UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata
Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata
Ciclo XXXVI

Novel methods to extract cosmological information from galaxy redshift surveys

Settore Scientifico Disciplinare FIS/05

Supervisore: Dott. Benjamin R. GRANETT

Cosupervisore: Prof. Luigi GUZZO

Coordinatore: Prof. Roberta VECCHI

Tesi di Dottorato di:
Marina Silvia CAGLIARI

Anno Accademico 2022-2023

External referees:

Prof. Henk HOEKSTRA
Dr. Ariel SANCHEZ

PhD evaluation committee:

External Members:
Prof. Henk HOEKSTRA
Dr. Annalisa PILLEPICH
Internal Member:
Prof. Marco BERSANELLI

Final examination:

27 March 2024 - Dipartimento di Fisica, Università degli Studi di Milano, Italy

To my family

Ai miei genitori
Susanna e Stefano

Ai miei fratelli
Francesco e Michele

Ai miei nonni
Rosaria, Bruno,
Bruna, e Gabriele

Cover illustration:

Euclid's view of the Perseus cluster of galaxies
Credits: ESA/Euclid/Euclid Consortium/NASA, image processing by J.-C. Cuillandre (CEA Paris-Saclay), G. Anselmi

Back cover illustration:

Sketch of Euclid observing galaxies
Credits: illustration by Marta Barretta

MIUR subjects:

FIS/05

Abstract

After decades of successes, the ΛCDM standard cosmological model is facing the first cracks in its structure. The nature of the two most abundant components of the Universe, namely dark energy and dark matter, still eludes our understanding and we started observing consistent discrepancies between the early and late-time measurements of some cosmological parameters. To clarify if these tensions are indicating a deeper problem in the ΛCDM model and hopefully understand the meaning of its key ingredients, a new generation of cosmological surveys has just started. A key probe of the cosmological model is provided by the large-scale distribution of structures in the Universe. The cosmic web contains information related to late-time parameters, such as the cosmological constant or the equation of state of dark energy, and gives means to determine, among others, the fraction of matter in the Universe, the linear matter power spectrum amplitude, or the neutrino mass.

For this reason, starting from the 80s of last century, the amount of data available for large-scale structure studies has steadily increased. It is now about to make a further leap forward thanks to the fourth-generation galaxy surveys, such as *Euclid*, the dark energy spectroscopic instrument (DESI), or the Vera C. Rubin Observatory legacy survey of space and time (LSST). In comparison to previous surveys, these experiments will observe larger volumes and will measure photometric and spectroscopic information for an unprecedented number of galaxies. Standard analysis methods will become suboptimal in terms of data management, both memory and time-wise, data modelling, and information extraction capabilities.

To achieve such ambitious goals, it is mandatory to develop new methods to study the data and improve their management at all levels of the analysis pipelines. In order to meet the requirements on the precision and accuracy of cosmological parameters, we need, in particular, to efficiently select the samples to be analysed, to measure redshifts with high confidence, and to correctly model summary statistics at all scales. The primary interest of my work is the development of alternative algorithms to improve the extraction of scientific information from large-scale galaxy surveys. The focus is on machine learning-based models, but I also study the potential of more standard methods, such as optimal quadratic estimators.

In the first part of this thesis, I develop and discuss two algorithms that exploit galaxy photometric information to measure redshifts and select samples for clustering analyses. First, I present a novel method that exploits the angular correlation of galaxies to improve photometric redshift measurements. We worked on a graph neural network that classifies angular close pairs of galaxies based on their photometric properties as

true or false physical neighbours. The algorithm is especially useful when the spectroscopic information of one of the galaxies in the pair is known. In this case, the graph neural network helps identify catastrophic errors in the redshift measurements reducing the dispersion of the final photometric sample by a factor of 2 and the fraction of catastrophic errors by a factor of ~ 4. This method is complementary to traditional techniques based on spectral energy distribution fitting and it also helps break the degeneracies in colour-redshift space the standard algorithms are prone to.

