

- The particular set of observations defined to extract the properties of the model
- Pre-defined knowledge about the possible results
- Identified sources of systematic and random errors which can contaminate observations
- Metrics used to compare the observed results against their expected values.

There is an immediate restriction which makes the frequentist approach unsuitable for cosmological experiments like weak lensing: repeating an experiment *an infinitely large number of times*, or even more than once under similar conditions is impractical because of the sheer scale of the experiment (one purposely build CCD camera, 6 years of observations, 200+ collaborators, etc), and the impossibility to have independent realisations since we only have one Universe to draw observations from. The Bayesian approach fits in nicely because it gives us an interpretation of probability which relies on a description of our own limitations and expectations, which is certainly more practical. In what follows, we describe the foundations of Bayesian inference and how it is applied to cosmological analysis using today's computational tools.

4.1.1 Bayes' theorem

Bayes' theorem quantifies the probability of a set of parameters θ being a good description of the model M, given a set of observations $\mathbf{x}$ as $P(\theta|\mathbf{x}, M)$, called the *posterior*:

$$P(\theta|\mathbf{x}, M) = \frac{\mathcal{L}(\mathbf{x}|\theta, M)\pi(\theta|M)}{p(\mathbf{x}|M)}. \tag{4.1}$$

$\mathcal{L}(\mathbf{x}|\theta, M)$ defines the parameter *likelihood*, which is a quantification of the similarity or concordance between the observables and their expected values for a set of parameters θ under model M. This concordance usually factors in the associated measurement errors on the observations. A common approach is to assume those errors are Gaussian-distributed. Under that assumption, we define the Gaussian likelihood $\mathcal{L}(\theta)$ as:

$$\mathcal{L}(\mathbf{x}|\theta, M) = P(\mathbf{x}|\theta, M) = \frac{1}{(2\pi)^{m/2}|\mathbf{C}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_t)^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{x}_t)\right], \tag{4.2}$$

where $\mathbf{C}$ is the covariance matrix describing the amplitude of the expected errors and their correlations with the observations of $\mathbf{x}$, m is the dimensionality of the observed data vector, and $\mathbf{x}_t$ is the expected vector obtained at a the model parameters θ. $\pi(\theta|M)$ is the parameter *prior*, which quantifies the previous expectation about the values of the parameters. Its definition can have physical motivations coming directly from limits characteristic of the employed model, or encode the subjective description of parameter restrictions, which can be based, for instance, on observations made by independent previous experiments. $P(\mathbf{x}|M)$ is the *evidence*, denoted $\mathcal{Z}$ which quantifies to total contribution by the data to constrain the model, and it is of central importance when using Bayesian inference to perform *model selection*. Because of its independence to the model, it can be interpreted as a factor required to normalize the posterior over the model parameters θ, such that

$$\mathcal{Z} = \int \mathcal{L}(\theta)\pi(\theta)d\theta. \tag{4.3}$$

This description helps understanding the notion of evidence as that of the goodness of the fit to a model, since it is the average of the likelihood over the prior, which grows for a more compact parameter space (represented by the prior $\pi(\theta)$), or when the data and model agree well (represented by the likelihood $\mathcal{L}(\theta)$). The ratio of the evidence between two models, called the *Bayes Factor*, can help comparing two models given an observed dataset x and their respective likelihoods.

We saw in section 1.1 that the standard cosmological model can be described by a small set of parameters, some of which are heavily constrained by cosmic shear and galaxy clustering. Moreover, additional parameters are used to describe different sources of systematics or model extensions, all of which are usually inferred simultaneously in the cosmological inference pipeline. The result is a multivariate probability distribution which we want to use to infer statistical properties of the model parameters, and need to integrate to obtain the evidence $\mathcal{Z}$. The naïve approach to evaluate these functions would be to discretize the space of parameters and ask a computer to evaluate the function on each of the grid points. For a cosmic shear experiment this is typically of the order of 20 parameters to be simultaneously sampled. As an example, the likelihood pipeline employed for DES typically takes around $\sim$6 seconds. For a **very** coarse grid of 10 points along each dimension, 10^{20} points would require 19.7 trillion CPU-**years**, which in Cori[1], one of the largest HPC computers in the US, would take approximately 298 millions years

[1] https://docs.nersc.gov/systems/cori/

assuming we are allocated all the nodes. And that the SSH link does not break.

In order to efficiently obtain samples describing the posterior or to calculate the evidence of a set of observations for a given model, which are both of very high dimensionality in the context of cosmology with weak lensing and galaxy clustering, we often employ a family of techniques, called Monte Carlo Markov Chain (MCMC) methods, which rely on efficient stochastic sampling of the space of parameters to map the posterior.

4.1.2 Monte Carlo Markov Chain methods

Monte Carlo methods refers to a broad family of mathematical and computational techniques which involve the use of pseudo-randomly[2] generated numbers to find the solutions to problems where deterministic approaches cannot be applied. The main idea behind these methods is that the results are computed based on statistical analysis of the random samples rather than via an attempt to do it analytically and exactly. In the context of probability distribution characterisation and multidimensional integration, which are the two main goals of Bayesian inference, the proposal of new samples and its acceptance/rejection are adjusted in a way that over long chains of randomly generated numbers and evaluations of the integrand, the statistical properties of the samples approach those of the true underlying distribution. These *adjustments* define a family of methods called Markov Chain Monte Carlo, where the distribution of generated samples converge to the underlying probability distribution. The general principle of these methods is based on the fact that the acceptance probability is equal to the ratio between the posterior values of the current value of the chain and the proposed value. This way all points from the parameter space have a probability to be sampled, which is proportional to its posterior value: for a sufficiently long chain, the multidimensional histogram of points is proportional to the posterior, which removes the need to resort to the actual values of the posterior to perform a statistical analysis of the parameters.

When dealing with parameter estimation, the most common methods all work based on a similar principle, on which samples are proposed from a specific probability distribution and accepted or rejected according to properties of the underlying target posterior distribution. The **Metropolis-Hastings** algorithm (Metropolis et al., 1953; Hastings, 1970) for example, proposes values drawn from a generic distribution conditional on the current value of the chain. A special case, the **Metropolis algorithm** defines the distribution as a symmetric Gaussian characterized by a covariance matrix which strongly determines the general efficiency of the methods. A variant of this method, called **Hamiltonian Monte Carlo Sampling** (Duane et al., 1987) uses a physical argument to propose samples assuming their posterior value is analogous to a *potential energy* and their *kinetic* energy is drawn from a known normal distribution. Using the Canonical probability distribution for energy levels of a system and the Hamilton equations of motions, it provides samples from the posterior which can be then accepted or rejected following a similar criteria to

[2]Every computational method to generate numbers is ultimately deterministic. Pseudo random number generator only approximate the properties of truly random numbers, such as having very long periods of repetition or low correlation between successive samples.

that of Metropolis Hastings. If the posterior distribution is unknown but the conditional probability of the model parameters are known or can be sampled, then a Markov chain of samples can be obtained using an iterative process called **Gibbs sampling** (Geman & Geman, 1984).

For the case of evidence evaluation, perhaps the most popular approach related to MCMC algorithms is **Thermodynamic integration** (Kirkwood, 1935), by which the logarithm of the Bayes factor for an unknown distribution with respect a known one is estimated by running a series of MCMC chains at regular intervals, defining a series of continuous variations of the proposal distribution. This method is effective but very computationally intensive, as it essentially means running a separate independent MCMC several times.

Most MCMC methods mentioned above have some serious disadvantages which become more noticeable at higher dimensions. Most notably, they require fine-tuning on some of the parameters that control the acceptance / rejection probabilities and the distributions used to propose new samples. They also either fail to provide a reliable formalism to estimate the errors on the computation of the evidence, or require large number of likelihood evaluations or realisations of the computation. A family of techniques which overcome some of these difficulties are called **Nested sampling** algorithms, which provide an improved error calculation of evidences and produce samples of the posterior as a by-product.

4.1.3 Nested Sampling

Nested sampling techniques, initially developed by Skilling (2004), are primarily targeted at efficient computation of the evidence by exploiting the relation between the likelihood and prior volume to transform the multidimensional evidence integral (4.3) into a one-dimensional integral. It starts by defining the prior volume X enclosed by an iso-likelihood surface as

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} \pi(\theta) d\theta. \tag{4.4}$$

Since the prior is a probability distribution, this function goes from 1 when considering the full prior volume, to 0 when λ reaches the maximum likelihood value for the posterior. Equation 4.3 can be rewritten in terms of a one-dimensional value:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX, \tag{4.5}$$

where $\mathcal{L}(X)$ is the inverse of equation 4.4, the likelihood value of a surface enclosing a prior volume X- Nested sampling finds an approximation of the evidence by finding samples of nested iso-likelihood surfaces $\mathcal{L}(X_i)$ (Hence the name nested sampling), and summing them using

$$\mathcal{Z} \approx \sum_{i=1}^{M} w_i \mathcal{L}(X_i), \tag{4.6}$$

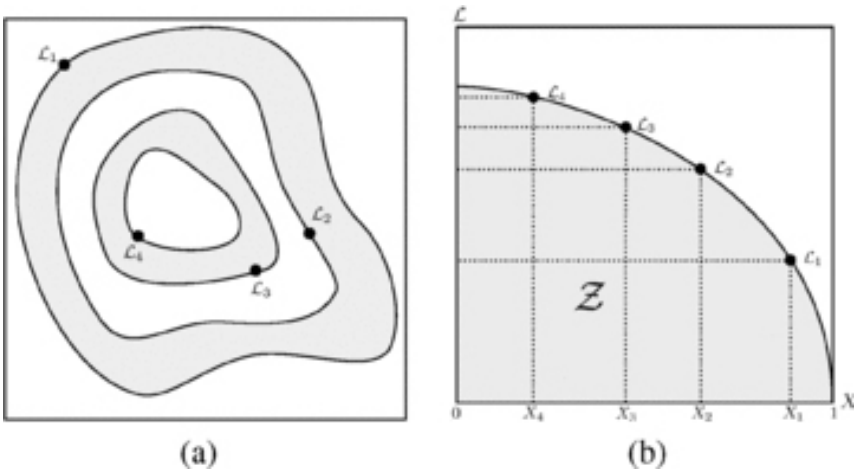

FIGURE 4.1: Graphical representation of the nested sampling algorithm, showing (a) the nested iso-likelihood surfaces progressively enclosing areas of higher likelihood from where new points can be added to the chain, and (b) the corresponding prior volumes enclosed by each iso-likelihood surface. The area under the curve corresponds to the evidence which can be integrated using quadrature methods.
Image credit: Skilling (2004)

where w_i is a weight based on the quadrature used for integration. A graphical representation of this concept is shown in figure 4.1. In order to obtain the samples, the procedure starts by drawing N points, called *live points* from the prior, and ranking them according to their likelihood values. The point in parameter space θ corresponding to the smallest likelihood is stored and replaced by a random new point, provided it has a larger likelihood value than the point to be replaced. After the new sample is selected, the replaced point is added to the chain and assigned a weight

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}} \tag{4.7}$$

which can then be used to characterize the posterior distribution if desired. One important problem with this approach is that as the volume enclosed by iso-likelihood surfaces decreases the fraction of samples from the prior which have likelihood values higher than $\mathcal{L}_i$ also decreases steadily as i grows. The iso-likelihood surface is never actually determined, and only enforced by the evaluation of the likelihood for a sample. **Ellipsoidal nested sampling** attempts to overcome this limitation by approximating the iso-likelihood surface as an ellipsoid whose axis are defined by the covariance matrix of the live points. Sampled points are first required to lie inside this ellipsoid, and then its likelihood is computed, significantly reducing the number of times a low likelihood sample is tested. MULTINEST (Feroz et al., 2009) proposes a more sophisticated nested sampling based on the concept of ellipsoidal nested sampling to overcome further limitations of the standard approach when the posterior is not unimodal or has pronounced banana-shaped degeneracies, by identifying these modes and/or separating the posterior into multiple subregions bounded by ellipsoid approximations of the iso-likelihood surface.

4.1.4 Sampling efficiency

We want our Monte-Carlo samplers to provide samples from the posterior both *accurately* and *efficiently*. Accurately in that their distribution closely resembles that of the underlying true distribution, and efficiently in that it gives a decent approximation in as few steps as possible. Typically these two requirements are in direct contradiction with each other: for a basic Metropolis algorithm, a narrow Gaussian proposal around the current step of the chain will result in a large rate of rejection but will ensure the samples actually map the true distribution. The opposite will result in samples being accepted despite their low likelihood, resulting on a distribution completely prior dominated. A similar behaviour occurs with nested samplers, where the proposals are accepted based on their likelihood being larger to that of the current live point. In order to correctly map odd-shaped non ellipsoidal posteriors and degeneracies between parameters, the ellipsoidal regions from where the trial samples are drawn is slightly expanded around the iso-likelihood surface $\mathcal{L}(X_i)$. This results in a slightly slower convergence towards high likelihood regions as the effective volume to sample is increased, but more accurate posterior samples as those regions are allowed to be explored.

Another important aspect of the sampling efficiency comes from the fact that most sampling techniques are designed under the assumption the posterior they are sampling is relatively smooth with respect to the typical distance between the trial samples and the current steps of the chain. This way, relatively small steps in parameter space result in controlled small increases in likelihood in a consistent way. If the posterior is not smooth, this can result in the sampler identifying only local peaks or difficulties in evaluating the conditions for convergence. An example of this is shown in figure 4.2 where the nested sampler must find a parameter θ with likelihood $\mathcal{L}(\theta) > \mathcal{L}^*$ for two posterior distributions with different smoothness. Since the probability of finding such value θ is proportional to the fraction of θ ranges where $\mathcal{L}(\theta) > \mathcal{L}^*$, the sampling of the smooth posterior will be significantly more efficient that the non-smooth case.

4.1.5 Propagation of uncertainty

In section 4.1 we described the process by which many modern scientific experiments evaluate their data and prior knowledge to constrain the characteristics of their models using Bayesian inference. The end result is often a multidimensional probability distribution encoding the level of confidence in a combination of parameters correctly describes the model, based on a series of observations and accompanied by previous information regarding the nature of these parameters. While we are mainly interested in the individual values of these parameters, these are not only correlated between them, but also with other less interesting parameters we have used to describe aspects like model extensions or systematic errors. The process by which we isolate the parameters of interest is called *marginalisation*, and refers to the process of finding the marginal probability distribution for the set of parameters of interest. Given a multivariate probability distribution $P(\boldsymbol{\theta})$, where theta is a vector of parameters, $\boldsymbol{\theta} = \boldsymbol{\alpha} \cup \boldsymbol{\beta} = \{\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m\}$, we can find

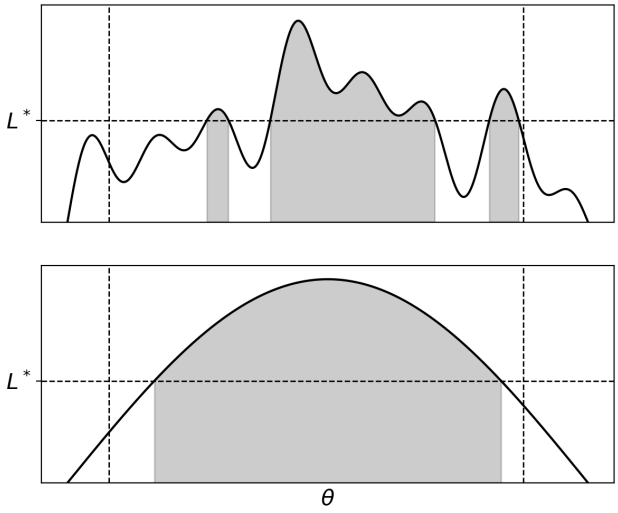

FIGURE 4.2: Toy example of showing the differences in sampling efficiency between a smooth (lower panel) and non-smooth posterior (upper panel). Horizontal dashed lines represent the iso-likelihood surfaces and likelihood thresholds $\mathcal{L}^*$ above which the proposal values $\mathcal{L}(\theta)$ are accepted into the chain. The vertical dashed lines define the elliptical boundaries (intervals in 1D) from where parameters can be initially sampled, including additional allowed areas on the edges to account for non elliptical posteriors. The shaded regions denote all θ intervals from where samples would be accepted, showing clearly the lower probability in the non-smooth case.

the joint distribution of a subset of parameters $\{\boldsymbol{\alpha} = \alpha_1^*, \ldots, \alpha_n^*\}$ marginalised over the subset of parameters $\{\boldsymbol{\beta} = \beta_1^*, \ldots, \beta_m^*\}$

$$P(\vec{\alpha}) = \int_{\Omega_\beta} P(\vec{\theta})\, d\beta_1 \ldots d\beta_m \tag{4.8}$$

This process accounts for all the possible effects the marginalised parameters can have over the target parameter distribution, and it returns a distribution which summarizes their effect regardless of the correlation between the parameters.

As mentioned in section 4.1.2, in practice the Bayesian inference techniques do not provide a functional form for these distributions, but samples from the posterior probability distribution, from where the marginal probability distribution can be approximated.

4.2 Hyperrank

In previous chapters we showed how the biases on the different measurements used in weak lensing and galaxy clustering can result in significant shifts on the cosmological parameter confidence contours. In section 4.1.5 we learned that in order to correctly account for the uncertainty associated to these measurements we must marginalize the posterior

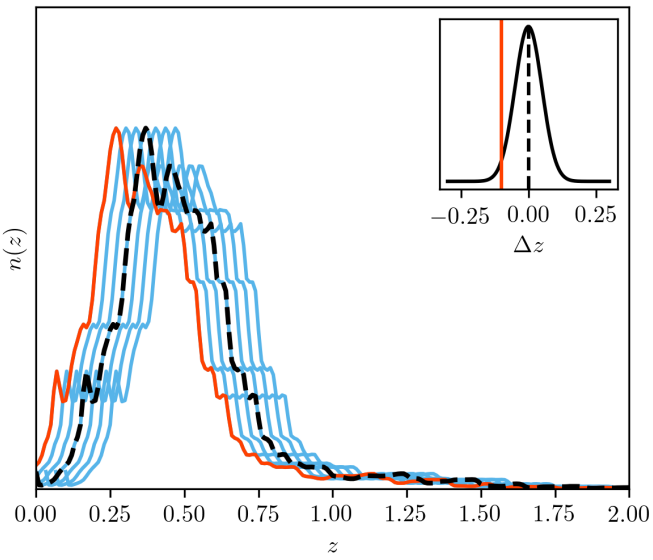

FIGURE 4.3: Graphic representation of the Δz marginalisation scheme, where a fiducial redshift distribution (black dashed) is shifted horizontally at each Monte Carlo step by a value drawn from a Gaussian distribution (inset, with draw from the 2σ tail highlighted in red). Any higher order moment and peculiarity of the distribution remains unchanged, which can lead to an underestimation of the total cosmological parameter uncertainty.

by integrating all the potential effects of these systematics can have on the parameters of interest.

An important aspect of this is the identification of the sources of uncertainty and what is their effect on the measurements used in the analysis. While these dependencies are typically intractable, simplified models can be made to describe their effect. One of these approximations is that uncertainty on the determination of photometric redshifts can be described by a set of nuisance parameters describing a shift on the mean of the redshift distributions, Δz_i:

$$n_i(z) \rightarrow n_i(z - \Delta z_i). \tag{4.9}$$

The motivation for this approximation, depicted in figure 4.3 is that uncertainty in the mean of the redshift distribution $n(z)$ have been shown to be the main contributors to the uncertainty of inferred values for σ_8.

Future large surveys will depend more on the precise determination of systematics and the effect of systematics on higher order moments than the mean may become important to keep under control. Unfortunately a simultaneous parametric description of the uncertainty for multiple high order moments is not possible, since the skewness and kurtosis of a distribution are highly degenerate with its mean and standard deviation, even for a well defined Gaussian distribution.

Here we present HYPERRANK, an alternative approach to marginalize over redshift systematics in which, instead of propagating the uncertainty using a set of parameters characterizing the statistical properties of the $n(z)$ posterior, we directly use an ensemble of samples of this posterior by choosing a new redshift distribution in each likelihood evaluation. The main concern is the potential loss in sampling efficiency as the finite

size of the ensemble can result in an uneven exploration of the posterior distribution. As described in section 4.1.2, a smooth posterior as a function of the sampled parameters results in a faster convergence, specially in nested samplers which rely on shrinking the sampled parameter space.

Once a discrete set of realisations for $n_i(z)$ have been generated using a method such as the DES Y3 approach described in section 3.2.1, we wish to use these realisations in the Monte Carlo algorithms used to generate samples from the joint posterior of cosmological and nuisance parameters, which we then use to do inference. Here, instead of modifying a fiducial distribution according to a nuisance parameter like the one in equation 4.9, the sampling strategy chooses a completely new distribution from the set of realisations to evaluate the likelihood. We introduce the idea of HYPERRANK-ing in which the full set of realisations is mapped onto a small (~ 1) number of rank parameters $\mathcal{H}$, which are *a priori* expected to correlate strongly with values of the cosmological parameters of interest. The realisations are then placed in rank order according to a set of descriptive values $\mathbf{d}$, and the rank parameters $\mathcal{H}$ become the nuisance parameters which are sampled (and subsequently marginalised over) in the cosmological analysis. The descriptive values $\mathbf{d}$ are chosen to allow the likelihood to vary as smoothly as possible along each of the rank parameters. Also, the ordering must be such that realisations with similar descriptive values are mapped close to each other and all realisation have the same volume of $\mathbf{d}$ parameter space assigned to them to avoid the introduction of an implicit prior.

4.2.1 One dimensional case

We initially consider the case in which a single HYPERRANK parameter is used to rank all realisations. Since realisations are comprised of a fixed combination of tomographic bins, we consider a basic approach which maps distributions based on the weighted average of a combination of values describing each tomographic bin,

$$\mathbf{d} = \frac{\sum w_i d_i}{\sum w_i}, \tag{4.10}$$

where i is the index of each tomographic bin and w_i is a weight, which can be defined based on number of assigned galaxies to each tomographic bin, or their relative contributions in the likelihood computation, and d_i is a descriptive value for each tomographic redshift distribution. The realisations are then ranked according to their descriptive value $\mathbf{d}$ and mapped to a continuous hyper-parameter $\mathcal{H} \in [0, 1)$, which is sampled by the MC on each likelihood evaluation. For a set of N_p ranked proposal $n(z)$, the rank of the realisation to be used in the likelihood evaluation is

$$rank = \lfloor \mathcal{H} \times N_p \rfloor, \tag{4.11}$$

where the brackets are the floor function. This is demonstrated in figure 4.4 which shows a small sample of realisations color-coded by their mean redshift and assigned a range of $\mathcal{H}$ values depending on their ranked position.

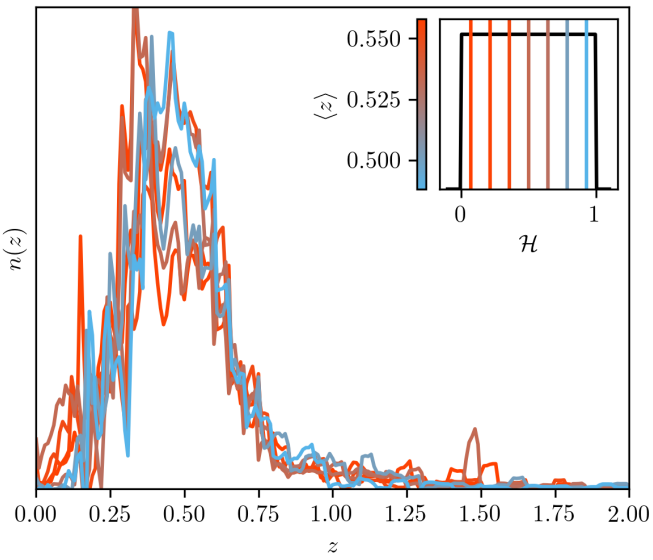

FIGURE 4.4: Discrete realisations of possible $n(z)$ are shown with colours corresponding to the mean redshift of each realisation $\langle z \rangle$, which can be mapped to a ranking hyper-parameter $\mathcal{H}$ which is then marginalised over on the Monte Carlo chain. Inset shows the uniform distribution for $\mathcal{H}$ which is sampled from, and the centres of the regions corresponding to each coloured $n(z)$ realisation.

It has been shown (Huterer et al., 2006; Ma et al., 2006) that the mean of the redshift distributions have the biggest impact in derived cosmology, so a good first choice for d_i is the mean redshift,

$$\langle z_i \rangle = \int_{z_{\min}}^{z_{\max}} z\, n_i(z) dz. \tag{4.12}$$

An alternative set of descriptive values are the mean inverse comoving distance of sources, $\langle 1/\chi \rangle$, which is expected to be more closely correlated with the eventual posterior value calculated by the analysis pipeline, since it is closely related to the shape of the lensing efficiency functions used in the Born and Limber approximations, which can be written as,

$$P_\kappa(\ell) = \frac{9H_0^4\Omega_m^2}{4c^4} \int_0^{\chi_{\rm H}} g^2(\chi) \frac{P_\delta(\ell/\chi;\chi)}{a^2(\chi)} \,\mathrm{d}\chi\,, \tag{4.13}$$

where $\chi_{\rm H}$, $a(\chi)$ and P_δ are the comoving horizon, scale factor and matter power spectrum, respectively and the lensing efficiency $g(\chi)$ at comoving distance χ is defined as:

$$g(\chi) = \int_\chi^{\chi_{\rm H}} n(\chi') \frac{\chi' - \chi}{\chi'} \,\mathrm{d}\chi'\,, \tag{4.14}$$

and depends on the comoving distance distribution $n(\chi)$ of sources, or equivalently their redshift distribution $n(z)$. By evaluating at $\chi = 0$ and differentiating the above definition for the lensing efficiency we obtain

$$g(\chi)|_{\chi=0} = 1 \tag{4.15}$$

$$g'(\chi)|_{\chi=0} = -\langle 1/\chi \rangle_n, \tag{4.16}$$

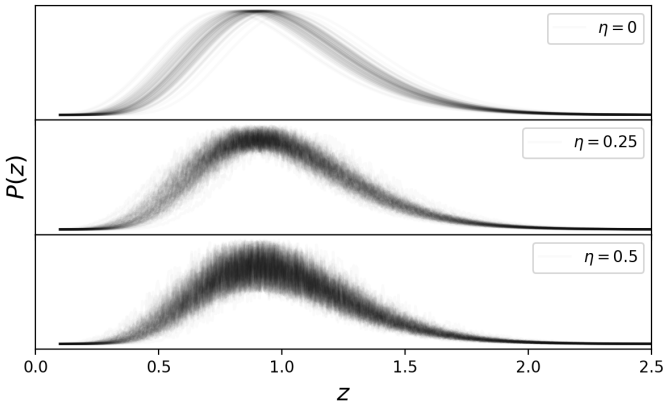

FIGURE 4.5: Three sets of 100 gamma probability distributions, obtained from shifting a fiducial distribution along the horizontal axis with a value Δ_z drawn from a Gaussian probability distribution. From top to bottom the distributions are randomly scaled at different z values to simulate increasing effects of noise.

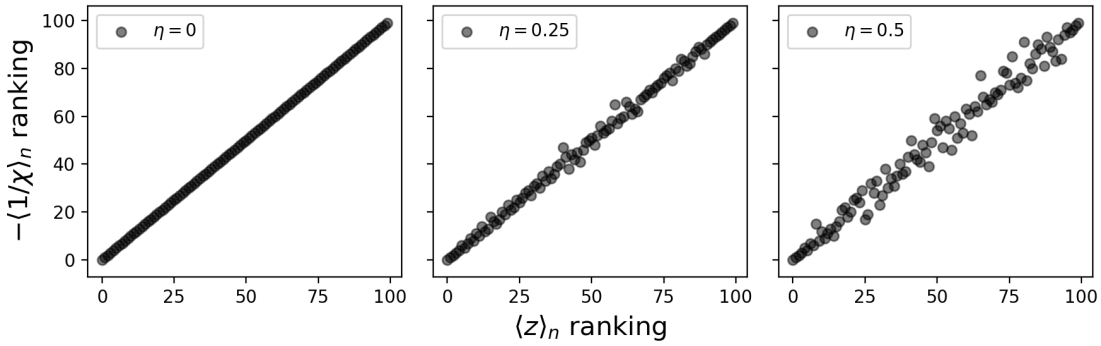

FIGURE 4.6: Scatter plot showing the relative ranking of the three sets of distributions shown in figure 4.5. For realisations without noise, this translates into all of them having the same values for higher moments, which ultimately results in the ranking with $\langle z \rangle$ and $\langle 1/\chi \rangle$ being equivalent. As the noise added to each realisation goes up, their higher order moments start becoming more different, which translates into the two sets of descriptive values d yielding different rankings.

which are boundary conditions for the lensing efficiency functions and their overall shape.

The two sets of descriptive values above not necessarily result on the same ordering. To visualize this difference we generate a set of probability distribution functions resembling the shape of a tomographic redshift distribution and compute the $\langle z \rangle$ and $\langle 1/\chi \rangle$ rankings. We shift a single gamma distribution according to equation 4.12, drawing 100 Δz values from a zero-centered Gaussian distribution with a $\sigma = 0.05$ standard deviation. Gamma distributions are characterized by a probability density function parameterized by a shape parameter k and a scale parameter θ such that

$$n(z; k, \theta) = \frac{z^{k-1} e^{-z/\theta}}{\theta^k \Gamma(k)} \tag{4.17}$$

and allow us to generate a distribution which vanishes for negative redshift values while

behaving reasonably well on the tail end of the redshift range. We then generate separate additional ensembles by adding a different noise realisation to each value $n(z_i)$ according to the following formula

$$n(z_i) \to (1 + f_i(\eta))n(z_i) \tag{4.18}$$

where $f(\eta)$ is a random factor drawn from a $[-\eta, \eta)$ uniform distribution. In the case of noiseless shifts ($\eta = 0$), the realisations retain their shape, which results in a one-to-one monotonic transformation between $\langle z \rangle$ and $\langle 1/\chi \rangle$. When noise is added ($\eta > 0$) this results on a non-monotonic transformation which can alter the relative ordering of the distributions. We generate three ensembles with the values $\eta = [0, 0.25, 0.5]$, shown in figure 4.5, and rank them using the above descriptive values. The resulting dispersion of rankings is shown in figure 4.6, where it can be seen that for *noisier* distributions the impact on ranking is more noticeable. In chapter 5 we test the impact of the different ranking schemes on sampling efficiency.

4.2.2 Multi-dimensional case

While the one-dimensional approach presents a clean and simple strategy to arrange and select realisations on each likelihood evaluation, it does not shield against cases where two sets of realisations with very different d_i values (i.e. the mean per tomographic bin) are assigned a similar rank due to having similar $\mathbf{d}$ (See equation 4.10) over all the tomographic bins. Here we describe a generalisation to rank distributions using multiple dimensions, which allows to use more than one descriptive parameter $\mathbf{d}$ to assign the proposal $n(z)$ realisations to a space of hyper-parameters $\mathcal{H}$.

The set of N_p proposal $n(z)$ is assigned a position in a uniform multi-dimensional grid of $N_1 \times \cdots \times N_{N_d} = N_p$ points, $\mathbf{u}$, according to a set of N_d descriptive values $\mathbf{d} = d_1, ..., d_{N_d}$. This grid is contained inside a N_d-dimensional unit cube, and the coordinates of the hyper-cube are the continuous hyper-parameters $\mathcal{H}_j \in [0, 1)$ which are sampled in the MCMC chain. Each time a set of hyper-parameters is chosen by the sampler, the method returns the closest point in the grid, which has been previously assigned to a $n(z)$ realisation. The dimensions of the grid depend on the number of proposal $n(z)$, which ideally must be a product of similar integers. In the extreme case N_p is a prime number, the multidimensional grid will have dimensions of $1^{N_d-1} \times N_p$, which will result in an ordering equivalent to the one dimensional case. If, in turn, N_p can be decomposed in N_d non trivial integers, we want them to be as similar and large as possible, since we want each of the hyper-parameters sampled by the MCMC to span the largest possible uncertainty encoded by the descriptive parameters d_i they represent. While ideally one would want to use a large number of dimensions to help constructing a space where the posterior is as smooth as possible, this comes at the expense of having to construct a grid with a low number of points per dimension. This can result in a noisy posterior as a function of the hyper-parameter $\mathcal{H}$ since all the realisations in the same row or column of grid points have no further ordering along that dimension. All of the realisations located inside each sub-interval of the grid are essentially randomly sampled from the

perspective of their corresponding hyper-parameter. If the sub-interval is a significant fraction of the total $\mathcal{H}$, this can result in a low sampling efficiency due to the sampler not being able to properly reduce the volume of parameters along that dimension. To find the optimal integer decomposition for a number of samples N_p we use a simple algorithm:

```
def factors(f, dim=N):
    if dim == 1:
        return f
    p = np.zeros(dim, dtype=int)
    s = int(f**(1/dim))
    for i in range(s,1,-1):
        if f % i == 0:
            p[0] = i
            p[1:] = factors(int(f/i), dim= dim-1)
    return p
```

In the case of a $N_d = 1$, where a single characteristic value describes each realisation and the arrangement of points is done over a grid in the interval $[0, 1)$, the optimal distribution is the one which ranks the points in order, which is equivalent to the one-dimensional scheme described in 4.2.1.

Linear sum assignment

One approach to find the optimal relative positions of the descriptive values is to use the solution to the *linear sum assignment problem* (Kuhn, 1955a). Given a set of N_p workers (points in the descriptive value space) we want to find an assignment to N_p fixed jobs (Fixed grid positions in the unit hyper-cube) such that the sum of the cost to assign each worker to one and only one job (the distance from descriptive value space to hyper-cube position) is minimised:

$$\min \sum C_{ij} X_{ij}, \tag{4.19}$$

where C_{ij} is the cost matrix of assigning each sample $\mathbf{d_i}$ to each point $\mathbf{u}_j$ of the grid, and X_{ij} is a binary matrix indicating which position is assigned to each set of descriptive values. If we use an Euclidean distance metric such that $C_{ij} = |\mathbf{d_i} - \mathbf{u_j}|^2$, the resultant assignment minimises the total distance moved by the points to the positions on the grid ensuring that any notion of neighbourhood between points in the original space of descriptive parameters is preserved in their new unit hyper-cube grid positions. Figure 4.7 shows a basic example of this ordering in two dimensions with 16 random points being assigned to the grid coordinates of the hyper-cube.

To solve the linear sum assignment, we first normalize the positions of the points in descriptive value space to have unit variance and a mean located at the center of the unit hyper-cube, and use the SCIPY.OPTIMIZE implementation of the Hungarian Algorithm (Kuhn, 1955b) to find their final positions. The algorithm finds row and column permutations of the cost matrix such that the sum of the diagonal, representing the assigned positions of the realisations to the points of the multidimensional grid, is minimized. To do this, a constant value can be added or subtracted to each row or column which does

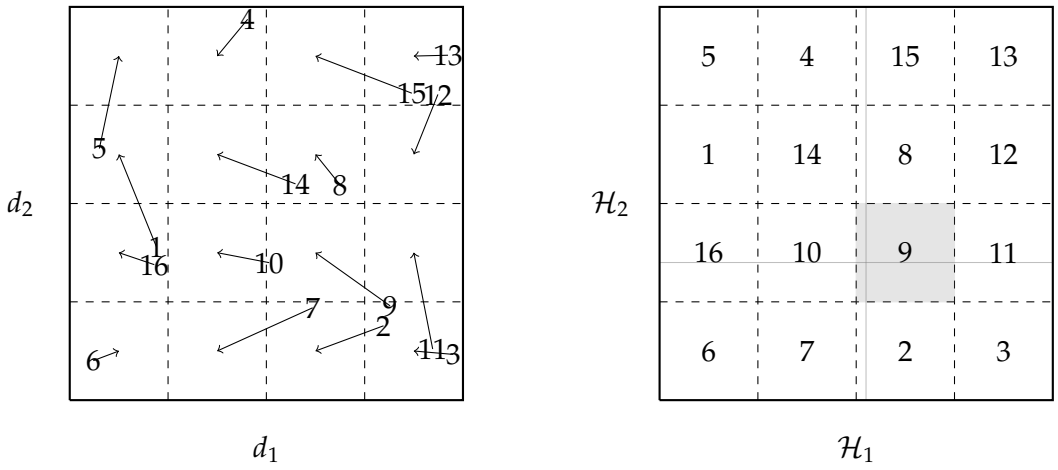

FIGURE 4.7: Basic example showing the concept of the *linear sum assignment* ranking scheme. A set of 16 random points, whose positions are marked by numbers on the (d_1, d_2) plane are shifted to the position of fixed grid points such that the sum shifts (length of the arrows) is minimized. This minimization results in no arrows crossing paths, which indicates the preservation of the neighborhood of the points in the original distribution. Right panel shows the sampling procedure, in which a pair of values $(\mathcal{H}_1, \mathcal{H}_2) = (2.1, 3.4)$ results in realisation 9 being sampled by HYPERRANK.

not change the optimal solution, but only the total cost associated to it. Then the method becomes a matter of finding a zero-diagonal. The method employed to solve the assignment problem scales as $\mathcal{O}(n^3)$ and can't be parallelized, which makes it only viable for values of N_p of the order of a few thousands, with $\sim$ 3000 realisations taking up to 24h in a single core. While this is close to the expected number of realisations for the DES Y3 analysis, larger number of realisations require faster alternatives to map realisations to the multi-dimensional grid.

Uneven grid assignment

One such alternative is to use an uneven grid of dimensions $N_1 \times \cdots \times N_{N_d} = N_p$ where N_d is the number of dimensions and descriptive values used for ranking. Samples are ranked according to the first descriptive value, d_1, and separated into N_1 subsets of N_p/N_1 realisations each. (See figure 4.8) Because of the constraitn on the dimensions of the grid, this number must be an integer. Each of the subsets will then be assigned in corresponding order to the grid along its first dimension. Each subset of N_p/N_1 samples is then ranked according to their second descriptive value, d_2, and separated into N_2 smaller subsets containing $N_p/(N_1 \times N_2)$ samples each. The procedure is then repeated until all descriptive values are used to separate the samples, or equivalently, until each sample has been isolated into a single grid, corresponding to a coordinate which can be mapped to the uniform grid and a single combination of hyper-parameters $\mathcal{H}$. It must be noted that while this scheme can result in equivalent orderings to the one using the linear sum assignment approach, this is not always the case, and even for the same set of points in descriptive value space, the order each of them is used to rank along the

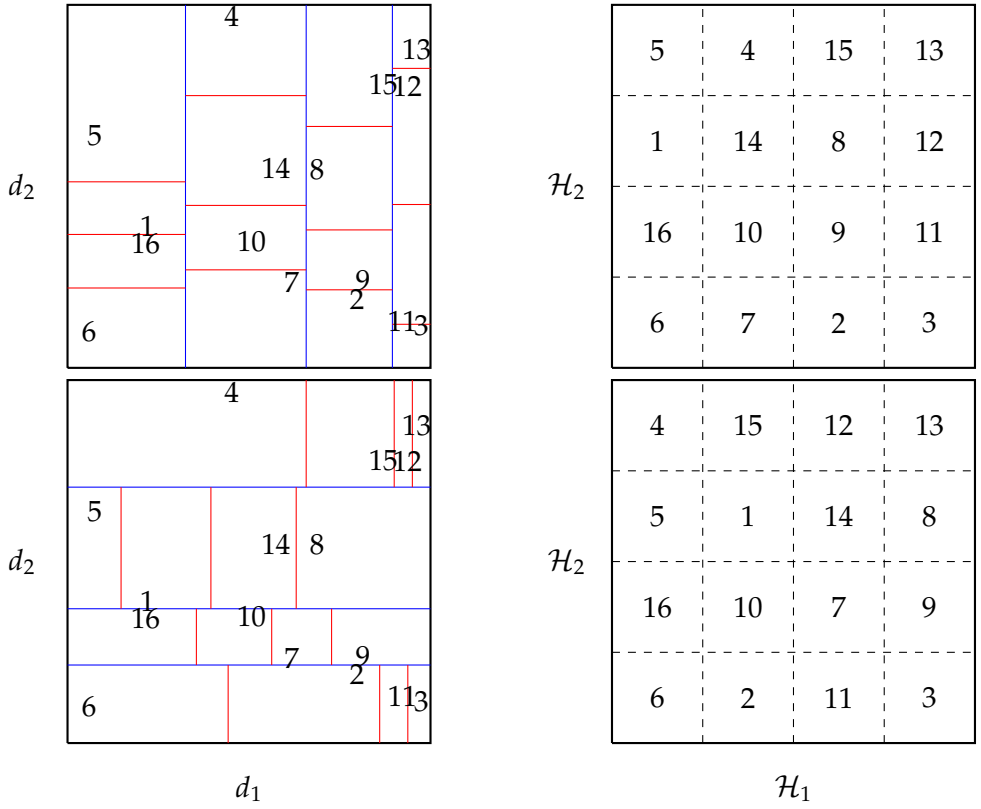

FIGURE 4.8: Similar to 4.7, this shows a basic example of the *uneven grid* approach to map the same set of points randomly distributed and marked by their respective number. Top panel shows the resultant arrangement once the ranking first uses descriptive value d_1 to generate subsets of realisations which are then ranked according to their d_2 values. This results in exactly the same ordering as in the linear sum assignment scheme. Bottom panel shows the resultant ordering if the order of the rankings is inverted to use d_2 first and then d_1, which results in a different ordering to the one used before.

dimensions can significantly change the results, which is evident from the second panel of figure 4.8. One disadvantage of this approach is that for a finite set of $n(z)$ samples, an attempt to map descriptive values $\mathbf{d}$ in high dimensions becomes less effective for the last dimensions to be mapped, because of the variance on the lower and upper limits and distribution of d_i on each subset of the uneven grid. This can result on two samples being assigned contiguous points in the grid and having large differences in the values for their last descriptive values d_{N_d}. A palliative measure to reduce the effect of this variance is to start the ranking with the values whose equivalent uncertainty have the biggest impact on the posterior.