Secondly, I explore the efficiency of machine learning classifiers for galaxy photometric selection tasks. The aim of this work is to improve the purity and completeness of the *Euclid* galaxy clustering spectroscopic sample using photometric information. I conduct a performance comparison among six machine learning classifiers and traditional photometric selection methods based on colour and magnitude cuts. The results reveal that machine learning algorithms, especially neural networks and support vector classifiers, can identify more intricate boundaries in the multidimensional colour-magnitude space compared to standard techniques. Demonstrating the efficacy of combining spectroscopic selection with neural network photometric selection, I observe an improvement in the redshift purity of the final sample by approximately 20% and 50% when using *Euclid* photometry alone and *Euclid* in combination with ground-based photometry, respectively.

In the second part of the thesis, I report my work on cosmological parameter measurements with galaxy clustering data. I present two alternatives to traditional approaches. I first illustrate my work with the optimal quadratic estimator of the signal of local primordial non-Gaussianities (PNG), parameterised by $f_{\rm NL}$, from the large-scale structure of the Universe. The analysis makes use of optimal redshift weights that maximise the response of the tracers to the possible presence of non-zero PNG. Analysing the power spectrum monopole of the quasar sample of the latest data release of the extended baryon oscillation spectroscopic survey (eBOSS), I obtain one of the most stringent constraints on local PNG from large-scale structure data up to date. This method not only mitigates the bias in the results, but also yields more precise bounds, with an estimated error on $f_{\rm NL}$ of $\sigma_{f_{\rm NL}} \sim 16$. This corresponds to an improvement of approximately 13% compared to the standard approach. In scenarios where quasars exhibit a lower response to local PNG, the optimal constraint gives $\sigma_{f_{\rm NL}} \sim 21$, representing an improvement of around 30% over standard analyses. This work is a first step in the direction of high-precision $f_{\rm NL}$ measurements from large-scale structure data, which will enable us to better understand the dynamics of inflation.

Finally, I discuss a preliminary study on the application of convolutional neural networks for a field-level analysis of large-scale structure data. This investigation is currently confined to the analysis of dark matter halo distributions. However, it applies a realistic survey geometry to generate training data and utilises observational information, such as halo angular positions and redshifts, to construct the network inputs. A novelty is that the training data for the convolutional neural network are generated using a third-order Lagrangian perturbation theory (3LPT) code, which is faster in producing halo catalogues than an N-body simulation. I assess the neural network performance on both 3LPT and N-body simulations to determine its generalisation ability across simulation types. Preliminary findings indicate that, in both real and redshift space, with a field pixelisation of approximately $\sim 10\,{\rm Mpc}\,h^{-1}$, the convolutional neural network consistently produces comparable results for both 3LPT and N-body simulations. The possibility to train machine learning algorithms for field-level analyses with fast simulations is of major importance. It would greatly reduce the computational costs of these methods making them a competitive alternative to traditional approaches.

Organisational note

The present thesis consists of seven chapters divided into two main parts. The second and third chapters of Parts I and the first chapter of Part II have appeared as refereed publications in scientific journals or have been submitted for publication; the co-authors of the relevant articles are mentioned below. Some variations have been made in the presentation of previously published results, to maintain consistency of style and content structure throughout the manuscript.

Chapter 1. Modern cosmology: introduction to the most important concepts of modern cosmological theory.

Chapter 2. Machine learning: introduction to machine learning and neural networks.

Chapter 3. Galaxy distances and redshifts in cosmology: description of the different methods used to measure distances in cosmology.

Chapter 4. Augmenting photometric redshift estimates using spectroscopic nearest neighbours: development and testing of a graph neural network that classifies pairs of angular neighbour galaxies as true or false redshift neighbours. This work has been completed in collaboration with F. Tosone, L. Guzzo, B. R. Granett, and A. Crespi and has been published as an article in Astronomy & Astrophysics (Tosone et al., 2023), on which the Chapter is based.

Chapter 5. *Euclid*: Testing photometric selection of emission-line galaxy targets: detailed study of applications of photometric machine learning classifiers for the selection of *Euclid* spectroscopic galaxy clustering sample. This work has been completed in collaboration with B. R. Granett, L. Guzzo, M. Bertermin, M. Bolzonella, S. de la Torre, P. Monaco, M. Moresco, W. J. Percival, C. Scarlata, Y. Wang, M. Ezziati, O. Ilber, V. Le Brun et al., the paper will be submitted to Astronomy & Astrophysics on behalf of the Euclid Collaboration and is currently under review by the Euclid Consortium Publication Board.