A quick example of the two schemes in two dimensions is presented in figure 4.9, where 400 random points are drawn from a three-dimensional Gaussian distribution in the XYZ volume and their XY coordinates used to train the uniform grid. The first panel shows the distribution of X and Y coordinate values arranged into a 20×20 uniform grid using the linear sum assignment solution, showing that similar points are located close to each other and the variation of the coordinate values is smooth along the dimensions of the hyper-cube.

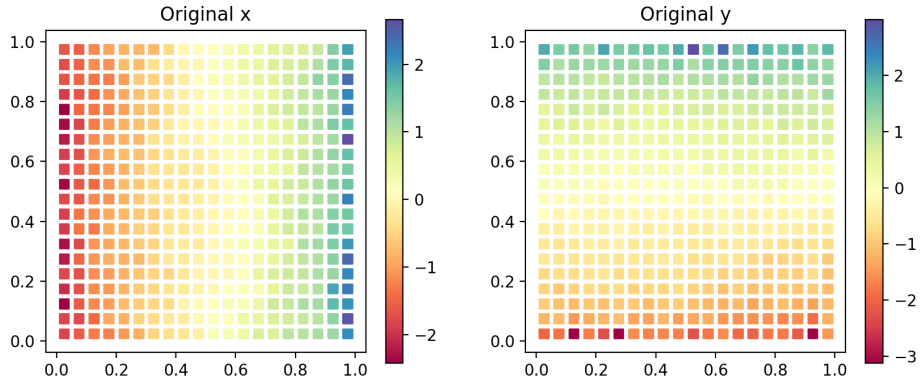

FIGURE 4.9: 20×20 ranking map generated using the X and Y coordinates of randomly points from a 2D Gaussian. Each panel shows the locations of the points on the uniform grid, but the color scale shows the values for the corresponding coordinate It can be seen that the mapping scheme permits the points to remain close to others with similar descriptive values used for the mapping, and has a smooth variation on the directions of the hyperparameters mapped to each dimension of the grid.

Analogous to the one dimensional case, various descriptive values can be used to define the points to be arranged into the hyper-cube, with $\langle z \rangle$ and $\langle 1/\chi \rangle$ still being suitable options. As mentioned above, the big advantage is the fact that multiple descriptive values from the ensemble of realisations can be used to arrange the distributions into the hyper-cube, which can help break the degeneracy between $\mathbf{d}$ and the posterior distribution.

An additional sophistication is to directly identify the most important features of the $n(z)$ variation in determining the inferred cosmology. To do this we find which components (or linear combinations of components) of the cosmic shear data vector (in our case the two-point correlation function of galaxy shapes as a function of angular scale $\xi(\theta)$) are the main contributors to the variation in χ^2 goodness of fit (at a fixed cosmology) computed between the observed data vector and simulated data vectors using each available $n(z)$ realisation.

The different samples $n_i(z)$ influence the likelihood through their impact on the theoretical data vector ξ_i that is produced when sample i is chosen. Let's consider a nominal data vector ξ_0 to come from a model with a reference cosmological and nuisance parameters and value of $n(z)$. Then we have the χ^2 shift induced by changing from nominal to sample i as

$$\chi^2_i = (\mathbf{d}_i - \mathbf{d}_0)^T C^{-1} (\mathbf{d}_i - \mathbf{d}_0) \tag{4.20}$$

where C is the adopted covariance matrix for the data vector. If we define the matrix D such that D_{ij} is equal to the jth element of $\xi_i - \xi_0$, then the total variation of χ^2 over all samples is

$$V \equiv \sum_{i=1}^{M} \chi^2_i = \mathrm{Tr}\left(DC^{-1}D\right). \tag{4.21}$$

If we can factor the symmetric square matrix in the trace by taking its eigenvalues and eigenvectors such that

$$DC^{-1}D = U\Lambda U^T, \tag{4.22}$$

where U is an orthonormal matrix and

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) \tag{4.23}$$

(where N is the length of the data vector now), then we obtain

$$V = \sum_{i=1}^{M} \chi_i^2 = \mathrm{Tr}\left(U\Lambda U^T\right) = \mathrm{Tr}\left(\Lambda U^T U\right) = \sum_{j=1}^{N} \lambda_j = \sum_{i=1}^{M}\sum_{j=1}^{N} \lambda_j U_{ij}^2 \tag{4.24}$$

Now let's assume that the λ_j are in decreasing order (they are all positive when $M > N$), and let's assume that we want our descriptive values $\mathbf{d}$ to have K elements. Then following the usual logic of principal component analysis, the strategy which would leave the smallest residual V after controlling for $\mathbf{d}$ would be to choose the first K coefficients of the PCA as the indicator variables:

$$d_i = \{U_{i1}, U_{i2}, \ldots, U_{iK}\}. \tag{4.25}$$

In this case the sampler, choosing to sample at some $\mathbf{d}$, can expect that the χ^2 value (relative to nominal data) to be

$$\chi^2 = \sum_{j=1}^{K} \lambda_j x_j^2 + r \tag{4.26}$$

where the "roughness" function r has an RMS value over the samples of

$$\langle r \rangle = \frac{1}{M} \sum_{j=K+1}^{N} \lambda_j. \tag{4.27}$$

This formula also tells us that, ideally, we should increase K until the quantity on the previous line is $\ll 1$.

4.3 Summary

In this chapter we have described the Bayesian approach to inferring model parameters using a family of techniques which stochastically samples the space of parameters. Using specific acceptance rules, high dimensionality distributions can be sampled efficiently not only giving insight into the best fit parameters of a model, but also on the degree of confidence given our prior expectations and their associated uncertainties. We described the concept of nested sampling, how it can be employed to compute evidences and obtain samples from the posterior distribution, and how specific conditions can impact the efficiency of such techniques. In this context, and in view of the proposed scheme for the estimation of source redshift distributions for the DES Y3 analysis, presented in chapter

3 we presented HYPERRANK, a novel technique to propagate photometric redshift uncertainties in weak lensing and galaxy clustering experiments. By sampling from a uniform space of hyper-parameters mapped to an ensemble of proposal redshift distributions, HYPERRANK is able to propagate any type of uncertainty described by the ensemble. This provides an alternative to simpler approaches where the uncertainty is only described as a shift of the distributions along the redshift direction.

In chapter 5 we will describe a series of tests devised to validate and explore the performance of HYPERRANK in simulated data and compare it to the more traditional Δz marginalisation approach. In chapter 6 we will present the results of applying HYPERRANK to the data products of the DES Y3 analysis.

Chapter 5

Testing and validation of HYPERRANK in the Buzzard simulations

In chapters 3 and 4 we presented the plan for the estimation of the redshift distribution of weak-lensing sources in the DES Y3 analysis and the process to characterize and propagate their uncertainties into the inferred cosmological parameters. As with all the methods employed in the pipeline these steps must be validated in order to understand potential sources of biases and to characterize their associated uncertainties. We start this chapter describing the main characteristics of BUZZARD in section 5.1, a set of N-body simulations purposely built to mimic the properties of real DES data and to be used for validation of several aspects of the full pipeline, including the SOMPZ scheme and HYPERRANK. Then we present the tests for validation of the HYPERRANK marginalisation technique applied to the BUZZARD simulations.

The author of this thesis designed and conducted the tests to validate and quantify HYPERRANK sampling efficiency under different equivalent uncertainties, number of realisations, choice of descriptive values $\mathbf{d}$ and ranking schemes. The author also generated the samples describing the different types of uncertainty and configures CosmoSIS to run HYPERRANK on those ensembles. The designed tests and their results shown here are also presented in Cordero et al. (2020).

5.1 BUZZARD simulations

The BUZZARD simulations are a set of mock DES Y3 surveys created from a suite of 18 dark-matter N-body simulations using a memory-optimised version of L-GADGET2 (DeRose et al., 2019, 2020) which separates the full redshift range into three boxes which are simulated independently. A single set of these simulations is able to provide a light-cone with an area of 10,413 square degrees up to a redshift of $z = 2.35$. Galaxies and their main morphological properties are added using ADDGALS (DeRose et al., 2019).

Each galaxy is assigned an absolute magnitude based on an empirical luminosity function from obtained from the Sloan Digital Sky Survey Main Galaxy Samples (SDSS-MGS), and their positions assigned based on clustering statistics from the same survey. Based on the absolute magnitude, a spectral energy distribution is assigned to each galaxy and DES *ugriz* and VISTA *JHK* photometry is obtained from the simulated spectral energy distributions. Intrinsic ellipticity for the simulated galaxies are assigned based on the obtained empirical luminosity function and then lensed and magnified by the simulation large scale structure using ray tracing. The samples of galaxies are then "observed" by DES, by rotating their positions to the footprint of DES, and applying a noise model to the fluxes. This is necessary since the noise model is position dependent, based on Galaxy extinction and the depth at each band on the real observations of DES.

The end result is a large catalog of sources with true photometry in any arbitrary band, true redshifts, ellipticity values and noisy simulated values of the photometry as observed by DES. These are fundamental ingredients for the validation of the methodologies in the DES Y3 analysis, since comparison between obtained and true values can be used to assess the performance and associated uncertainties of each method.

In the case of HYPERRANK, BUZZARD presents an appropriate testing ground to evaluate its performance and suitability for use in the pipeline of the DES Y3 analysis by generating different sets of $n(z)$ samples covering the types of uncertainty for which HYPERRANK is expected to provide a better marginalisation than the fiducial Δz approach.

5.1.1 $n(z)$ Ensemble from the BUZZARD simulations

In 3.3 we described the process to characterize the uncertainty associated with the SOMPZ scheme, which employs a smart discretization of the deep and wide colour spaces of two galaxy samples to build a transformation between the two by assigning galaxies with real and simulated photometry available for both samples. In section 3.2.2 we also described the 3SDIR process in which the histogram uncertainties due to cosmic variance and shot noise are incorporated into an ensemble of $n(z)$ obtained from the best-estimate SOMPZ $n(z)$ distribution using a three-step Dirichlet sampling process. The pipeline described above was applied to the simulated galaxy catalogs generated from the BUZZARD simulations as part of the validation process described in (Myles et al., 2020) and resulted in 500 $n(z)$ distributions (Figure 3.10) which we use here as the base for the validation and performance tests. While this set of realisations does not span the full range of types of uncertainties we expect HYPERRANK to be effective over, they do represent a realistic scenario expected for DES Y3 data. We can use them as a base to construct additional sets of realistic realisations, keeping their consistency with the simulated data-vectors we use in our analysis. We use this ensemble of realisations to define a fiducial redshift distribution, $n_{\texttt{Fid}}(z)$ by averaging the values for each histogram bin across all 500 realisations. We then use this distribution as the base to construct several additional distribution ensembles encoding different types of uncertainty.

5.2 Validation and performance tests

We aim to validate the HYPERRANK method for marginalising over redshift distribution uncertainty and explore its configuration, with the target of using it for the weak lensing source redshift distributions in the DES Year 3 cosmological analysis. To do so we explore the effects of marginalizing over ensembles of $n(z)$ distributions constructed to represent four main types of uncertainty:

- Uncorrelated Gaussian mean shifts: Distributions from these ensembles are shifted versions of $n_{\texttt{Fid}}(z)$ along the z-direction according to equation 4.9, with shifts for each tomographic bin corresponding being drawn from a Gaussian distribution which is uncorrelated between tomographic bins. This is similar to the Δz approach described in 4.2, with the exception that Δz_i values are limited to the values drawn before marginalization, and not obtained arbitrarily by the sampler.

- Uncorrelated non-Gaussian mean shifts: Similar to the above, but Δz values are drawn from an asymmetric Gamma distribution with long tail towards high redshift, parameterized by a set of shape and scale parameters to match the variance of an equivalent reference Gaussian distribution.

- Correlated Gaussian mean shifts: Δz values are drawn from a multi-dimensional Gaussian distribution characterized by a covariance matrix which can be adjusted to result in different levels of correlation between pairs of tomographic bins.

- Amplified peculiarities: The histogram values $n(z_i)$ of each of the 500 SOMPZ +3SDIR realisations are amplified with respect to their average values in $n_{\texttt{Fid}}(z)$ to accentuate the peculiarities of each realisation.

We investigate the HYPERRANK method's ability to marginalise over these types of uncertainty and compare it against the results obtained from marginalizing uncertainty using the standard Δz approach, which is only able to accurately describe Gaussian Δz shifts. We also investigate whether hyperrank is able to marginalise these types of uncertainty both *correctly* in terms of the coverage of the input space of possible $n(z)$ which each ensemble represents and *efficiently* by requiring as few likelihood evaluations in the MCMC as possible to converge. We test the correctness by comparing the error bars on Ω_m, σ_8 and $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$ parameters obtained from a cosmological parameter inference pipeline. Efficiency is tested by comparing the number of likelihood evaluations under several configurations of HYPERRANK including the one- and multi-dimensional implementations described in 4.2.1 and 4.2.2 respectively, as well as descriptive values $\mathbf{d}$ used to rank the $n(z)$ realisations.

An important goal of these validation tests is to provide a baseline configuration to be used in the DES-Y3 cosmology pipeline, and the expectation regarding the input $n(z)$ ensemble. We estimate an approximate minimum number of $n(z)$ realisations required before systematic errors on the cosmology parameters from the discreteness introduced by HYPERRANK become negligible. Throughout these tests we use the DES-Y3 modelling

TABLE 5.1: Summary of cosmological, systematic, and astrophysical sampling parameters used in the fiducial analysis and their priors. In the case of flat priors, the prior bound to the range indicated in the Value column while Gaussian priors are described by their mean and 1σ width.

Parameter	Symbol	Type	Value
Cosmological			
Matter density	Ω_m	Flat	[0.1, 0.9]
Baryon density	Ω_b	Flat	[0.03, 0.07]
Scalar spectrum ampl.	$A_s \times 10^{-9}$	Flat	[0.5, 5.0]
Hubble parameter	h	Flat	[0.55, 0.91]
Spectral ind.	n_s	Flat	[0.87,1.07]
Neutrino mass	$\Omega_\nu h^2$	Flat	[0.0006,0.00644]
Curvature	Ω_k	Fixed	[0.0]
Optical depth	τ	Fixed	[0.0697]
Observational			
Shear calibration 1	m^1	Gauss.	(0, 0.005)
Shear calibration 2	m^2	Gauss.	(0, 0.005)
Shear calibration 3	m^3	Gauss.	(0, 0.005)
Shear calibration 4	m^4	Gauss.	(0, 0.005)
Intrinsic alignments			
Tidal alignment ampl.	a_1	Flat	$[-5, 5]$
Tidal torque ampl.	a_2	Flat	$[-5, 5]$
Tidal alignment redshift ind.	η_1	Flat	$[-5, 5]$
Tidal torque redshift ind.	η_2	Flat	$[-5, 5]$
Tidal alignment bias	$b_{\rm ta}$	Flat	$[0, 2]$

choices, likelihood and pipeline software and configuration, which are described in detail in Amon et al. (2020); Secco et al. (2020). The choice of parameters limits and priors is similar to the analysis choices presented on those references and is presented in table 5.1.

We only consider cosmic shear in our data vector, which reduces the dimensionality of the space of parameters to be sampled in the MCMC inference and enhances the effect of redshift systematics in the source sample. Nevertheless, this method can also be applied to the distribution of lensing sources and used simultaneously in both redshift distributions.

5.2.1 Sampling Efficiency

In addition to the correct exploration of the uncertainties, we also wish to see the effect of the HYPERRANK procedure on the efficiency of mapping the posterior of cosmological and nuisance parameters. For a randomly sampled set of distributions the likelihood is not a smooth function of the parameters being sampled. Therefore, the parameter space volume cannot be sampled consistently in higher likelihood regions since there is no correlation between the sampled nuisance parameter and cosmology posterior. In this case, any proposal step in the Monte Carlo algorithm typically do not have the intended effect, since proposed jumps in the redshift nuisance parameters are now across a random,

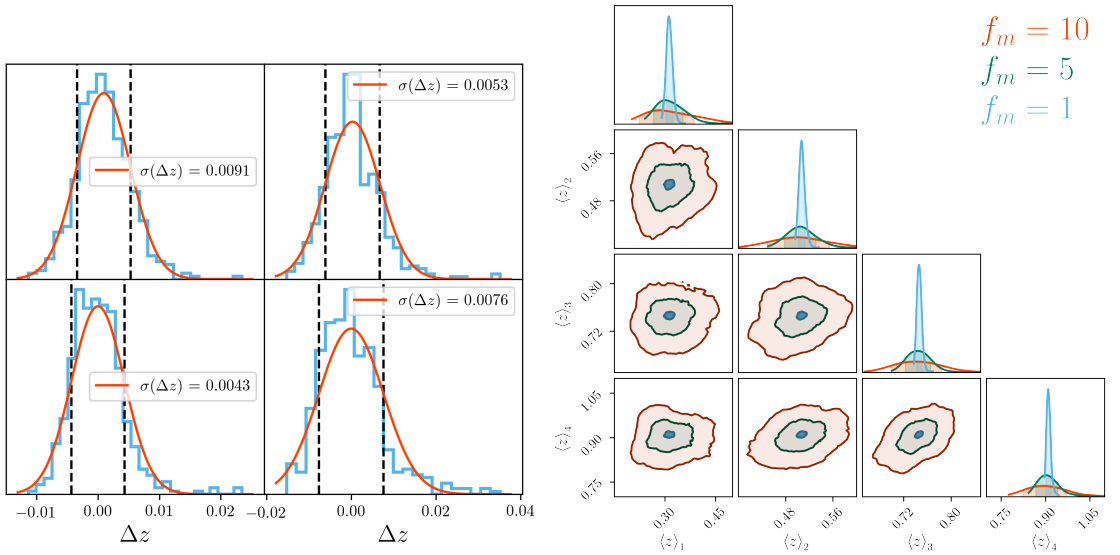

FIGURE 5.1: *Left:* Histogram of Δz values (blue) for the ensemble of 500 SOMPZ +3SDIR realisations obtained from the BUZZARD simulation. Each Δz value is obtained from computing the difference between mean redshift from each tomographic bin from each realisation and the mean redshift from the fiducial distribution $n_{\texttt{Fid}}(z)$, which is computed from averaging all 500 realisations. A Gaussian fit (solid orange line) describes the corresponding equivalent Gaussian prior, characterized by $\sigma(\Delta z)$. *Right:* Corner plot of mean redshift values for all four tomographic bins for the three $n(z)$ ensemble generated by amplifying the equivalent $\sigma(\Delta z)$ from the 500 SOMPZ +3SDIR realisations. The amplification preserves the correlation and non-Gaussianity of the Δz distributions.

discontinuous likelihood. This leads to the sampler requiring many more likelihood evaluations to find new samples of the posterior. In the case of nested samplers like the ones described in 4.1.3, a non-smooth posterior can result on its inability to consistently reduce the volume of parameters it samples, since the replacement of live points becomes highly non-deterministic. HYPERRANK mapping of distributions using descriptive values $\mathbf{d}$ seeks to arrange the ensemble of $n(z)$ realisations such that the posterior is a smooth function of the space of hyper-parameters $\mathcal{H}$.

To quantify the relative performances we define the sampling efficiency η as the number of replacements (samples of the posterior) made by MULTINEST over the total number of likelihood evaluations required for convergence. We choose this instead of just using the number of likelihood evaluations, as this quantity is highly dependent on the dimensionality and volume of the space of sampled parameters. We test the different mapping schemes described in sections 4.2.1 and 4.2.2 (1D and 3D $\langle z \rangle$, 3D $\langle 1/\chi \rangle$ and 3D-PCA) and compare the sampling efficiency between them and against a naïve random sampling of realisations. To do this we employ three different sets of realisations, the 500 SOMPZ +3SDIR $n(z)$ and two additional sets obtained by shifting each tomographic bin from the realisations using 4.12 by a value Δz proportional to the difference between their mean that of the $n_{\texttt{Fid}}(z)$ bins.

$$\Delta z_i = f_m \left[\langle n_{\texttt{Fid}}(z) \rangle - \langle n_i(z) \rangle \right] \tag{5.1}$$

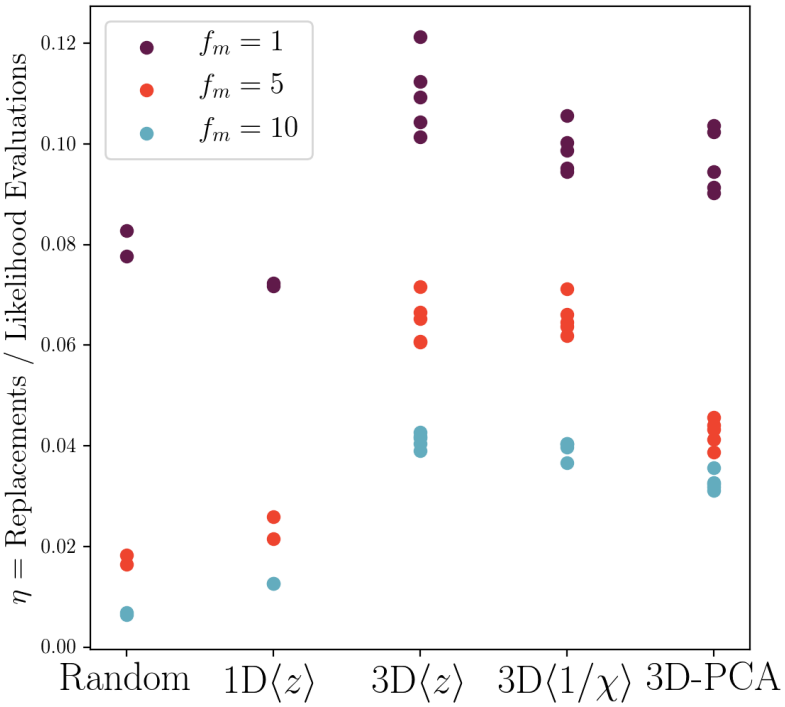

FIGURE 5.2: The sampling efficiency η for the different mapping schemes described in sections 4.2.1 and 4.2.2, for three $\sigma(\Delta z)$ amplifications values f_m ($f_m = 1$ corresponds to the original BUZZARD redshift distributions). Ranking schemes are sorted from left to right according to its perceived complexity, with "random" corresponding to the simplest scheme in which realisations are sampled completely at random. Each ranking scheme and prior amplification is repeated 5 times to better discern the effect of the ranking over random sampling noise.

where f_m is a proportionality constant we arbitrarily set to 1, 5 and 10. This preserves the correlations between tomographic and histogram bins, small peculiarities of each independent realisations, and non-Gaussianity of the sample (See figure 5.1) while allowing us to explore the effect of larger equivalent uncertainties. Through this chapter we use the concept of *equivalent uncertainty* of an ensemble of $n(z)$ distributions, which we define as standard deviation of the distribution of mean redshift values, $\sigma(\Delta z)$. To reduce the effect of sampling noise due to the stochastic nature of the sampler, we repeat each run five times with different initial random seeds for the sampler. Figure 5.2 shows the individual sampling efficiencies η for each run at different f_m values and as a function of descriptive values $\mathbf{d}$. For all f_m values it is clear that the more complex choices of $\mathbf{d}$ using multiple dimensions are more efficient at exploring the space of uncertainties, with $3D\langle z \rangle$ and $3D\langle 1/\chi \rangle$ leading at all f_m values. This is expected since the addition of more dimensions helps breaking the degeneracy of the posterior values present when a single parameter is used and all the information of the $n(z)$ realisations is compressed into a single hyper-parameter $\mathcal{H}$.

The PCA approach, also tested in three dimensions, provides an improvement over

random and 1D sampling, but does not reach the same levels of efficiency for methods of equal dimensionality. This may be caused by the fact that the principal components are computed for the variations in $\xi_k^l(\theta)$ with respect to a reference data vector obtained at a fixed cosmology, and the relative importance of each data point reflected in the principal components of the $\Delta\xi_{i,j}^l(\theta_k)$ matrix can change as the sampler moves in cosmology.

Perhaps one surprising result occurs when comparing the random approach against $1D\langle z\rangle$ in the un-amplified case ($f_m = 1$), in which the former does not appear to be consistently less efficient. We believe this is caused by the relatively small contribution of $n(z)$ uncertainty to the posterior in the $f_m = 1$ regime, as all realisations have very similar mean redshift values in all tomographic bins. This can lead to a very small change of smoothness of the posterior at a fixed cosmology when moving from a random ordering to a 1D ordering, resulting on similar efficiencies. While we do not show the effect of additional dimensions for a similar type of descriptive value $\mathbf{d}$ (i.e. 4D $\langle z\rangle$), some test runs suggest their efficiency is not noticeable better than a 3D approach, at the expense of noisier posteriors on the $\mathcal{H}$ parameters.

Based on these results we consider a 3D approach an appropriate default configuration, with a preference for $\langle z\rangle$ since its computation does not involve the use of a fiducial cosmology, unlike $\langle 1/\chi\rangle$.

5.2.2 Minimum number of samples

In HYPERRANK, discrete samples from the posterior over the sub-set of redshift nuisance parameters are generated outside of the main chain used to sample over the cosmological and other nuisance parameters. This means a limited and discrete set of values of the nuisance parameters are available to the main sampling, as opposed to the continuous range of parameters within a specified prior which would be available otherwise. There will be a transition from the regime in which there are two few realisations of $n(z)$ available to effectively explore the redshift distribution uncertainty, and the limit where infinitely many realisations would be available, corresponding to the continuous case. Here we investigate the convergence of HYPERRANK marginalisation with respect to the number of $n(z)$ samples generated, for the case of our DES-Y3 simulated data set.

We first generate several sets of distributions where each realisation is a shifted version of the fiducial $n_{\texttt{Fid}}(z)$, and the shifts are drawn from a Gaussian prior, following a similar approach to the Δz method. We generate 8 sets of redshift distributions, each containing $3^3, 4^3, 5^3, 6^3, 7^3, 8^3, 9^3, 10^3$ realisations which are then ranked using the 3 dimensional default configuration described at the end of 4.2.2.

Since we expect the approximate minimum number of realisations required for this convergence to depend on the level of uncertainty in the $n(z)$, we generate two additional sets of proposal distributions by multiplying the $\sigma(\Delta z)$ obtained above, by a factor $f_m = 5, 10$ We then repeat the generation of proposal realisations with five different random seeds for each of the three f_m values, and for each of the 8 sets of realisations containing different number of proposals. By comparing the standard deviation on the central, lower and upper confidence values for the derived $S8$ parameter as a function of the number

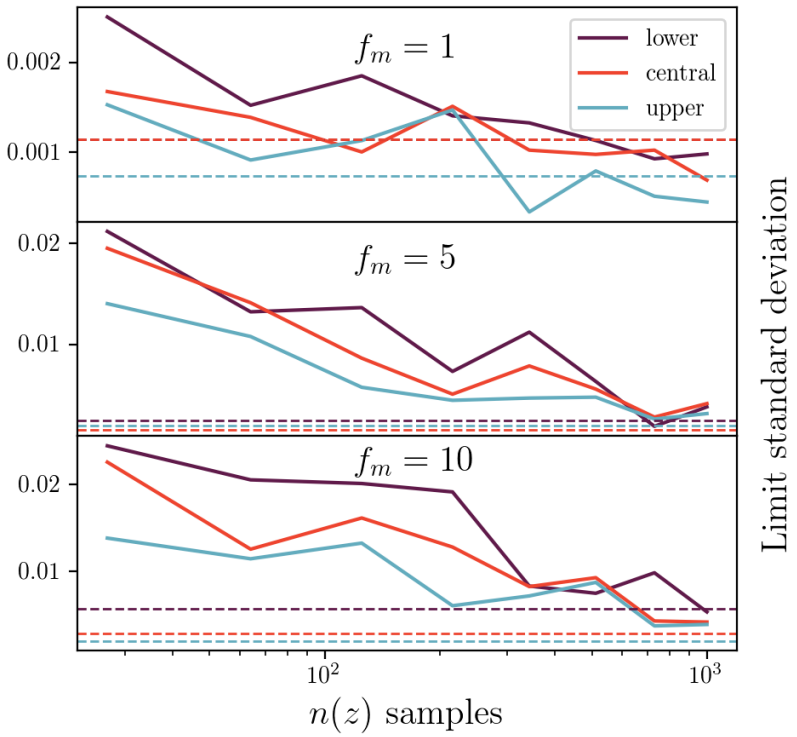

FIGURE 5.3: Standard deviation of the lower (purple), central (red), and upper (cyan) values for the $S8$ parameter obtained using HYPERRANK for 5 realisations of the ensemble of $n(z)$ samples, as a function of the total number of distributions to form the ensemble. From top to bottom, the equivalent $\sigma(\Delta z)$ width is amplified by a progressively larger number, f_m, with respect to the original distributions of BUZZARD samples. Horizontal dashed lines indicate the typical standard deviation for runs using the traditional Δz marginalisation approach

of realisations, we can find an approximate minimum number of realisations required for the standard deviation of error bars from HYPERRANK converged to that obtained using the Δz approach (which is formally correct for this set of realisations). In 5.3 we can observe that for all three levels of uncertainty, described by the amplification factor f_m, 1000 realisations yield standard deviation of the error bars obtained using HYPERRANK comparable to the ones using the Δz approach (typical deviations with respect to the stacked contours are around 0.09σ on all parameters).

5.2.3 Correct marginalisation of uncertainty

In order to test the correctness of HYPERRANK when propagating the uncertainty described by the input proposal redshift distributions into the cosmological parameters, we generate a series of $n(z)$ ensembles each describing a different type of uncertainty. In each case we explore what is the impact of the uncertainty in the inferred cosmological parameters and compare the performance of HYPERRANK with results obtained by using the standard marginalisation approach Δz using a Gaussian prior obtained directly from each ensemble of distributions.

Gaussian distributions for Δz

A simple representation of uncertainty with which redshift distributions can be biased by a constant shift Δz along the z-direction assumes these effects are uncorrelated between tomographic bins. The first test we devise utilizes an ensemble of $n(z)$ distributions shifted by the same distribution used in the standard Δz approach. Within each tomographic bin we draw 1000 values of Δz from a Gaussian distribution with width $\sigma(\Delta z)$. Realisations for $n(z)$ are then generated by shifting the fiducial $n_{\texttt{Fid}}(z)$ along the redshift axis by the drawn Δz. Aside from the fact that the Δz values are defined prior to the MCMC sampling, unlike the case of the standard nuisance parameter marginalisation approach, the application of HYPERRANK to this ensemble of realisations is the closest to an *apples to apples* comparison between the two. The results of this test can serve as our first confirmation that the code and the approach to utilize a transformation between the descriptive value $\mathbf{d}$ space and the hyper-parameters $\mathcal{H}$ yields the expected contours for that particular uncertainty.

In order to assess performance and convergence we test HYPERRANK for three different levels of uncertainty described by $\sigma(\Delta z)$. For each level of uncertainty we amplify the width of the Gaussian prior by multiplying $\sigma(\Delta z)$ by a multiplicative factor f_m. We arbitrarily select three values for $f_m = 1$ (No amplification), 5 and 10.

For our fiducial $\sigma(\Delta z)$ we use the values appropriate obtained from the ensemble of 500 SOMPZ +3SDIR realisations ($\sigma(\Delta z) = \{0.0091, 0.0063, 0.0043, 0.0076\}$, see left panel of figure 5.1). For reference, the values obtained from the DES Y1 source redshift calibration (Hoyle et al., 2018) are $\sigma(\Delta z) = \{0.016, 0.013, 0.011, 0.022\}$ ($f_m \lesssim 3$), and the ones obtained for the Y3 analysis, shown in section 3.3, are $\sigma(\Delta z) = \{0.015, 0.011, 0.008, 0.015\}$ ($f_m \lesssim 2$). The unaltered case reflects a very optimistic calibration, $f_m = 5$ describes a

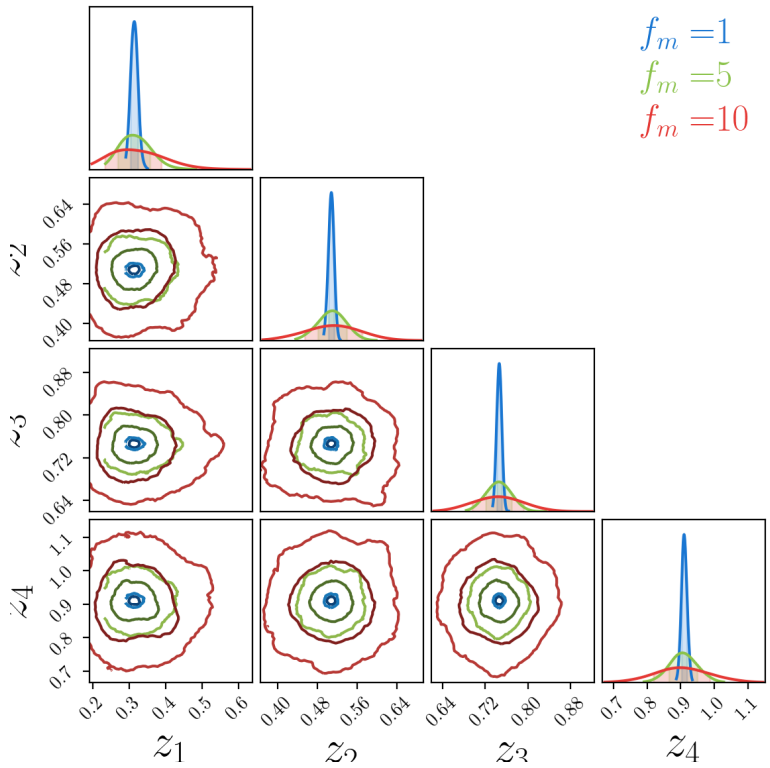

FIGURE 5.4: Corner plot showing the distribution of mean redshift values $\langle z \rangle$ from the ensemble of 1000 $n(z)$ distributions generated using a set of uncorrelated Gaussian distributed Δz values. Blue, green and red contours show the distributions for the three levels of amplification of the equivalent width $\sigma(\Delta z)$ by the factors $f_m = \{1, 5 \text{ and } 10\}$, respectively. It is evident that for large values of the equivalent prior width $\sigma(\Delta z)$, significant fractions of the distribution of Δz values are clipped at low redshifts, as a consequence of the imposed limit at $z = 0$.

somewhat pessimistic but still realistic uncertainty given the typical values from current and past DES analysis and $f_m = 10$ represents a scenario with highly uncertain redshift distributions. It is important to note that large negative values of Δz can result in a high probability mass at low redshift and potentially a relatively large fraction of it shifting to negative ranges of z, especially for the first tomographic bin. When shifting realisations we employ the same shifting algorithm used by CosmoSIS, in which any negative value of $n(z)$ is set to zero and the probability distribution is clipped to be zero for any negative z value. The distribution of Δz values with respect to the fiducial means from $n_{\texttt{Fid}}(z)$ for all three amplification values can be seen in figure 5.4.

Independent of the amplification value f_m used the tomographic bin with the smallest $\sigma(\Delta z)$ corresponds to that of tomographic bin 3. Following the recommended configuration we proposed in the previous section we use the mean redshift $\langle z \rangle$ from tomographic bins 1, 2 and 4 as descriptive values $\mathbf{d}$. We then run the full cosmological parameter estimation pipeline on the simulated data vector using these redshift distributions, marginalising over the uncertainties using both the Δz method and HYPERRANK. For the Δz values we recompute the $\sigma(\Delta z)$ from the generated $n(z)$ ensemble instead of using $f_m \sigma(\Delta z)$. This ensures the effects of clipping at $z = 0$ are accounted for and more

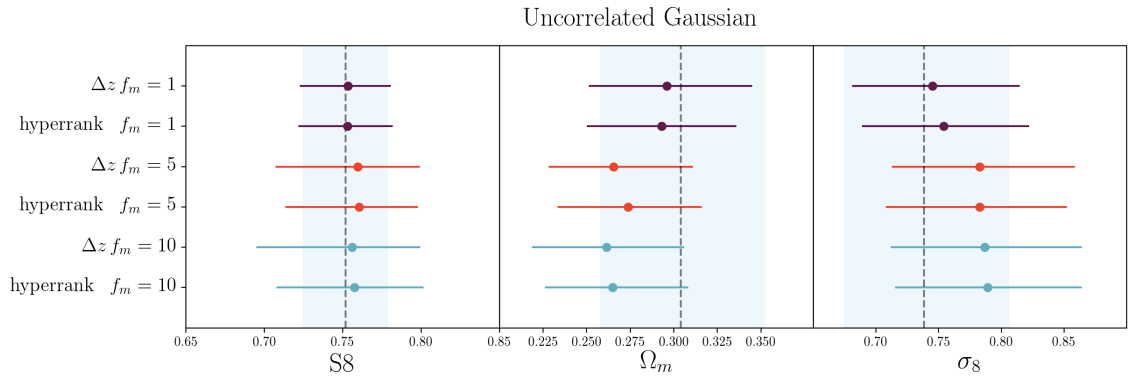

FIGURE 5.5: Summary plots showing the 1D marginalised $S8$, Ω_m and σ_8 error bars obtained from running the cosmic shear MCMC pipeline using HYPERRANK and Δz marginalisation schemes. The ensemble of $n(z)$ distributions used in HYPERRANK describes an uncertainty characterized by uncorrelated Gaussian Δz shifts to the fiducial distribution $n_{\texttt{Fid}}(z)$. Purple, red and blue error bars correspond to the three amplification cases described by the f_m values 1, 5 and 10, respectively. Vertical dashed line and light blue filled region represent the 1-σ error bars from running the MCMC assuming no redshift uncertainty with the fiducial redshift distribution $n_{\texttt{Fid}}(z)$.

closely follows what the standard procedure would be, in which the $\sigma(\Delta z)$ are computed **from** the available samples.

Figure 5.5 shows the 1D marginalized error bars for each run used to compare HYPERRANK and the Δz approach, for the three levels of uncertainty described both by the ensemble of $n(z)$ distributions and the equivalent prior widths $\sigma(\Delta z)$. We see that HYPERRANK gives consistent error bars in S_8, Ω_m and σ_8 when comparing to those obtained using the standard Δz marginalisation approach. We also plot the contours obtained from running the analysis using $n_{\texttt{Fid}}(z)$ as the input distribution and assuming it to be noiseless ($\sigma(\Delta z) = 0$) and observe that for wider equivalent priors the Ω_m and σ_8 drift as much as $1-\sigma$. This is most likely caused by the condition imposed at $z = 0$ resulting on true sampled distributions having slightly larger mean redshift than what is predicted from the Δz shifts alone, especially in the first tomographic bin.

While at first glance this is a trivial example, it shows that the method is at the very least able to recover the same effects of redshift uncertainty when samples describe the same type of uncertainty we typically describe by means of a Δz nuisance parameter.

Non-Gaussian distributions for Δz

Modelling the distribution of Δz for each tomographic redshift bin as a Gaussian constitutes an approximation to the true level of uncertainty which can potentially result in an incorrect propagated uncertainty depending on the true Δz distribution shape. In the left panel of 5.1 we show histograms of the Δz between $n_{\texttt{Fid}}(z)$ and the 500 realisations generated using the full uncertainty model for DES Y3. These show appreciable non-Gaussianity, with skews and heavy tails which can be accentuated by the hard boundary at $z = 0$ for all distributions. An alternative is to describe the Δz prior using a more complex functional form to account for these high order effects, or sample directly from

the histogram of mean values obtained from an ensemble of distributions. HYPERRANK is able to deal with these high order effects as it doesn't require any functional form to be fitted to the uncertainty description, abandoning the idea of uncertainty only being described by a shift Δz.