Chapter 6. Optimal constraints on Primordial non-Gaussianity with the eBOSS DR16 quasars in Fourier space: power-spectrum analysis of the latest eBOSS quasar sample to measure local primordial non-Gaussianities using a cosmological signal optimal quadratic estimator. This work has been completed in collaboration with E. Castorina, M. Bonici, and D. Bianchi and has been accepted for publication as an article in the Journal of Cosmology and Astroparticle Physics (Cagliari et al., 2023), on which the Chapter is based.

Chapter 7. Preliminary applications of machine learning to LSS analysis: first results of a field level machine learning algorithm applied to dark matter halo simulated catalogues.

Appendix A. Photometric selection additional tests: additional plots and discussion related to Chapt. 5.

Appendix B. Fitting $b_\phi\, f_{\rm NL}$: appendices related to Chapt. 6. Results of the analyses that fit the product $b_\phi\, f_{\rm NL}$.

Contents

Introduction

CHAPTER 1

Modern cosmology

Cosmology is the subject that studies the Universe as its whole. It aims to understand the physics, the dynamics, and the evolution of the Universe and its content. Cosmology is a relatively new branch of physics for its development started in the second decade of the last century after the publication of Einstein's General Theory of Relativity (GR, Einstein, 1915).

1.1 A homogeneous Universe

Cosmology pioneers, such as Friedmann and Lemaître, took as an assumption that the Universe is isotropic and homogeneous. This axiom is known as the *cosmological principle*. Nowadays, the idea of the homogeneity and isotropy of the Universe is based on strong observational evidence, which includes the isotropy of the cosmic microwave background (CMB, Planck Collaboration et al., 2020a) radiation and the distribution of galaxies on large scales measured in galaxy surveys (Gonçalves et al., 2021).

Another observational fact that is fundamental in cosmology is the expansion of the Universe. In 1929 Hubble published the evidence of a relation between the distance of the galaxies and their radial velocity (Hubble, 1929). This was just the proof of what was independently derived by Friedmann (Friedmann, 1922) and Lemaître (Lemaître, 1927) starting from the metric of an expanding Universe.

1.1.1 The Friedmann-Lemaître-Robertson-Walker Universe

The cosmological principle and the expansion of the Universe are backed by observational evidence, while the underlying assumption of cosmology is that the Universe dynamics are described by the General Theory of Relativity. The fundamental equations of GR link the curvature of space-time, which is encoded in the metric $g_{\mu\nu}$, and its content, which is described by the energy-momentum tensor $T_{\mu\nu}$. Einstein's equations read as follows

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}\, g_{\mu\nu}\, R = \frac{8\pi\, G}{c^4}\, T_{\mu\nu} - \Lambda\, g_{\mu\nu}\,, \tag{1.1}$$

where $G_{\mu\nu}$ is the *Einstein tensor*, $R_{\mu\nu}$ and R are the *Ricci tensor* and *Ricci scalar*, G is the Newton's gravitational constant, and c is the speed of light. Equation (1.1) also features the cosmological constant Λ, which was originally introduced by Einstein in the equations to obtain a static solution (Einstein, 1917). Nowadays, we know that the cosmological constant can explain the accelerated expansion of the late Universe (Riess et al., 1998).

The metric that describes a homogeneous, isotropic, and expanding Universe is the *Friedman-Lemaître-Robertson-Walker metric* (FLRW),

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \frac{a(t)}{c^2}\left[\mathrm{d}r^2 + \mathcal{R}^2\,\sin^2\left(\frac{r}{\mathcal{R}}\right)\left(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2\right)\right]. \tag{1.2}$$

The time coordinate t that appears in Eq. (1.2) is the *cosmic time*. In the spatial term of the right-hand side of the equation, $a(t)$ is the *scale factor*, which describes the expansion of the Universe and is normalised to 1 at the present epoch t_0, $\mathcal{R}$ is the spatial curvature of the Universe at the present epoch, and r is the *comoving radial distance*, which is the proper distance of a galaxy at the present epoch. Then, the proper distance at epoch t is

$$r(t) = a(t)\,r\,. \tag{1.3}$$

With a change of coordinate, $r_1 = \mathcal{R}\,\sin(r/\mathcal{R})$, Eq. (1.2) becomes

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \frac{a(t)}{c^2}\left[\frac{\mathrm{d}r_1^2}{1-k\,r_1^2} + r_1^2\left(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2\right)\right], \tag{1.4}$$

where k is a real number that encodes the spatial curvature of the Universe. If $k = 0$ the Universe is flat, if it is positive the Universe is close with a spherical geometry, and if it is negative the Universe is open and the geometry is hyperbolic.