The test presented here serves an illustrative goal since the case for a non-Gaussian uncertainty can be easily solved with the above alternatives. We investigate the impact of the Gaussian approximation by generating sets of proposal distributions by sampling Δz values from a highly skewed Gamma distribution to shift our fiducial distribution $n_{\texttt{Fid}}(z)$, and then marginalising over the uncertainty in a cosmological chain using HYPERRANK. We employ the *shape-scale* parameterisation for the Gamma distributions:

$$f(\Delta z; k, \theta) = \frac{\Delta z^{k-1} e^{x/\theta}}{\theta^k \Gamma(k)} \tag{5.2}$$

where $\Gamma(k)$ is the Gamma function. The distribution is defined for $\Delta z > 0$ and parameterized by the shape parameter k and scale parameter θ, both having positive values. The shape parameter dictates how skewed the distribution is: small values of k define a distribution with a long tail towards positive Δz values and a steep cutoff at 0, while larger values make the distribution appear more symmetrical. The scale parameter θ typically describes how "spread out" the distribution is. The convenience of using this parameterisation comes from the fact that its first moments are easy to compute in terms of k and θ, with its mean being $k\theta$ and its variance equal to $k\theta^2$.

Using the prior definitions based on $\sigma(\Delta z)$ we obtain the scale parameter θ such that the standard deviation of our Gamma distribution is equal to that of a Gaussian by fixing the shape parameter to a set of arbitrary low values $k = 1, 2, 3$ to ensure the distribution of mean shifts of all tomographic bins have a long tail to high values and to explore the effect of different degrees of non-Gaussianity. The distribution of values is then centered so that the mean shift value is equal to zero, which generates a set of Gamma distributed Δz with the same variance and mean of that of a Gaussian, but with a skewness that cannot be captured by the use of a standard Gaussian prior. We saw in the previous section when using uncorrelated Gaussian draws that the effect of the smallest prior ($f_m = 1$) barely inflates the contours with respect to the case where no redshift uncertainty is considered. For this reason, we decide to draw the Δz values from a distribution characterized by the largest amplification values, $f_m = 10$. The distributions of Δz values for the three shape parameters used here are shown in figure 5.6

The effect of uncertainty being better described by a Gamma distribution with a high skewness does not seem to significantly affect the S_8 error bars shown on figure 5.7, but is evident on individual cosmological parameters, with Ω_m values typically $\sim 0.2\sigma$ above the ones obtained using the Δz, and $\sim 0.3\sigma$ below for σ_8. These values are larger than the typical sampling noise for 1000 realisations, characterised in 5.2.2, and apparently independent of the shape parameter k, although it is worth noting that the three selected values represent a fairly asymmetric distribution, as opposed to the Gaussian distributions characterised by $k \to \infty$. The differences observed seem to suggest a moderate but

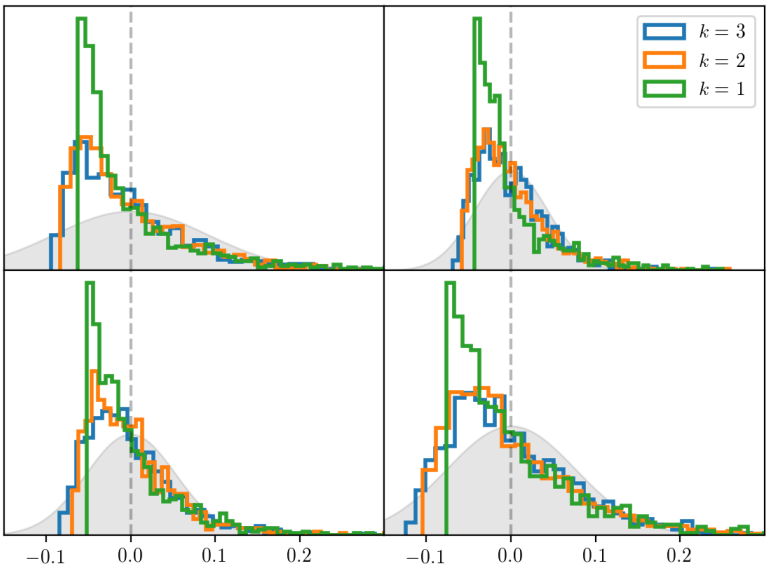

FIGURE 5.6: Histogram of Δz values computed with respect to the fiducial distribution $n_{\texttt{Fid}}(z)$ for the ensemble of 1000 $n(z)$ distributions generated using a set of uncorrelated Gamma distributed Δz shifts. Blue, orange and green histogram show the distributions for the increasingly non-symmetric distributions described by the shape parameters $k = \{3, 2$ and $1\}$. Gray filled plot shows a Gaussian distribution with an equivalent $\sigma(\Delta z)$, which would have been used in the standard Δz marginalisation approach.

non-negligible impact of non-Gaussianity on cosmological parameters, and are consistent with the chosen direction of the tail of the distributions towards higher redshifts, which favours higher values for Ω_m, but are also obtained for an extremely wide equivalent $\sigma(\Delta z)$ prior, with typical deviations for smaller equivalent priors expected to be even smaller.

Correlations between tomographic bins

Another aspect of uncertainty the standard Δz scheme fails to account for is the potential correlation between the uncertainty from different tomographic bins and histogram bins. Since each tomographic bin is shifted independently, combinations of Δz values which would not be expected to appear in multiple realisations of the survey or photo-z analysis are equally sampled. In addition to this, the use of a single fiducial shifted $n(z)$ negates the potential effect of correlation at the histogram bin level. Correlation can come from the binning of galaxies and from how the shapes of the distributions and their moments can change when galaxies are re-assigned to another histogram or tomographic bin in a different realisation of a photo-z analysis. Depending on the nature of the colour-redshift degeneracy, correlation can also appear between non contiguous tomographic bins. In this case, the standard Δz scheme can not be expected to preserve the effects of such correlations, as the set of N_{tomo} Δz nuisance parameters are sampled independently from their corresponding priors in the Monte-Carlo chain. This has been previously mentioned as a potential source of tension between DES Y1 and KiDS results (Joudaki et al., 2020). A way to address the issue of correlated uncertainty between tomographic bins is to sample

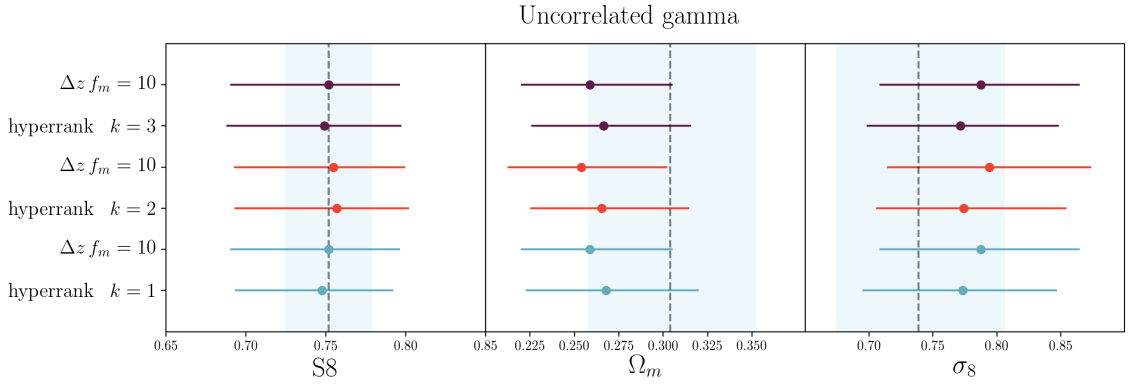

FIGURE 5.7: Summary plots showing the 1D marginalised $S8$, Ω_m and σ_8 error bars obtained from running the cosmic shear MCMC pipeline using HYPERRANK and Δz marginalisation schemes. The ensemble of $n(z)$ distributions used in hyperrank describes an uncertainty characterized by uncorrelated Δz shifts to the fiducial distribution $n_{\texttt{Fid}}(z)$, drawn from a Gamma distribution. Purple, red and blue error bars correspond to the non-Gaussian cases described by the k values 3, 2 and 1, respectively. Vertical dashed line and light blue filled region represent the 1-σ error bars from running the MCMC assuming no redshift uncertainty with the fiducial redshift distribution $n_{\texttt{Fid}}(z)$.

the Δz values jointly from a multidimensional Gaussian distribution characterized by an adequate covariance matrix.

Drawing a value of the HYPERRANK parameter(s) in a chain jointly specifies the $n(z)$ to be used in all tomographic bins and preserves these correlations, which can potentially lead to tighter contours on the cosmological parameters since the space of Δz values is restricted to those allowed by the samples. Depending on the sign of the correlation, this can also result in a shift of the contours if the Δz values favour a combination of high or low mean redshift only (positive correlation), instead of a combination of low and high mean redshift (negative correlation).

To explore the potential effects of these correlations at the tomographic bin level on inferred cosmological parameters, we generate three sets of mean-shifted realisations of the fiducial $n_{\texttt{Fid}}(z)$ with values of Δz sampled from a covariance matrix with increasing correlation between arbitrarily selected tomographic bin pairs. To generate the covariance matrices which describe these correlations we separate the correlation into two pairs of bins: (1,2) and (3,4). This way we can control the correlation via the Pearson correlation coefficient:

$$\rho_{i,j} = \frac{cov(\Delta z_i, \Delta z_j)}{\sigma(\Delta z_i)\sigma(\Delta z_j)} \tag{5.3}$$

We arbitrarily set $\rho_{1,2} = \rho_{3,4} = \{0.25, 0.5, 0.75\}$ and generate each pair of correlated Δz values independently making sure their final $\sigma(\Delta z)$ correspond to the ones for the desired prior. To better visualise the effects of these correlations once again we use an amplified $\sigma(\Delta z)$ prior to describe the diagonal of the covariance matrix, equal to the $f_m = 10$ prior described in section 5.2.3. The correlations between Δz values can be seen in figure 5.8 for the three choices of ρ.

The effect of correlated uncertainties can be seen in figure 5.9, where small deviations

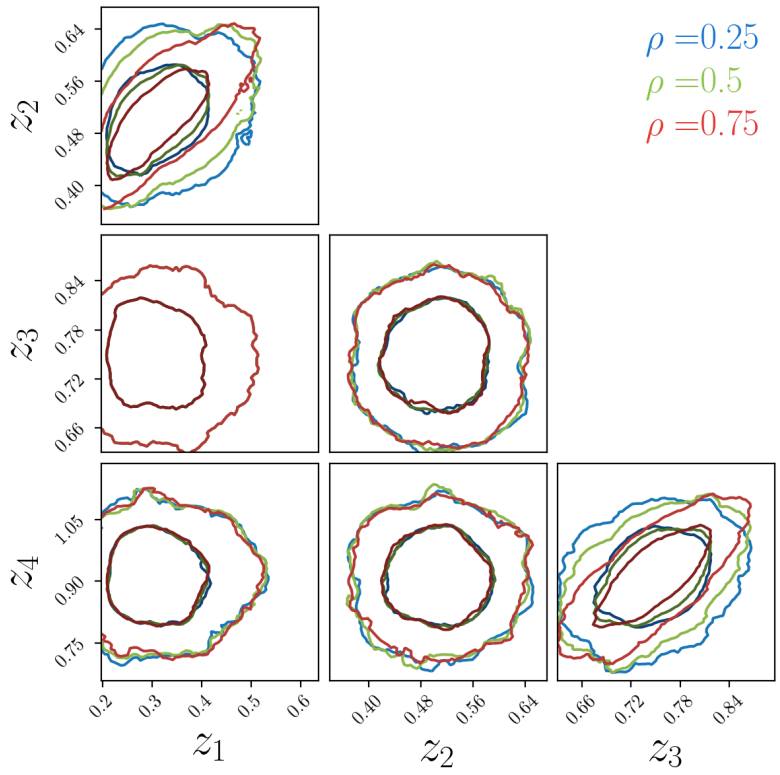

FIGURE 5.8: Corner plot showing the distribution of mean redshift values $\langle z \rangle$ from the ensemble of 1000 $n(z)$ distributions generated using a set of correlated Gaussian distributed Δz values. Blue, green and red contours show the distributions for the three levels of correlation described by the Pearson coefficients $\rho = \{0.25, 0.50 \text{ and } 0.75\}$, respectively. Despite their correlation, all samples of Δz have the exact same equivalent $\sigma(\Delta z)$ widths. All ensembles of distribution have the same values for the Δz_1 and Δz_3, wich are used as the base to generate the correlated values Δz_2 and Δz_4 respectively. Hence, their distribution appear identical for all three ensembles.

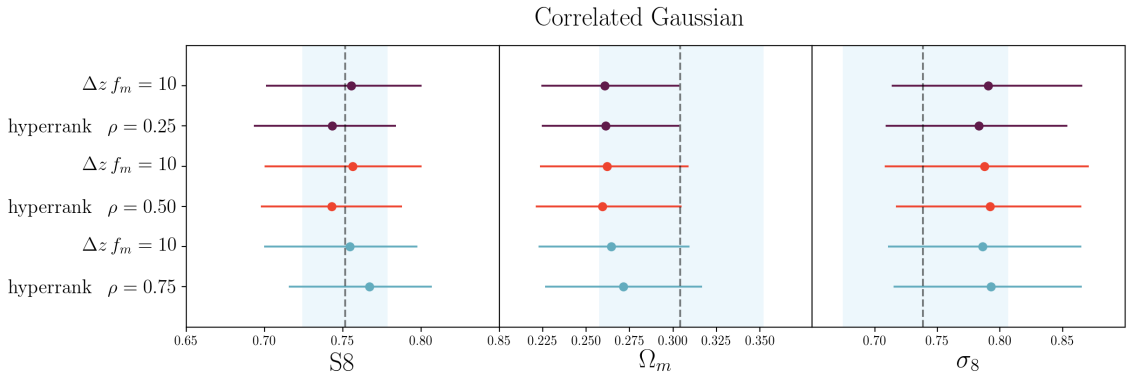

FIGURE 5.9: Summary plots showing the 1D marginalised $S8$, Ω_m and σ_8 error bars obtained from running the cosmic shear MCMC pipeline using HYPERRANK and Δz marginalisation schemes. The ensemble of $n(z)$ distributions used in hyperrank describes an uncertainty characterized by correlated Gaussian Δz shifts to the fiducial distribution $n_{\texttt{Fid}}(z)$. Purple, red and blue error bars correspond to the three levels of correlation described by the ρ values 0.25, 0.5 and 0.75, respectively. Vertical dashed line and light blue filled region represent the 1-σ error bars from running the MCMC assuming no redshift uncertainty with the fiducial redshift distribution $n_{\texttt{Fid}}(z)$.

on the $S8$ parameter ($\sim 0.3\sigma$) can be observed, above the typical sampling noise value expected for 1000 realisations, observed in section 5.2 It is worth noting that there does not seem to be a correlation between the deviation of the central values and the level of correlation, although for some tests done with an even larger correlation between the pairs of tomographic bins the contours obtained are heavily biased and their $1-\sigma$ confidence intervals shrink significantly. A notable feature of these results is the fact that the direction of the deviation of the $S8$ parameters changes to a higher value than the one obtained for the Δz approach when increasing the correlation between tomographic bins, to a Pearson correlation coefficient of 0.75. This is perhaps caused by a strong self-calibration effect, as the lower values of equivalent Δz on one tomographic bin are accompanied by a similarly low value on the corresponding pair, resulting in low posterior values each time that part of the Δz space is visited. In all cases, $\rho = 0.25 - 0.75$ are relatively large values compared to the correlation observed for the 500 BUZZARD realisations $\rho \sim 0.34$ for the bin pair with highest correlation), more in line with the values expected in real data.

Higher order modes of uncertainty

Finally, we create as set of realisations of $n(z)$ which represent a fully flexible model of the uncertainty in $n(z)$, following the approach in which histogram bin heights as a function of redshift within a tomographic bin are treated as the nuisance parameters to be inferred. These are the 500 realisations of possible $n(z)$ generated using the SOMPZ +3SDIR. As above for the different values of f_m, we additionally apply a procedure to these realisations to artificially increase the level of uncertainty they represent. Starting from the set of 500 realisations, we amplify the difference between each of the $n(z_i)$ values and the value of the fiducial distribution $n_{\texttt{Fid}}(z)$, such that $n'(z_i) = n(z_i) + \lambda\left[n(z_i) - n^f(z_i)\right]$. For this test we generate three sets of distributions: one with no amplification, $\lambda = 0$;

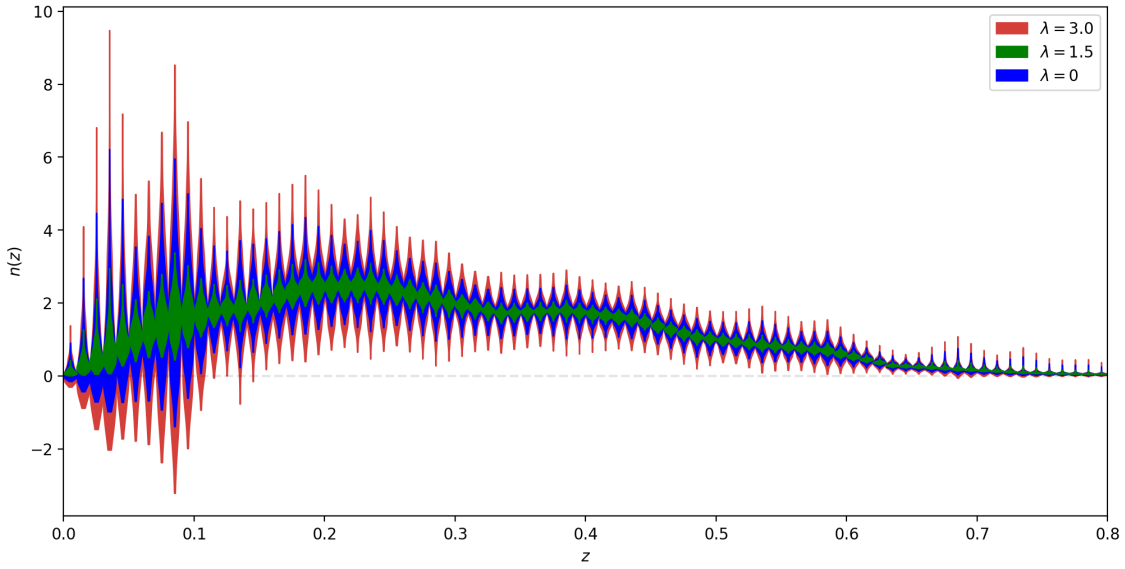

FIGURE 5.10: Violin plot showing an example of the amplified peculiarities for the three levels of amplification of tomographic bin 1, described by $\lambda = 0$ (No amplification, green), 1.5 (blue) and 3.0 (red).

and two with amplified peculiarities, $\lambda = \{1.5, 3\}$. While the average $n(z)$ obtained from the amplified realisations remains unaltered, this procedure can result in a slightly wider equivalent Gaussian prior $\sigma(\Delta z)$ to those of the un-amplified realisations. Thus, we also obtain the $\sigma(\Delta z)$ values for each set of distributions and use them to compare HYPERRANK to the standard Δz marginalisation. An illustration of this amplification can be seen in the violin plots presented in figure 5.10.

From figure 5.11 it can be seen that the effects on the $S8$ parameters are fairly small, compared to the typical sampler noise seen for 500 realisations in the $f_m = 1$ case. Contours for the $\lambda = 0$, consistent with the typical uncertainties obtained using the SOMPZ scheme on BUZZARD show very little differences to the contours obtained the standard Δz marginalisation approach. The case for larger amplification values the contours appear to shrink slightly in $S8$, by a larger factor than the typical value expected from sampling noise.

5.3 Summary

In chapter 4 we presented the formalism for a new approach to marginalising redshift systematics in weak lensing and galaxy clustering experiments by ranking and mapping a set of proposal $n(z)$ distributions encoding the photometric redshift uncertainty to a continuous hyper-parameter which is then sampled in the MC chain. In this chapter we showed that this approach, while resulting in a $\sim$ 50% penalty in total number of likelihood evaluations, provides an equivalent result to standard marginalisation using nuisance parameters when the uncertainty can be described by just a shift in mean redshift Δz. It can also marginalise over higher order uncertainties caused by method systematics, photometric errors or other sources of error which are difficult to parameterize

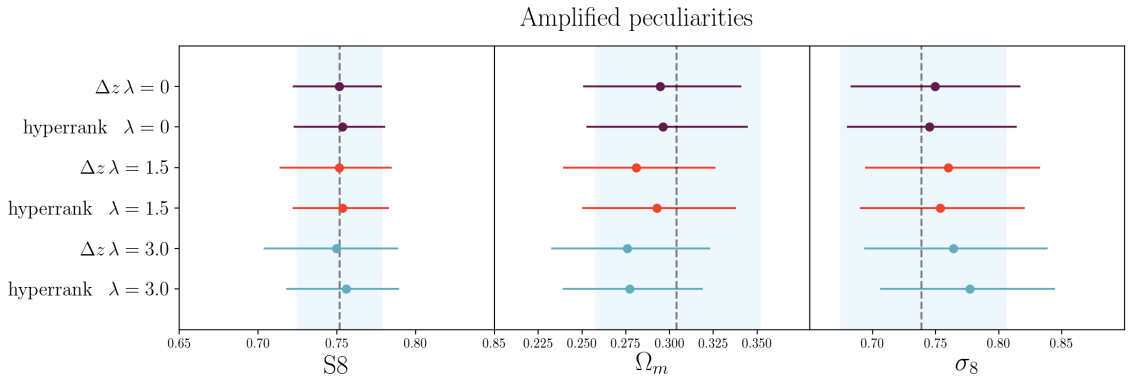

FIGURE 5.11: Summary plots showing the 1D marginalised $S8$, Ω_m and σ_8 error bars obtained from running the cosmic shear MCMC pipeline using HYPERRANK and Δz marginalisation schemes. The ensemble of $n(z)$ distributions used in hyperrank describes an uncertainty characterized by amplified peculiarities of the SOMPZ +3SDIR $n(z)$ ensemble. Purple, red and blue error bars correspond to the three amplification cases described by the λ values 0, 1.5 and 3.0, respectively. Vertical dashed line and light blue filled region represent the 1-σ error bars from running the MCMC assuming no redshift uncertainty with the fiducial redshift distribution $n_{\texttt{Fid}}(z)$.

to be marginalised as nuisance parameters. Additionally, this approach preserves the correlation between these effects at the tomographic and histogram bin level which, if unaccounted for, can lead to an overestimation of the confidence levels in cosmological parameters, previously ignored in weak lensing experiments.

We used this approach to test the effect of these high order variations on the redshift distributions used for weak lensing by generating a series of $n(z)$ ensembles representative of different potential descriptions of systematics typically ignored when using nuisance parameters to describe uncertainty as a Δz shift in the mean redshift of each tomographic bin. We showed that while there are observable differences in the resultant contours exceeding their expected variance caused by sampling noise, these are only appreciable when uncertainty is large in comparison to the expected levels for current weak lensing experiments.

A set of tests were conducted to obtain an approximate optimal configuration for the descriptive values used to rank the distributions and its effect on sampling efficiency, resulting in the use of mean redshift of a sub-set of tomographic bins, $\langle z \rangle_n$ being the ranking with the lowest typical number of likelihood evaluations required for convergence.

In chapter 6 we will explore the performance of HYPERRANK for the particular levels of uncertainty expected for DES Y3 analysis and discuss its applicability to in the fiducial DES Y3 data analysis, based on the difference in obtained confidence contours with respect to the standard approach using Δz shifts.

Chapter 6

Results of using HYPERRANK on DES Y3 data

In chapters 3 and 4 we presented the plan to estimate the line-of-sight distribution of weak lensing sources using the SOMPZ scheme, quantify the associated uncertainty to these estimates and the strategy to propagate this uncertainty into the inferred cosmological parameters using HYPERRANK.

In this chapter we discuss the application of HYPERRANK to propagate uncertainty of the photometric redshift distributions into the cosmological parameter inference in the DES Y3 weak lensing analysis. In section 6.1 we briefly discuss the differences between HYPERRANK and standard Δz marginalisation approach by comparing the sampling efficiency and cosmological parameter contours obtained on a set of BUZZARD $n(z)$ realisations closely resembling the uncertainty expected for the DES Y3 real data. Then, in section 6.2 we describe the main DES Y3 cosmic shear pipeline, its data products and the application of HYPERRANK to the ensemble of realisations obtained using the SOMPZ scheme, presented in chapter 3. The results of applying HYPERRANK to the DES Y3 data realisations are also presented in Amon et al. (2020) in which the author of this thesis has contributed as co-author. The contributions of the author of this thesis include the configuration of the HYPERRANK module to accommodate for the $n(z)$ ensemble coming from the SOMPZ + 3SDIR scheme and configuration of the standard marginalization modules using Δz.

6.1 Forecasts on efficiency and correctness

Here we discuss the recommended baseline configuration for DES Y3 data analysis based on the results obtained in chapter 5 for the BUZZARD simulations, and specifically for the set of realisations that better describes the expected magnitude of photo-z systematics from real data. This configuration is being applied to real DES Y3 data, including the analysis of cosmic shear data (Amon et al., 2020; Secco et al., 2020) and cosmic shear plus galaxy-galaxy lensing and clustering (3x2pt; DES Collaboration et al., 2020). Given the magnitude of these systematics, we discuss the benefits of using HYPERRANK as an

alternative to marginalise photo-z uncertainty and whether the Δz approximation is still a compelling option given its better sampling efficiency.

6.1.1 Configuration

In chapter 4 we described HYPERRANK as a dimensional reduction of all the modes of uncertainty described by a set of N_p $n(z)$ realisations, into a set of n hyper-parameters $\mathcal{H} \in [0,1)^n$ using a set of descriptive values $\mathbf{d}$. These hyper-parameters can be sampled continuously by the Monte Carlo sampler and act as the set of nuisance parameters to marginalise from the joint posterior distribution. This reduction is then a function of three main components: N_p, n, and $\mathbf{d}$, from which we can only tune the last two, while providing a minimum recommended value for N_p. This value is ultimately determined by the ability of the photometric redshift pipeline to generate the $n(z)$ samples.

We have argued in section 5.2.2 that the number of realisations N_p ultimately determines how well the different modes of uncertainty are described and has a large impact on the marginalised contours of cosmological parameters. For DES Y3, the 3sDir+HMC sampling scheme is able to provide enough samples to overcome this limitation so the only concern is whether we can find the positions of the $n(z)$ in the multi-dimensional grid in a reasonable time. For the level of uncertainties expected for DES Y3 ($f_m \sim 2$; see Myles et al., 2020) we expect $N_p \sim 1000$ realisations to be an adequate minimum number. Still, the linear sum assignment can to map up to 16^3 realisations in a time comparable to what a full MC chain takes to run, taking approximately a couple days on a single core. If the ensemble of distributions and the set of descriptive values $\mathbf{d}$ does not change, then the computation of the mapping can be done only once and then provided to each MC run as a pre-computed array. This is useful if multiple chains are run using a different datavector (e.g. 3x2pt analysis) or sets of priors (ωCDM versus ΛCDM).

Alternative approaches to speed-up the mapping of $n(z)$ realisations to the multi-dimensional hyper-cube include solutions of the linear sum assignment using parallel computing (e.g. Date & Nagi, 2016), or the use of simpler space discretization methods like the uneven grid assignment we presented in 4.2.2. We do not explore those methods here, although we expect alternative arrangements to only have a minor impact on sampling efficiency and not on the shape of the final contours.

The choice of descriptive values $\mathbf{d}$ and number of dimensions n plays a significant role in achieving optimal sampling efficiency. Results from our tests presented in 5.2.1 suggest that choosing an appropriate set of descriptive values can result in a reduction in likelihood evaluations over a purely random approach (See 5.2), and moving from a single dimension to 3 has a large effect too. From those results we find that while $\langle z \rangle$ and $\langle 1/\chi \rangle$ trade the lead in terms of efficiency, for the case that better represents the set of uncertainties expected for DES Y3 the use of three tomographic bin mean redshifts from the bins with the highest expected $\sigma(\Delta z)$ show the best sampling efficiency, and we suggest it to be used in DES Y3 data.

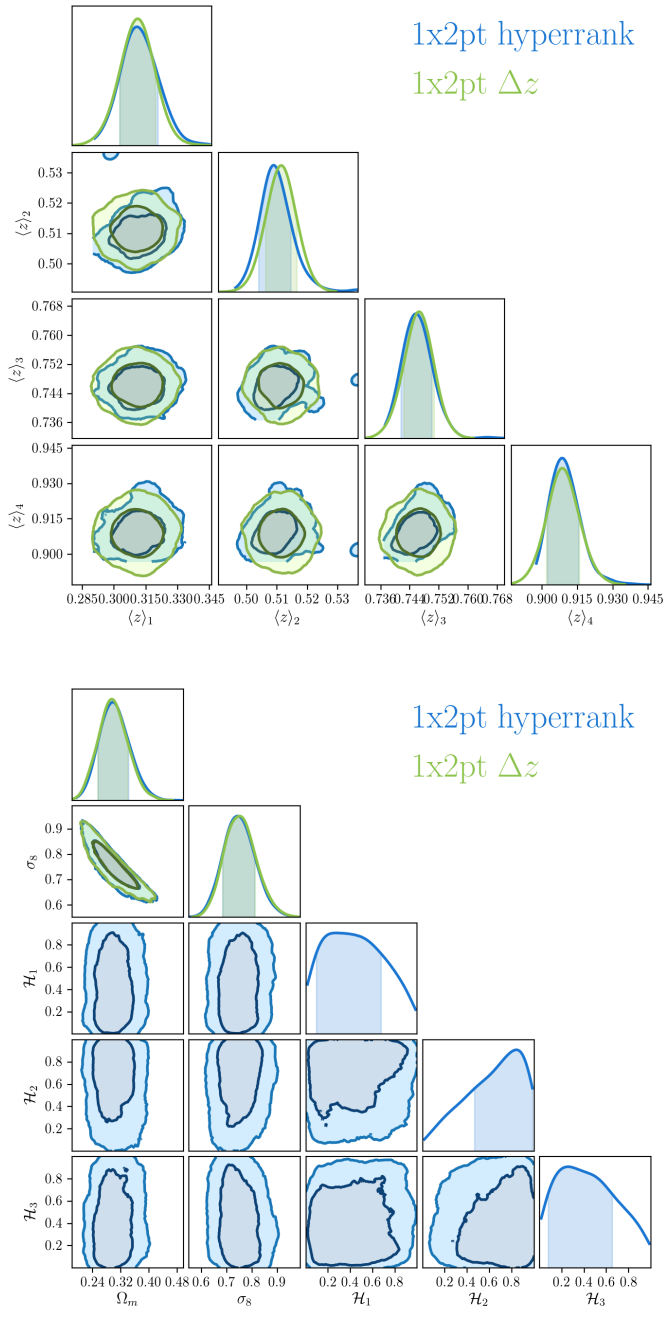

FIGURE 6.1: Figure showing the correlation between redshift distribution uncertainty nuisance parameters in the BUZZARD simulated DES-Y3 analysis, comparing the standard Δz approach (red) with the HYPERRANK approach presented in this work (blue). *Top* shows the recovered posteriors on mean redshifts of redshift distributions within the tomographic bins considered. *bottom* shows the recovered cosmological parameters for both approaches, and the HYPERRANK ranking parameters. Both show good agreement between the two approaches for the modelled uncertainty expected in DES-Y3.

6.1.2 Forecasts

The left panel of 6.1 shows the distribution of Δz nuisance parameters, and the posterior-weighted mean redshifts of the $n(z)$ realisations used in HYPERRANK. This explicitly shows the important result of HYPERRANK exploring the same space of uncertainty as the Δz for these simulated data and realisations. The right panel of 6.1 shows the joint posteriors for σ_8 and Ω_m obtained from the ensemble of 500 $n(z)$ distributions obtained with BUZZARD, using an equivalent $\sigma(\Delta z)$ prior for the Δz scheme, and the full ensemble of realisations for HYPERRANK, with the mean redshift from tomographic bins 1, 2 and 4 as descriptive values in a 3D hyper-cube configuration (as in our suggested configuration). We can see that for the level of uncertainty expected for DES Y3, the differences between the Δz approach and HYPERRANK are minor, and consistent with the sampling noise for a similar set of realisations at an equivalent $\sigma(\Delta z)$ regime. We also show the joint posterior for the HYPERRANK parameters $\mathcal{H}$ which should be contrasted with the uniform prior on each $\mathcal{H}$, showing smooth variation of the posterior and favouring of a particular sub-set of $n(z)$ realisations. Together, the results in 6.1 indicate that the Δz model appears adequate for the redshift uncertainties which are thought to be present in the DES-Y3 data, meaning it may be used in place of HYPERRANK without bias or loss of constraining power on cosmological parameters. In this particular run, HYPERRANK requires close to 50% more likelihood evaluations per posterior sample than the Δz approach.

These results suggests that for the level of uncertainty expected for DES Y3, where the correlation between tomographic bins is expected to be small and the mean redshift values of each tomographic bin of the $n(z)$ ensemble are close to Gaussian-distributed, the Δz marginalisation approach is still a compelling option. Given these observations, it has been decided that the fiducial analysis presented in the DES Y3 publications will use Δz instead of HYPERRANK, but the comparison and test of robustness to the choice of systematic modelling will still be shown.

6.2 Application to DES Y3 data

The methodological infrastructure and robustness of the DES Y3 cosmic shear is presented in detail in Secco et al. (2020) and Amon et al. (2020). Secco et al. (2020) primarily deals with the robustness to the modeling of intrinsic alignments, baryon physics, high order lensing effects and neutrinos. It also presents consistency tests with external datasets. Amon et al. (2020) in turn presents the analysis choices and testing of methodologies related to shear measurements, image simulations and redshift distributions. Both of these are complementary works and present the constraints obtained using cosmic shear, which are combined with the estimates of galaxy clustering (Rodríguez-Monroy et al., 2020) and galaxy-galaxy lensing (Prat et al., 2020) into a joint 3x2pt analysis which will be presented in DES Collaboration et al. (2020).

The summary of the complete analysis can be given in context with the supporting publications of the DES Y3 analysis:

- New PSF modelling (Jarvis et al., 2020) combined with weak lensing shape measurement based upon METACALIBRATION (Sheldon & Huff, 2017) gives a catalogue of 100 million selected galaxies that are validated in Gatti et al. (2020b). The shear correlation functions $\xi_{\pm}$ are computed using TREECORR (Jarvis, 2015) in 20 angular bins logaritmically spaced between 2.5 and 250 arcmins.

- Redshift calibration methodology is summarised in Myles et al. (2020) and in chapter 3. This calibration scheme employs data from narrow-band photometric and spectroscopic sources and DES deep observations presented in (Hartley et al., 2020a) to estimate the line-of-sight distribution of wide sources using a transition matrix characterizing the deep-to-wide color relation, characterized by BALROG. The full scheme incorporates new independent methods, a two-step reweighting with self-organising maps (Buchs et al., 2019), clustering redshifts (Gatti et al., 2020a), and small-scale shear ratios (Sánchez et al., 2020a), to precisely constrain the redshift distributions.

- Alternative techniques for modelling and marginalising over the uncertainty on the tomographic distributions are tested including HYPERRANK (Cordero et al., 2020), presented in chapter 4. The method is compared against a standard marginalisation approach where uncertainty is described by a set of independent tomographic bin shiftt Δz, drawn from a Gaussian prior computed from the uncertainty presented in table 3.1.

- State-of-the-art shear calibration with realistic image simulations and new methodology to account for the impact of blending on the effective redshift distribution for lensing measurements in MacCrann et al. (2020a).

- The general methodology, likelihood analysis and covariance are presented in Krause et al. (2020) and Friedrich et al. (2020) independently validated using realistic simulations in DeRose et al. (2020). The MCMC chains are run using COSMOSIS (Zuntz et al., 2015), a modular cosmological analysis software.

- The statistical framework to assess the internal consistency of the DES data and measurements is presented in (Doux et al., 2020) and the consistency with independent, external data in (Lemos et al., 2020).

Three main aspects of the redshift calibration pipeline are to be tested in Amon et al. (2020): (i) the robustness to the choice of redshift information to inform the SOMPZ scheme (spectra, COSMOS, COSMOS+PAU), (ii) the consistency between the three independent methods used to constrain the redshift distributions (SOMPZ, clustering redshifts and shear rations), and (iii) robustness to the how the uncertainty is modelled and propagated (Δz vs HYPERRANK).

We are primarily interested in the robustness to the choice of uncertainty modelling, and whether our prediction that for the level of uncertainty expected for DES Y3 data the use of a Δz marginalisation provides similar results to those of using HYPERRANK

is correct. We then proceed to compare the cosmological parameters inferred using the fiducial choices of the DES Y3 cosmic shear analysis. These include parameter boundaries and priors described in table 5.1. The ensemble of redshift distributions consists of 1000 realisations obtained using the 3SDIR + HMC sampling procedure presented in section 3.2.2 and depicted in figure 3.8. Following the recommended configuration proposed in chapter 5, we map the mean redshift from tomographic bins 1,2 and 4 to three hyper-parameters, since these tomographic bins are the three with the largest equivalent uncertainty $\sigma(\Delta z)$. The standard marginalisation Δz nuisance parameters are drawn from Gaussian priors described by $\sigma(\Delta z)$ from table 3.1.

Figure 6.2 shows the cosmological parameters inferred from the two marginalisation approaches described above. Both runs employ the same datavector describing the two-point shear correlation functions and the same sets of nuisance parameters describing intrinsic alignments and multiplicative shear biases. It can be seen that the inferred cosmology is robust to the choice of marginalisation scheme. Figure 6.3 shows a comparison of the sampled uncertainty for each case, where the consistency between the standard Δz approach and HYPERRANK can also be seen. These results confirm the predictions made in 6.1: modelling only the mean of the redshift distribution captures the full effect of photo-z bias uncertainty as analysed with HYPERRANK, illustrating that differences in the shape of the redshift distribution are sub-dominant for cosmic shear at the uncertainty of DES Y3.

6.3 Summary

In this chapter we have briefly discussed the applicability of the HYPERRANK scheme to propagate source redshift distribution uncertainty in the DES Y3 cosmic shear analysis. While HYPERRANK allows us to include higher order uncertainty effects than the standard Δz approach, the observed differences in inferred cosmological parameters for the levels of uncertainty expected for DES Y3 are small. This, in addition to the increased number of likelihood evaluations required to achieve convergence suggests that the use of nuisance parameters describing uncertainty as an independent shift Δz in the z direction for each tomographic bin is a good enough approximation to be the fiducial choice of marginalisation.

To test the robustness of this conclusion, we compared the two marginalisation approaches on real data form the DES Y3 cosmic shear analysis and indeed confirm that the inferred cosmology is robust to the choice of uncertainty modelling.