As mentioned above, *Hubble's law*

$$v = H_0\,r\,, \tag{1.5}$$

which is an observational relation, can be also derived from the expansion of the Universe as it is described in Eq. (1.3). If we derive this equation with respect to time and substitute it again we obtain

$$\dot{r}(t) = \dot{a}(t)\,r = \frac{\dot{a}(t)}{a(t)}r(t) \longrightarrow v(t) = H(t)\,r(t), \tag{1.6}$$

where $H(t) = \frac{\dot{a}(t)}{a(t)}$, usually called *Hubble parameter*, is a measure of the expansion rate of the Universe at a given epoch t. Therefore, the *Hubble constant* represents the expansion rate at the present epoch, $H_0 = H(t_0)$, and $H(t)$ defines a Hubble parameter for each epoch. The Hubble constant has the dimension of the inverse of a time, but it is usually measured in $\mathrm{km\,s^{-1}\,Mpc^{-1}}$. It is also useful to define the dimensionless Hubble constant

$$h = \frac{H_0}{100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}}\,. \tag{1.7}$$

In cosmology, we usually model the different Universe components as non-interacting perfect fluids that are at rest, in thermodynamical equilibrium, and have energy density ρ_i and pressure P_i. Given their energy-momentum tensor, $T_{\mu\nu}$, we can derive the Einstein's equations for the FLRW metric (see Eq. 1.4) and obtain

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi\,G}{3}\sum_i \rho_i - \frac{c^2\,k}{a^2} + \frac{\Lambda}{3}\,, \tag{1.8}$$

$$\frac{\ddot{a}}{a} = -4\pi\,G\sum_i\left(\rho_i + \frac{3\,P_i}{c^2}\right) + \frac{\Lambda}{3}\,. \tag{1.9}$$

These two equations are also known as the *first* and *second Friedmann's equations* and describe the dynamics of the Universe expansion. An additional equation is the local conservation of energy,

$$\dot{\rho}_i + 3\,\frac{\dot{a}}{a}\,\left(\rho_i + \frac{P_i}{c^2}\right) = 0\,, \tag{1.10}$$

which is derived from the local conservation of the energy-momentum tensor.

1.1.2 The Universe bricks

To close the equation system of Eqs. (1.8), (1.9), and (1.10) we need a link between the pressure and the energy density, which is given by the equation of state,

$$\frac{P_i}{c^2} = w_i\,\rho_i\,. \tag{1.11}$$

Substituting Eq. (1.11) in Eq. (1.10) we obtain an expression that can be integrated

$$\dot{\rho}_i = -3\,\frac{\dot{a}}{a}\,\left(1 + w_i\right)\,\rho_i\,, \tag{1.12}$$

$$\rho_i \propto \exp\left[-3\int_a \frac{\mathrm{d}a'}{a'}\,\left[1 + w_i\,(a')\right]\right]\,. \tag{1.13}$$

If w_i is time-independent, Eq. (1.13) becomes

$$\rho_i = \rho_{i,0}\,a^{-3\,(1+w_i)}\,, \tag{1.14}$$

where $\rho_{i,0}$ is the value of the energy density at the present epoch. Then, to describe the Universe expansion we need to know its content and the equation of state of its components.

First, there is the *cold matter*, which is how we refer to all the non-relativistic components of the Universe. Cold matter includes cold dark matter (CDM) and baryons and is pressureless. Therefore, $w_m = 0$ and

$$\rho_m = \rho_{m,0}\,a^{-3}\,. \tag{1.15}$$

Equation (1.15) translates in the fact that the matter-energy density scales as the particle density if their number is conserved in a comoving volume. As second comes *radiation*, which comprises all the relativistic components of the Universe such as photons and relativistic particles. The relativistic equation of state reads

$$w_r = \frac{1}{3}\,, \tag{1.16}$$

and corresponds to

$$\rho_r = \rho_{r,0}\,a^{-4}\,. \tag{1.17}$$

Radiation gets diluted faster than matter because of the *cosmological redshift*. This effect corresponds to a stretch of the particle wavelengths by a factor a^{-1}, which, combined with the volume dilution ($\propto a^{-3}$), leads to the dependence in Eq. (1.17). Equations (1.15) and (1.17) describe the evolution of the matter and radiation density as functions of the