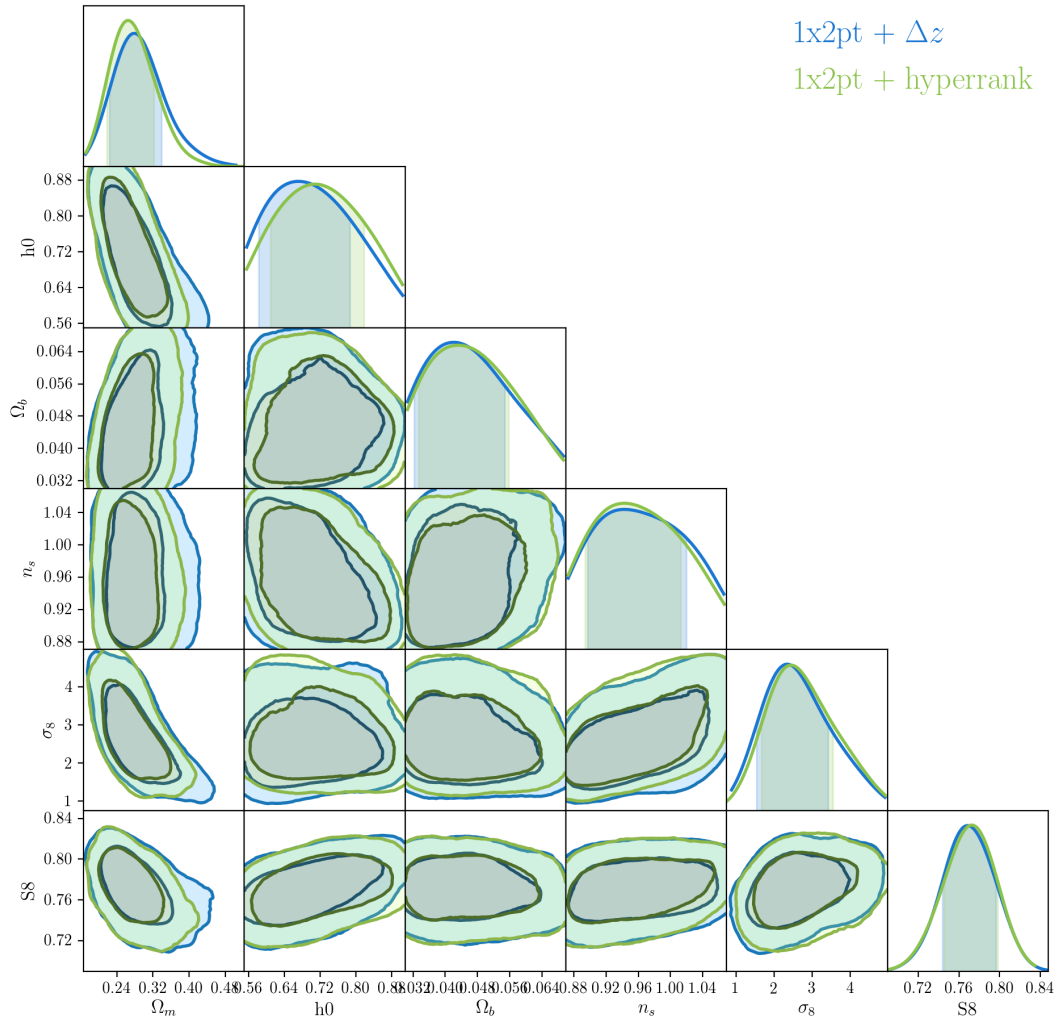

FIGURE 6.2: Comparison of DES Y3 cosmic shear cosmological parameter constrains obtained using HYPERRANK (blue) versus using Δz to model uncertainty. Consistency of the contours shows that the Δz is a good approximation for the level of uncertainty expected for DES Y3. The slightly tighter posteriors can be explained as a consequence of the correlations between tomographic bins not considered by the standard Δz approach, and the slightly narrower Δz posteriors seen in figure 6.3

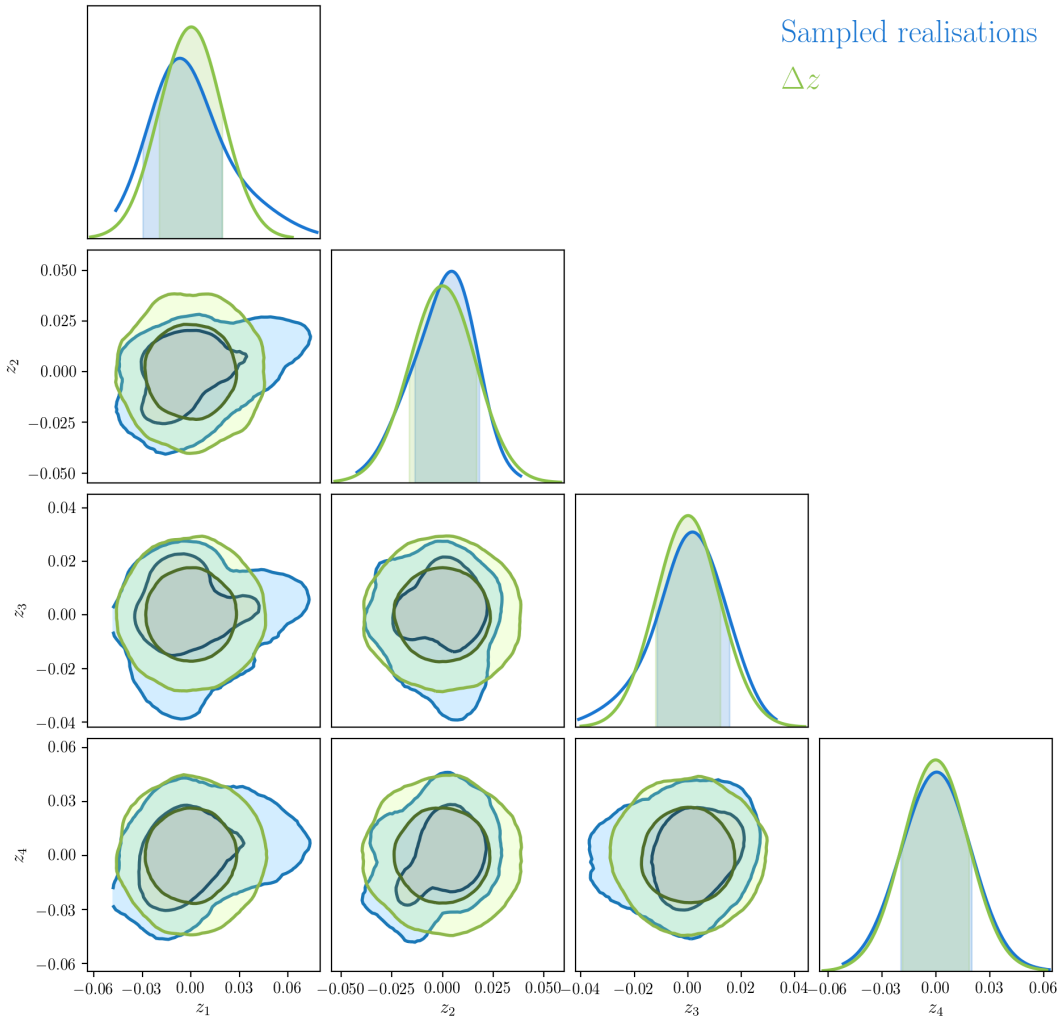

FIGURE 6.3: Comparison between the sampled Δz nuisance parameters and the distribution of shifts from the sampled realisations using HYPERRANK with respect to the fiducial realisation $n_{\texttt{Fid}}(z)$. The shifts for the HYPERRANK chain are computed as $\Delta z_{\text{Hyp}} = \langle n_{\text{Hyp}}(z) \rangle - \langle n_{\texttt{Fid}}(z) \rangle$ for all of the sampled realisations in the HYPERRANK chain. Figure shows that HYPERRANK explores uncertainty of the mean redshift of the tomographic bins similarly to the Δz approach.

Chapter 7

Final remarks

In this chapter we give a brief overview of the work presented in this thesis along with some remarks regarding specific areas where the approaches employed can be improved, and the perspectives for their importance in the wider field.

7.1 Summary and remarks about the current work

In chapter 1 we presented the basic concepts of the standard model of cosmology, which describes the evolution of the largest structures in the Universe and the formalism to statistically describe their distribution. This evolution is driven by the interactions between the Universe components within the fabric of space-time where they are embedded, as described by General Relativity. Since a large fraction of the energy distribution is believed to come from two components for which we have little knowledge about their origins, dark energy and dark matter, understanding their impact on the evolution of the large-scale structure is a necessary step to unveil their fundamental properties. We also described the observational probes at our disposal, which allow us to constrain the properties of the large-scale structure by observing the effects of its distribution both at high and low redshift.

In chapter 2 we described the formalism of weak gravitational lensing, one of the most rapidly evolving techniques to characterize the large-scale structure in the late Universe. By measuring the correlated shapes of observed distant galaxies sheared by the gravitational lensing effect of the large structures, weak lensing is able to directly probe the mass distribution irrespective of its nature (Dark or Baryonic). Because the effect of cosmic shear is very small compared to the intrinsic shapes of galaxies, its effect can only be detected by measuring the shapes of many galaxies and under the assumption that they do not orientate towards any preferential direction. In order to determine the relative lensing effect of the large-scale structure along the line of sight, the distribution of distances to the lensed sources must be obtained accurately. The determination of this distribution using photometric redshifts is a major source of uncertainty, and modest biases in the low order moments of the tomographic distribution can potentially lead to

large biases in the derived cosmological parameters (Huterer et al., 2006), which can potentially lead to incorrect model selection to describe the importance of Dark Energy and Dark Matter.

In chapter 3 we described a family of techniques to estimate these distances using measurements of the flux from distant galaxies in many bands, called photometric redshifts. Photometric redshifts are a suitable alternative to spectroscopy since they provide efficient estimations for surveys containing millions of galaxies while not suffering from the same incompleteness effects at faint magnitudes. We described the main types of photometric redshift codes and introduced the SOMPZ scheme, a machine learning algorithm to estimate photometric redshift distributions using Self-organizing maps, developed for use in the analysis of the first three years of observations of the Dark Energy Survey (DES Y3). We described the identified sources of systematic effects in the photometric redshift pipeline of the DES Y3 analysis, and described the methods to estimate its associated uncertainty in the $n(z)$ distributions, presenting the result on the effects of sample variance and the uncertainty associated to the random training of SOMPZ scheme. Then we described the 3SDIR + HMC, a three step Dirichlet sampling method to incorporate the estimates from the SOMPZ scheme, clustering redshifts and the measured uncertainty. This method generates samples from the redshift distribution posterior which can be used to describe the full uncertainty of the photometric redshift estimation pipeline.

In chapter 4 we introduced HYPERRANK, a novel technique to propagate the uncertainty encoded as an ensemble of $n(z)$ realisations like the one obtained using the 3SDIR + HMC sampling into the cosmological parameter estimates. To put HYPERRANK in context, we described the process of Bayesian inference, the use of nested samplers to draw samples from a high dimensional posterior distribution, and how some conditions can impact sampling efficiency. HYPERRANK maps the proposal distributions to a set of continuous hyper-parameters which are then marginalized in the Monte Carlo sampling pipeline. The aim is to generate a transformation between some descriptive values of the $n(z)$ ensemble and a set of hyper-parameters which allows the sampler to efficiently explore the set of distributions, while being able to propagate descriptions of uncertainty typically neglected by simpler nuisance parameter approaches.

In chapter 5 we devised a series of test to validate this new technique and to compare its performance against a standard marginalisation approach using a set of nuisance parameters describing a shift Δz of the redshift distributions along the redshift direction. We applied these tests to a ensemble of $n(z)$ distributions obtained by applying the SOMPZ scheme to BUZZARD, a suite of N-body simulations that replicate the expected observational conditions of DES. A series of $n(z)$ ensembles were also generated to describe different types of uncertainty and the ability of HYPERRANK to propagate all of them to the inferred cosmological parameters. We investigated the optimal set of descriptive parameters and number of hyper-parameters for a set of realisations described by different levels of uncertainty, including similar uncertainty to the one expected in the DES Y3 analysis.

In chapter 6 we discussed the possibility that HYPERRANK becomes the fiducial marginalisation approach for the DES Y3 analysis, considering the sampling efficiency penalty introduced in comparison with the standard marginalisation and given the small observed differences in inferred cosmological parameter contours. We concluded that for the level of uncertainty expected for DES Y3, the standard approach to marginalize uncertainties using a shift in the redshift distributions, Δz, is still a valid approximation. Finally we briefly described the DES Y3 cosmic shear pipeline and presented the results of applying HYPERRANK to propagate the source redshift distribution uncertainties of real DES Y3 data.

Perhaps one immediate observation that can be made is the fact that the HYPERRANK approach can be applied to any type of systematic to be marginalized, as long as a procedure to generate proposal samples exist. An example of this is presented in (MacCrann et al., 2020a) and (Amon et al., 2020) where a set of multiplicative shear biases m are inferred from each $n(z)$ realisation. On the other hand, it is harder to justify the use of a scheme like HYPERRANK for models which can be easily encoded as single parameters, where arbitrary models of uncertainty are easier to sample from. Cases where this can be useful is when uncertainty is associated to a heterogeneous systematic, such as the different sets of spectroscopic samples used for calibration of the SOMPZ scheme, or when the uncertainty of a given systematic effect is characterized by multiple parameters as is the case of, for example, the intrinsic alignment model used in DES Y3 which utilizes five parameters. The caveat is that for sets of parameters covering different numeric ranges the descriptive values $\mathbf{d}$ and mapping scheme have to be carefully chosen.

On the topic of mapping schemes, we have presented two alternatives which have the commonality of mapping the descriptive values $\mathbf{d}$ to the set of hyper-parameters $\mathcal{H}$ such that there is a monotonic relation between d_i and $\mathcal{H}_i$. An third alternative was considered early on the development of HYPERRANK, where we used Self-organized maps (SOM, the same machine learning algorithm used for SOMPZ) to arrange distributions into a m-dimensional grid in a similar way to the linear sum assignment and uneven grid work. One disadvantage of this approach was that the SOM scheme generated a grid to which distributions were not assigned in a one-to-one basis, but using a metric which could potentially assign more than one distribution to a single grid point. While each hyper-parameter $\mathcal{H}$ can be easily assigned to the coordinates of the SOM map, an additional step would be required to select a realisation from a cell with multiple $n(z)$ realisations, or to provide one when the sampled cell is empty. An interesting follow-up to this project can explore more efficient ways to assign realisations to the uniform grid. In its current fiducial implementation HYPERRANK uses the solution to the linear sum assignment problem using the Hungarian algorithm (Kuhn, 1955a) as implemented by SCIPY. This algorithm scales as $\mathcal{O}(n^3)$, and quickly becomes unmanageable for more than a few thousand realisations[1]. The uneven grid method presented in 4.2.2 is significantly faster, but suffers from poor ranking performance for higher dimensions. This can be potentially harmful in cases where many more bins are used for a tomographic

[1]The method required 14 hours to find the solution for a set of 5000 realisations

analysis, such as in the case of the DES Y3 lens sample redshift calibration for galaxy clustering and galaxy-galaxy lensing where up to 6 tomographic bins are used (Porredon et al., 2020, prep) although the effects of these mapping differences on sampling efficiency or correctness were not tested on this thesis.

Another aspect not explored in detail in this thesis is the fact that the standard Δz approach, described as an approximation of the total uncertainty of the source redshift distributions, can be adapted to marginalize over more complex representations of the uncertainty. Correlations between tomographic bins can be modelled by a covariance matrix and the individual Δz_i values for each tomographic bin sampled using its Cholesky decomposition. Arbitrary non-Gaussian Δz distributions can be sampled directly with CosmoSIS by providing a tabulated normalized prior $\pi(\Delta z)$. We have plans to test these modifications in (Cordero et al., 2020) before publication.

7.2 Prospects for the future

At the time of writing this thesis, the Dark Energy Survey is preparing to submit a series of publications describing the analysis of the first three years of observations carried out to measure the cosmic shear, galaxy clustering and galaxy-galaxy lensing signals from more than 4,000 square degrees of sky. As stated in Abbott et al. (2019a), where constrains of the dark matter equation of state, matter density and amplitude of fluctuations of the matter density field were obtained from a single homogeneous set of calibrated data, optical surveys are starting to rival estimates obtained from early Universe made from the Cosmic Microwave Background. With triple the area covered and close to four times the number of observed galaxies than Y1, DES Y3 is expected to produce one of the most competitive constraints of the standard model parameters, setting the ground for the final analysis to be carried out on the full six years of observations, which perhaps will mark the end of the Stage III era of cosmological surveys. Many of the novel techniques applied at this stage will carry on to become standard procedures in large synoptic surveys of Stage IV, expected to begin operations during this decade.

A particular aspect which will surely continue to generate debate is the use of spectroscopy and photometric redshifts in the analysis of weak lensing and galaxy clustering. The calibration scheme presented in DES Y3 employs a hybrid approach where both types of redshift estimates are used simultaneously. Still being too early to conclude whether this choice helps reconciling the observed tensions between similar early universe probes, we remain expectant to what the final published results and their interpretations will be.

Regardless of what aspect of the analysis is to be tested and validated, a framework to generate realistic simulations of the observed properties of optical surveys will continue to be a cornerstone of any similar analysis. In the case of DES, BUZZARD (DeRose et al., 2020) has been an invaluable tool providing high quality simulated catalogs. With the advent of more complex machine learning techniques and the need to generate reliable training data, simulations will continue to play this fundamental role. This is equally

true with new approaches to tackle the inference of model parameters such as the *forward modelling* process, which depends on the fine tuning of simulation parameters to replicate observed data. Significant progress has been made in this regards with alternative estimates of redshift distributions (Herbel et al., 2017) and weak lensing observables (Bruderer et al., 2018), In this context, the SkyPy package (SkyPy Collaboration et al., 2021) arises as a promising tool to generate powerful simulations to serve as testing ground of future optical surveys.

As Stage IV surveys begin their operations, characterization of systematics associated to their methodologies will become more important, as the statistical uncertainty becomes less dominant thanks to the large datasets which they will generate. Because of the overlap in terms of sky coverage and time of expected operation, these surveys will benefit from each other, joining forces to pursue common scientific goals. As an example, the large photometric redshift catalogs expected to be obtained from the Legacy Survey of Space and Time (LSST) can be used to enable studies of radio weak lensing from the Square Kilometer Array (SKA), while estimates of spectroscopic information of HI emission can be used to calibrate photo-z estimates from the LSST (Bacon et al., 2015). Many more combinations like this will result in a significant increase in the statistical power of large cosmological surveys, as independent observables will help break degeneracies. Several approximations made over different aspects of the analysis will have to be reviewed in order to minimize their effect on the inferred cosmology and in some cases the heterogeneous origin of the sources of data and their uncertainty will require complex tools to propagate them into the cosmological parameters. Hopefully HYPERRANK will help us move into that direction.

Appendix A

Deflection angle from Fermat's principle

Finding the deflection angle is a variational problem that starts with looking for the path, $\mathbf{x}(l)$, for which

$$\delta \int_a^b \eta(l)\mathbf{x}(l)dl = 0 \tag{A.1}$$

where a and b are fixed endpoints. We let

$$dl = \left|\frac{d\mathbf{x}}{d\chi}\right| \tag{A.2}$$

where χ is the comoving distance which we use it as curve parameter. Equation A.1 becomes

$$\delta \int_{\chi_a}^{\chi_b} L(\mathbf{x}, \dot{\mathbf{x}}, \chi)d\chi = 0 \tag{A.3}$$

and L is a Lagrangian

$$L(\mathbf{x}, \dot{\mathbf{x}}, \chi) = \eta(\mathbf{x}(\chi)) \left|\frac{d\mathbf{x}}{d\chi}\right| \tag{A.4}$$

which satisfies the Euler-Lagrange equations:

$$\frac{d}{d\chi}\frac{\partial L}{\partial \dot{\mathbf{x}}} = \frac{\partial L}{\partial \mathbf{x}} \tag{A.5}$$

Since $\frac{\partial L}{\partial \dot{\mathbf{x}}} = |\dot{\mathbf{x}}|\frac{\partial \eta}{\partial \mathbf{x}} = \nabla\eta|\dot{\mathbf{x}}|$ and $\frac{\partial L}{\partial \dot{\mathbf{x}}} = \eta\frac{\dot{\mathbf{x}}}{|\dot{\mathbf{x}}|}$, we have

$$\frac{d}{d\chi}(\eta\dot{\mathbf{x}}) - \nabla\eta = 0 \tag{A.6}$$

which implies

$$\eta\ddot{\mathbf{x}} = \nabla\eta - \dot{\mathbf{x}}(\nabla\eta \cdot \dot{\mathbf{x}}) \tag{A.7}$$

The right hand side is equal to the gradient of η perpendicular to the light path since the second term of the right hand side is equivalent to the derivative along the light path.