Universe expansion. It is possible to identify a time when matter density equals radiation density. This epoch is called *matter-radiation equality* and reads

$$a_{\rm eq} = \frac{\rho_{r,0}}{\rho_{m,0}} \,. \tag{1.18}$$

Before $a_{\rm eq}$ the Universe dynamic is dominated by radiation after it is dominated by matter. We refer to these two eras as radiation or matter domination. As we will see in Sect. 1.2.1, $a_{\rm eq}$ marks a transition in the rate of the large-scale structure evolution in the Universe.

Equation (1.8) suggests that also the *curvature* and the *cosmological constant* are part of the cosmic inventory. The curvature energy density is defined as follows

$$\rho_k = -\frac{3\,k}{8\pi\,G}\,a^{-2}\,, \tag{1.19}$$

which means that

$$w_k = -\frac{1}{3}\,. \tag{1.20}$$

Analogously, we can define the cosmological constant energy density

$$\rho_\Lambda = \frac{\Lambda}{8\pi\,G}\,, \tag{1.21}$$

and its equation of state

$$w_\Lambda = -1\,. \tag{1.22}$$

Equation (1.22) implies that the pressure of the cosmological constant is negative. Originally, the fact that the cosmological constant energy density does not vary with the Universe expansion (Eq. 1.21) led to the idea that Λ was related to the vacuum energy. However, that is not the case as the measured value of the cosmological constant is between 50 and 120 order of magnitude lower than expected if it was to be related to the energy of vacuum (e.g., Adler et al., 1995). Nowadays it is believed that the cosmological constant is related to the so called *dark energy* (DE), which is one of the biggest puzzles of modern physics.

The two Friedmann's equations (Eqs. 1.8 and 1.9) can be rewritten in terms of Hubble's parameter and the energy densities defined above becoming

$$H^2(t) = \frac{8\pi\,G}{3}\sum_i \rho_i\,, \tag{1.23}$$

and

$$\dot{H}(t) = -4\pi\,G\sum_i (1+w_i)\;\rho_i\,. \tag{1.24}$$

The equations can be further simplified by defining the *density parameters*. First we introduce the *critical density*,

$$\rho_{\rm cr} \equiv \frac{3\,H_0^2}{8\pi\,G}\,, \tag{1.25}$$

which corresponds to the present epoch energy density of a flat Universe. The density parameter of the Universe i-th species is its energy density normalised by the critical density,

$$\Omega_i = \frac{\rho_i}{\rho_{\rm cr}} = \frac{\rho_{i,0}}{\rho_{\rm cr}}\,a^{-3\,(1+w_i)} = \Omega_{i,0}\,a^{-3\,(1+w_i)}\,. \tag{1.26}$$

In Eq. (1.26) $\Omega_{i,0}$ is the density parameter evaluated at the present epoch. Combining Eqs. (1.23) and (1.26) we obtain a relation for the Hubble parameter as a function of the scale factor,

$$H^2(a) = H_0^2 \sum_i \Omega_i = H_0^2 \left(\Omega_{m,0}\, a^{-3} + \Omega_{r,0}\, a^{-4} + \Omega_{k,0}\, a^{-2} + \Omega_\Lambda\right) . \tag{1.27}$$

If we evaluate Eq. (1.27) at the present epoch, t_0, it becomes

$$\sum_i \Omega_{i,0} = 1 . \tag{1.28}$$

1.1.3 Distances in an expanding Universe

As the majority of information on astrophysical objects comes from their electromagnetic radiation we need to understand how the expansion of the Universe affects travelling light waves. In Sect. 1.1.2 I already introduced the concept of cosmological redshift, here I am going to describe the origin of this effect. In general, we define the *redshift*, z, as the relative difference between the emitted wavelength, λ_1, and the observed one, λ_0:

$$z \equiv \frac{\lambda_0 - \lambda_1}{\lambda_1} . \tag{1.29}$$

If we consider a wave packet, which travels along null cones, $\mathrm{d}s^2 = 0$, that is moving radially ($\mathrm{d}\vartheta = 0$ and $\mathrm{d}\varphi = 0$), Eq. (1.2) becomes:

$$\frac{c\,\mathrm{d}t}{a(t)} = -\mathrm{d}r . \tag{1.30}$$

Let us say that this wave packet was emitted between time t_1 and $t_1+\Delta t_1$ with frequency ν_1, and received by an observer at present time in an interval between time t_0 and $t_0+\Delta t_0$ and frequency ν_0. The leading edge of the wave packet travels the comoving distance, r, between time t_1 and t_0,

$$\int_{t_0}^{t_1} \frac{c\,\mathrm{d}t}{a(t)} = -\int_r^0 \mathrm{d}r , \tag{1.31}$$

while its end must travel the same comoving distance from time $t_1 + \Delta t_1$ and $t_0 + \Delta t_0$

$$\int_{t_0+\Delta t_0}^{t_1+\Delta t_1} \frac{c\,\mathrm{d}t}{a(t)} = -\int_r^0 \mathrm{d}r . \tag{1.32}$$

Combining Eqs. (1.31) and (1.32) we obtain

$$\int_{t_0}^{t_1} \frac{c\,\mathrm{d}t}{a(t)} + \frac{c\,\Delta t_0}{a(t_0)} + \frac{c\,\Delta t_1}{a(t_1)} = \int_{t_0}^{t_1} \frac{c\,\mathrm{d}t}{a(t)} . \tag{1.33}$$

Knowing that $a(t_0) = 1$, Eq. (1.33) becomes

$$\Delta t_0 = \frac{\Delta t_1}{a(t_1)} . \tag{1.34}$$

This result, known as *time dilation*, can be reduced to a relation between the redshift and the Universe expansion. If the emission interval is $\Delta t_1 = \nu_1^{-1}$ and the observed one is $\Delta t_0 = \nu_0^{-1}$, Eq. (1.34) can be rewritten as

$$\nu_0 = a(t_1)\nu_1 . \tag{1.35}$$

Substituting this expression in the redshift definition, Eq. (1.29) becomes

$$z = \frac{\lambda_0}{\lambda_1} - 1 = \frac{\nu_1}{\nu_0} - 1 = \frac{1}{a(t_1)} - 1 \longrightarrow a(t_1) = \frac{1}{z+1}\,. \tag{1.36}$$

The redshifting effect is only related to the expansion of the Universe and not to the relative velocity between the source and the observer, thus it is known as cosmological redshift. Cosmological redshift is a measure of the scale factor of the Universe at the epoch in which the radiation was emitted.

Now, combining Eq. (1.30) and the definition of the Hubble parameter we find a relation between the radial comoving distance of a source at scale factor a and the scale factor itself

$$r(a) = c\int_{t(a)}^{t_0} \frac{\mathrm{d}t'}{a(t')} = c\int_a^1 \frac{\mathrm{d}a'}{a'^2\,H(a')}\,. \tag{1.37}$$

From this expression, we understand that the comoving distance is the maximum distance light can travel between the time of emission $t(a)$ and the time of observation t_0. We call the *comoving horizon* the distance light could have travelled from $t = 0$. Since no information can travel faster than light, events that are further than the comoving horizon are causally disconnected. Alternatively, combining Eqs. (1.36) and (1.37) we can write a relation between the comoving distance and the redshift of the source,

$$r(z) = c\int_0^z \frac{\mathrm{d}z'}{H(z')}\,. \tag{1.38}$$

At the end of the previous section, I presented a relation between the Hubble parameter and the scale factor (Eq. 1.27). With the change of coordinate of Eq. (1.36), we can write a relation between the Hubble parameter as a function of the cosmological redshift. Therefore, given a cosmological model, we can solve the integral of Eq. (1.38) and measure the radial comoving distance of an object starting from its redshift. This makes the redshift one of the primary pieces of information we want to measure in any cosmological observation.