Therefore,

$$\ddot{\mathbf{x}} = \frac{1}{\eta}\nabla_{\perp}\eta = \nabla_{\perp}\ln(\eta) \tag{A.8}$$

Recalling that

$$\eta = \frac{c}{c'} = \frac{1}{1+\frac{2\Phi}{c^2}} \approx 1 - \frac{2\Phi}{c^2} \qquad \text{with} \qquad \frac{\Phi}{c^2} \ll 1 \tag{A.9}$$

Then equation A.8 becomes

$$\ddot{\mathbf{x}} \approx \frac{-2}{c^2}\nabla_{\perp}\Phi \tag{A.10}$$

Since the new deflection angle satisfies $\hat{\boldsymbol{\alpha}} = \int_{\chi_a}^{\chi_b} \ddot{\mathbf{x}} d\chi$ we can therefore write the deflection angle as

$$\hat{\boldsymbol{\alpha}} = -\frac{2}{c^2}\int \nabla_{\perp}\Phi d\chi \tag{A.11}$$

Appendix B

Convergence as a function of density contrast

We wish to find the connection between the convergence field $\kappa(\boldsymbol{\theta}, \chi)$ (Equation 2.19) and the matter power spectrum $P_\delta(\ell)$. We recall that the convergence κ is defined from the lensing potential Ψ as

$$\kappa = \frac{1}{2}\nabla\Psi = \frac{1}{2}\left(\frac{\partial^2}{\partial\theta_1^2} + \frac{\partial^2}{\partial\theta_1^2}\right)\Psi, \tag{B.1}$$

where

$$\Psi(\boldsymbol{\theta}, \chi) = \frac{2}{c^2}\int_0^\chi d\chi' \frac{f_K(\chi - \chi')}{f_K(\chi')f_K(\chi)}\Phi(f_K(\chi)\boldsymbol{\theta}, \chi'), \tag{B.2}$$

Combining B.1 and B.2 we obtain

$$\kappa(\boldsymbol{\theta}, \chi) = \frac{1}{c^2}\int_0^\chi d\chi' \frac{f_K(\chi - \chi')}{f_K(\chi)} f_K(\chi')\left[\frac{\partial^2}{\partial\theta_1^2} + \frac{\partial^2}{\partial\theta_1^2}\right]\Phi(f_K(\chi)\boldsymbol{\theta}, \chi'). \tag{B.3}$$

The terms in the square brackets are the two first components of the Poisson equation

$$\Delta\Phi = 4\pi G\rho \tag{B.4}$$

which, for a matter dominated universe, can be written in terms of the density contrast parameter δ and the matter density Ω_m as

$$\Delta\Phi = \frac{3H_0^2}{2a}\Omega_m\delta \tag{B.5}$$

We can add the third term $\partial^2\Phi/\partial\chi'^2$ to equation B.3 which vanishes from the full integral as it averages out from the integral along the line of sight χ', and use B.5 to obtain

$$\begin{aligned}\kappa(\boldsymbol{\theta}, \chi) &= \frac{3H_0^2\Omega_m}{2c^2}\int_0^\chi d\chi' \frac{f_K(\chi - \chi')}{a(\chi')f_K(\chi)} f_K(\chi')\delta(f_K(\chi)\boldsymbol{\theta}, \chi') \\ &= \frac{3H_0^2\Omega_m}{2c^2}\int_0^\chi \frac{d\chi'}{a(\chi')} W(\chi, \chi')\delta(f_K(\chi)\boldsymbol{\theta}, \chi').\end{aligned} \tag{B.6}$$

Bibliography

Abbott T., et al., 2016, Physical Review D, 94, 022001

Abbott T. M. C., et al., 2018a, Phys. Rev. D, 98, 043526

Abbott T. M. C., et al., 2018b, Physical Review D, 98, 043526

Abbott T. M. C., et al., 2019a, Physical Review Letters, 122, 171301

Abbott T. M. C., et al., 2019b, Monthly Notices of the RAS, 483, 4866

Abbott T. M. C., et al., 2019c, Astrophysical Journal, Letters, 872, L30

Alam S., et al., 2015, Astrophysical Journal, Supplement, 219, 12

Alarcon A., Sánchez C., Bernstein G. M., Gaztañaga E., 2020, Monthly Notices of the RAS, 498, 2614

Albrecht A., Steinhardt P. J., 1982, Physical Review Letters, 48, 1220

Albrecht A., et al., 2006, arXiv e-prints, pp astro–ph/0609591

Alimi J.-M., et al., 2012, arXiv e-prints, p. arXiv:1206.2838

Amon A., et al., 2020, To be submitted to

Anderson T., 2003, An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Statistics, Wiley, https://books.google.de/books?id=Cmm9QgAACAAJ

Ata M., et al., 2018, Monthly Notices of the RAS, 473, 4773

Babcock H. W., 1939, Lick Observatory Bulletin, 498, 41

Bacon D. J., Refregier A. R., Ellis R. S., 2000, Monthly Notices of the RAS, 318, 625

Bacon D., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 145 (arXiv:1501.03977)

Bartelmann M., Schneider P., 2001, Physics Reports, 340, 291

Benítez N., 2000, Astrophysical Journal, 536, 571

Bennett C. L., et al., 1996, Astrophysical Journal, Letters, 464, L1

Berger E., Fong W., Chornock R., 2013, Astrophysical Journal, Letters, 774, L23

Bernstein G., 2020, in preparation

Bond J. R., Jaffe A. H., Knox L., 1998, Physical Review D, 57, 2117

Bonnett C., et al., 2016, Physical Review D, 94, 042005

Bruderer C., Nicola A., Amara A., Refregier A., Herbel J., Kacprzak T., 2018, Journal of Cosmology and Astroparticle Physics, 2018, 007

Buchs R., et al., 2019, Monthly Notices of the RAS, 489, 820

Bull P., 2016, Astrophysical Journal, 817, 26

Carrasco Kind M., Brunner R. J., 2013, Monthly Notices of the RAS, 432, 1483

Cawthon R., et al., 2020, To be submitted to MNRAS

Chakrabarti S. K., Jin L., Arnett W. D., 1987, Astrophysical Journal, 313, 674

Chevallier M., Polarski D., 2001, International Journal of Modern Physics D, 10, 213

Choi A., et al., 2016, Monthly Notices of the RAS, 463, 3737

Cole S., et al., 2005, Monthly Notices of the RAS, 362, 505

Collister A. A., Lahav O., 2004, Publications of the ASP, 116, 345

Cordero J. P., Harrison I., et al., 2020, To be submitted to MNRAS

Crill B. P., et al., 2003, The Astrophysical Journal Supplement Series, 148, 527–541

DES Collaboration et al., 2020, To be submitted to

DESI Collaboration et al., 2016, arXiv e-prints, p. arXiv:1611.00036

Date K., Nagi R., 2016, Parallel Computing, 57, 52

Davis C., et al., 2017, arXiv e-prints, p. arXiv:1710.02517

Dawson K. S., et al., 2013, Astronomical Journal, 145, 10

Dawson W. A., Schneider M. D., Tyson J. A., Jee M. J., 2016, Astrophysical Journal, 816, 11

De Vicente J., Sánchez E., Sevilla-Noarbe I., 2016, Monthly Notices of the RAS, 459, 3078

DeRose J., et al., 2019, arXiv e-prints, p. arXiv:1901.02401

DeRose J., et al., 2020, To be submitted to

Desjacques V., Jeong D., Schmidt F., 2018, Physics Reports, 733, 1

Di Valentino E., et al., 2021, arXiv e-prints, p. arXiv:2103.01183

Doux C., et al., 2020, arXiv e-prints, p. arXiv:2011.03410

Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Physics Letters B, 195, 216

Dyson F. W., Eddington A. S., Davidson C., 1920, Philosophical Transactions of the Royal Society of London Series A, 220, 291

Efstathiou G., 2020, arXiv e-prints, p. arXiv:2007.10716

Eisenstein D. J., et al., 2005, Astrophysical Journal, 633, 560

Erben T., et al., 2013, Monthly Notices of the RAS, 433, 2545

Eriksen M., et al., 2019, Monthly Notices of the RAS, 484, 4200

Everett S., et al., 2020, arXiv e-prints, p. arXiv:2012.12825

Famaey B., McGaugh S. S., 2012, Living Reviews in Relativity, 15, 10

Feroz F., Hobson M. P., Bridges M., 2009, Monthly Notices of the RAS, 398, 1601

Foëx G., Chon G., Böhringer H., 2017, Astronomy and Astrophysics, 601, A145

Friedrich O., Seitz S., Eifler T. F., Gruen D., 2016, Monthly Notices of the RAS, 456, 2662

Friedrich A., et al., 2020, To be submitted to

Garcia-Fernandez M., et al., 2018, Monthly Notices of the RAS, 476, 1071

Gatti M., Giannini G., et al., 2020a, To be submitted to MNRAS

Gatti M., et al., 2020b, arXiv e-prints, p. arXiv:2011.03408

Gaztanaga E., Schmidt S. J., Schneider M. D., Tyson J. A., 2021, Monthly Notices of the RAS,

Geman S., Geman D., 1984, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721

Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, Astrophysical Journal, 715, 823

Gruen D., Brimioulle F., 2017, Monthly Notices of the RAS, 468, 769

Gupta A., Mathur S., Krongold Y., Nicastro F., Galeazzi M., 2012, Astrophysical Journal, Letters, 756, L8

Guth A. H., 1981, Physical Review D, 23, 347

Hamana T., et al., 2020, Publications of the ASJ, 72, 16

Harrison I., Camera S., Zuntz J., Brown M. L., 2016, Monthly Notices of the RAS, 463, 3674

Hartlap J., Simon P., Schneider P., 2007, Astronomy and Astrophysics, 464, 399

Hartley W. G., Choi A., et al., 2020a, To be submitted to MNRAS

Hartley W. G., et al., 2020b, Monthly Notices of the RAS, 496, 4769

Hastings W. K., 1970, Biometrika, 57, 97

Herbel J., Kacprzak T., Amara A., Refregier A., Bruderer C., Nicola A., 2017, Journal of Cosmology and Astroparticle Physics, 2017, 035

Heymans C., et al., 2013, Monthly Notices of the RAS, 432, 2433

Heymans C., et al., 2021, Astronomy and Astrophysics, 646, A140

Hildebrandt H., et al., 2017, Monthly Notices of the RAS, 465, 1454

Hildebrandt H., et al., 2021, Astronomy and Astrophysics, 647, A124

Hoekstra H., Franx M., Kuijken K., Squires G., 1998, Astrophysical Journal, 504, 636

Hosseini R., Bethge M., 2009, Technical report, Spectral Stacking: Unbiased Shear Estimation for Weak Gravitational Lensing

Hoyle B., et al., 2018, Monthly Notices of the RAS, 478, 592

Hu W., 1999, Astrophysical Journal, Letters, 522, L21

Hu W., White M., 2001, Astrophysical Journal, 554, 67

Huff E., Mandelbaum R., 2017, arXiv e-prints, p. arXiv:1702.02600

Huterer D., Takada M., Bernstein G., Jain B., 2006, Monthly Notices of the RAS, 366, 101

Ivezić Ž., et al., 2019, Astrophysical Journal, 873, 111

Jõeveer M., Einasto J., Tago E., 1978, Monthly Notices of the RAS, 185, 357

Jaffe A. H., et al., 2001, Physical Review Letters, 86, 3475–3479

Jarvis M., 2015, TreeCorr: Two-point correlation functions (ascl:1508.007)

Jarvis M. J., et al., 2013, Monthly Notices of the RAS, 428, 1281

Jarvis M., et al., 2020, arXiv e-prints, p. arXiv:2011.03409

Jarvis M., et al., 2021, Monthly Notices of the RAS, 501, 1282

Joachimi B., Schneider P., Eifler T., 2008, Astronomy and Astrophysics, 477, 43

Joudaki S., et al., 2020, Astronomy and Astrophysics, 638, L1

Kaiser N., Squires G., Broadhurst T., 1995, Astrophysical Journal, 449, 460

Kaiser N., Wilson G., Luppino G. A., 2000, arXiv e-prints, pp astro–ph/0003338

Kamionkowski M., Kovetz E. D., 2016, Annual Review of Astron and Astrophys, 54, 227

Kirk D., Rassat A., Host O., Bridle S., 2012, Monthly Notices of the RAS, 424, 1647

Kirkwood J. G., 1935, Journal of Chemical Physics, 3, 300

Kitching T. D., Verde L., Heavens A. F., Jimenez R., 2016, Monthly Notices of the RAS, 459, 971

Kitching T. D., Alsing J., Heavens A. F., Jimenez R., McEwen J. D., Verde L., 2017, Monthly Notices of the RAS, 469, 2737

Köhlinger F., et al., 2017, Monthly Notices of the RAS, 471, 4412

Kohonen T., 1982, Biological Cybernetics, 43, 59

Kohonen T., 2001, Self-organizing maps, 3rd edn. Springer series in information sciences, 30, Springer-Verlag, Berlin, Heidelberg, doi:10.1007/978-3-642-56927-2

Krause E., Eifler T., Blazek J., 2016, Monthly Notices of the RAS, 456, 207

Krause E., et al., 2020, To be submitted to MNRAS

Krist J. E., Hook R. N., Stoehr F., 2011, in Kahan M. A., ed., Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 8127, Optical Modeling and Performance Predictions V. p. 81270J, doi:10.1117/12.892762

Kuhn H. W., 1955a, Naval Research Logistics Quarterly, 2, 83

Kuhn H. W., 1955b, Naval Research Logistics Quarterly, 2, 83

Kuijken K., et al., 2015, Monthly Notices of the RAS, 454, 3500

LSST Science Collaboration et al., 2009, arXiv e-prints, p. arXiv:0912.0201

Laigle C., et al., 2016, Astrophysical Journal, Supplement, 224, 24

Le Fèvre O., et al., 2013, Astronomy and Astrophysics, 559, A14

Lemos P., Raveri M., Campos A., et al., 2020, to be submitted to PRD

Liaudat T., Bonnin J., Starck J. L., Schmitz M. A., Guinot A., Kilbinger M., Gwyn S. D. J., 2020, arXiv e-prints, p. arXiv:2011.09835

Lilly S. J., et al., 2009, Astrophysical Journal, Supplement, 184, 218

Limber D. N., 1953, Astrophysical Journal, 117, 134

Linde A. D., 1982, Physics Letters B, 108, 389

Linde A. D., 1983, Physics Letters B, 129, 177

Luppino G. A., Kaiser N., 1997, Astrophysical Journal, 475, 20

Lupton R. H., Gunn J. E., Szalay A. S., 1999, Astronomical Journal, 118, 1406

Ma Z., Hu W., Huterer D., 2006, Astrophysical Journal, 636, 21

MacCrann N., Becker M. R., McCullough J., Amon A., Gruen D., et al., 2020a, To be submitted to MNRAS

MacCrann N., et al., 2020b, arXiv e-prints, p. arXiv:2012.08567

Macaulay E., et al., 2019, Monthly Notices of the RAS, 486, 2184

Masters D. C., Stern D. K., Cohen J. G., Capak P. L., Rhodes J. D., Castander F. J., Paltani S., 2017, Astrophysical Journal, 841, 111

Masters D. C., et al., 2019, Astrophysical Journal, 877, 81

McCracken H. J., et al., 2012, Astronomy and Astrophysics, 544, A156

Ménard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, arXiv e-prints, p. arXiv:1303.4722

Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, The Journal of Chemical Physics, 21, 1087

Miller L., Kitching T. D., Heymans C., Heavens A. F., van Waerbeke L., 2007, Monthly Notices of the RAS, 382, 315

Myles J. T., Alarcon A., et al., 2020, to be submitted to MNRAS

Newman J. A., 2008, Astrophysical Journal, 684, 88

Nørgaard-Nielsen H. U., 2018, Astronomische Nachrichten, 339, 432

Padilla N. D., Strauss M. A., 2008, Monthly Notices of the RAS, 388, 1321

Padilla C., et al., 2019, Astronomical Journal, 157, 246

Padmanabhan N., et al., 2007, Monthly Notices of the RAS, 378, 852

Paraficz D., Hjorth J., 2010, Astrophysical Journal, 712, 1378

Peacock J. A., Smith R. E., 2000, Monthly Notices of the Royal Astronomical Society, 318, 1144–1156

Penzias A. A., Wilson R. W., 1965, Astrophysical Journal, 142, 419

Perlmutter S., et al., 1999, Astrophysical Journal, 517, 565

Persic M., Salucci P., Stel F., 1996, Monthly Notices of the RAS, 281, 27

Phillips M. M., et al., 2019, Publications of the ASP, 131, 014001

Planck Collaboration et al., 2016, Astronomy and Astrophysics, 594, A13

Planck Collaboration et al., 2020, Astronomy and Astrophysics, 641, A6

Porredon A., et al., 2020, arXiv e-prints, p. arXiv:2011.03411

Porredon A., et al., in prep., To be submitted to PRD

Prat J., et al., 2020, To be submitted to MNRAS

Ratra B., Peebles P. J. E., 1988, Physical Review D, 37, 3406

Refregier A., Bacon D., 2003, Monthly Notices of the RAS, 338, 48

Refsdal S., 1964, Monthly Notices of the RAS, 128, 307

Riess A. G., et al., 1998, Astronomical Journal, 116, 1009

Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, Astrophysical Journal, 876, 85

Rodríguez-Monroy M., et al., 2020, To be submitted to MNRAS

Rozo E., et al., 2016, Monthly Notices of the RAS, 461, 1431

Rubin V. C., Ford W. K. J., Thonnard N., 1980, Astrophysical Journal, 238, 471

Saha P., Coles J., Macciò A. V., Williams L. L. R., 2006, Astrophysical Journal, Letters, 650, L17

Samuroff S., Troxel M. A., Bridle S. L., Zuntz J., MacCrann N., Krause E., Eifler T., Kirk D., 2017, Monthly Notices of the RAS, 465, L20

Sánchez C., Prat J., et al., 2020a, To be submitted to MNRAS

Sánchez C., Raveri M., Alarcon A., Bernstein G. M., 2020b, arXiv e-prints, p. arXiv:2004.09542

Scheneider P. Kochanek C. W. J., ed. 2006, Gravitational Lensing: Strong, Weak and Micro (arXiv:astro-ph/0407232)

Schneider P., 2006, Extragalactic Astronomy and Cosmology

Schneider P., Seitz C., 1995, Astronomy and Astrophysics, 294, 411

Schneider P., van Waerbeke L., Kilbinger M., Mellier Y., 2002, Astronomy and Astrophysics, 396, 1

Scodeggio M., et al., 2018, Astronomy and Astrophysics, 609, A84

Secco L. F., Samuroff S., et al., 2020, To be submitted to

Seljak U., 1998, Astrophysical Journal, 506, 64

Sheldon E., 2015, NGMIX: Gaussian mixture models for 2D images (ascl:1508.008)

Sheldon E. S., Huff E. M., 2017, Astrophysical Journal, 841, 24

Sheldon E. S., Becker M. R., MacCrann N., Jarvis M., 2020, Astrophysical Journal, 902, 138

Simon P., King L. J., Schneider P., 2004, Astronomy and Astrophysics, 417, 873

Skilling J., 2004, AIP Conference Proceedings, 735, 395

SkyPy Collaboration et al., 2021, SkyPy, doi:10.5281/zenodo.4475347, https://doi.org/10.5281/zenodo.4475347

Slosar A., et al., 2013, Journal of Cosmology and Astroparticle Physics, 2013, 026

Spergel D. N., et al., 2003, Astrophysical Journal, Supplement, 148, 175

Springel V., et al., 2005, Nature, 435, 629

Square Kilometre Array Cosmology Science Working Group et al., 2020, Publications of the Astron. Soc. of Australia, 37, e007

Suyu S. H., et al., 2013, Astrophysical Journal, 766, 70

Tanaka M., et al., 2018, Publications of the ASJ, 70, S9

Tessore N., Harrison I., 2020, The Open Journal of Astrophysics, 3, 6

Troxel M. A., et al., 2018, Physical Review D, 98, 043528

Van Waerbeke L., et al., 2000, Astronomy and Astrophysics, 358, 30

Vázquez J. A., Padilla L. E., Matos T., 2018, arXiv e-prints, p. arXiv:1810.09934

Walsh D., Carswell R. F., Weymann R. J., 1979, Nature, 279, 381

Way M. J., Foster L. V., Gazis P. R., Srivastava A. N., 2009, Astrophysical Journal, 706, 623

White M., 2001, Monthly Notices of the Royal Astronomical Society, 321, 1–3

Wittman D. M., Tyson J. A., Kirkman D., Dell'Antonio I., Bernstein G., 2000, Nature, 405, 143

Wong K. C., et al., 2020, Monthly Notices of the RAS, 498, 1420

Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., 2020, Astronomy and Astrophysics, 637, A100

Yahya S., Bull P., Santos M. G., Silva M., Maartens R., Okouma P., Bassett B., 2015, Monthly Notices of the RAS, 450, 2251

Zuntz J., Kacprzak T., Voigt L., Hirsch M., Rowe B., Bridle S., 2013, Monthly Notices of the RAS, 434, 1604

Zuntz J., et al., 2015, Astronomy and Computing, 12, 45

Zwicky F., 1933, Helvetica Physica Acta, 6, 110

de Jaeger T., Stahl B. E., Zheng W., Filippenko A. V., Riess A. G., Galbany L., 2020, Monthly Notices of the RAS, 496, 3402

de la Hoz E., Vielva P., Barreiro R. B., Martínez-González E., 2020, Journal of Cosmology and Astroparticle Physics, 2020, 006