We can define two additional distances. From the FLRW metric, Eq. (1.2), it is straightforward to obtain the angular size of a source with proper length Δl perpendicular to the radial coordinate and at redshift z. The relevant spatial component in the metric is the angular term in $\mathrm{d}\vartheta$,

$$\Delta l = a(t)\mathcal{R}\sin\left(\frac{r}{\mathcal{R}}\right)\Delta\vartheta = a(t)r_1\,\Delta\vartheta = \frac{r_1\Delta\vartheta}{1+z}\,. \tag{1.39}$$

I already introduced the distance measure r_1 in Eq. (1.4). Now, we can give a physical interpretation of this definition. Let us consider an object that is expanding with the Universe. Its proper dimension at epoch t is $\Delta l(t) = a(t)\,\Delta l_0 = \Delta l_0\,(1+z)^{-1}$ and it subtends an angle

$$\Delta\vartheta = \frac{\Delta l(t)\,(1+z)}{r_1} = \frac{\Delta l_0}{r_1}\,. \tag{1.40}$$

The distance measure r_1 is the distance of a source with angular dimension $\Delta\vartheta$ that is expanding with the Universe and is called *comoving angular diameter distance*. We can also define the *angular diameter distance* $r_\mathrm{A} = r_1\,(1+z)^{-1}$ and reduce Eq. (1.40) to the standard Euclidean relation between distance and angle at any epoch t,

$$\Delta\vartheta = \frac{\Delta l}{r_\mathrm{A}}\,. \tag{1.41}$$

Finally, we can tackle the problem of the relation between the observed flux density $S(\nu_0)$, which is the energy per unit of time, area, and bandwidth, and the source luminosity $L(\nu_1)$ that is the total energy emitted over 4π steradians per unit of time and bandwidth. The luminosity of a source that emits $N(\nu_1)$ photons with energy $h_{\rm P}\,\nu_1$, where $h_{\rm P}$ is the Planck constant, in a bandwidth $\Delta\nu_1$, and in a proper time interval Δt_1 is

$$L(\nu_1) = \frac{N(\nu_1)\,h_{\rm P}\,\nu_1}{\Delta t_1\,\Delta\nu_1}\,. \tag{1.42}$$

The photons travel on the surface of a sphere centred on the source. The number of photons an observer will see depends on their telescope angular dimension with respect to the source, while the observed frequency and the observed time interval are related to the emitted ones by Eqs. (1.34) and (1.35). Let us say that the telescope, which is in the present epoch t_0, has a diameter Δl and it subtends an angular diameter $\Delta\vartheta$ for the source in epoch t_1, then

$$\Delta l = r_1 \Delta\vartheta\,. \tag{1.43}$$

The area of this telescope is $\pi\,\Delta l^2/4$, the solid angle it subtends is $\Delta\Omega = \pi\Delta\vartheta^2/4 = \pi\,\Delta l^2\,(4\,r_1^2)^{-1}$, and the number of photons it observes is $N(\nu_1)\,\Delta\Omega/(4\pi)$. These photons are observed in a time interval Δt_0 at frequency ν_0, thus the flux density of the source is

$$S(\nu_0) = \frac{4\,N(\nu_1)\,h_{\rm P}\,\nu_0\,\Delta\Omega}{4\pi\,\Delta t_0\,\Delta\nu_0\,\pi\,\Delta l^2}\,. \tag{1.44}$$

Substituting in this expression the luminosity at the source, Eq. (1.42), the relation between times and frequencies at different epochs, Eqs. (1.34) and (1.35), and rewriting $\Delta\Omega$ as above Eq. (1.44) reads

$$S(\nu_0) = \frac{L(\nu_1)}{4\pi\,r_1^2(1+z)}\,. \tag{1.45}$$

If we consider the case of bolometric luminosities and flux densities we can define a new distance measure called *luminosity distance*, $r_{\rm L}$, as follows

$$S_{\rm bol} = \frac{L_{\rm bol}}{4\pi\,r_1^2\,(1+z)^2} = \frac{L_{\rm bol}}{4\pi\,r_{\rm L}^2}\,, \tag{1.46}$$

where $r_{\rm L} = r_1\,(1+z)$.

1.2 The large-scale structure of the Universe

In the previous section, I discussed the dynamic of a homogeneous and isotropic Universe. However, nowadays we observe a Universe that has very strong anisotropies on small scales and contains stars, galaxies, and clusters of galaxies. We observe the seeds of these same anisotropies in the CMB radiation, which have a very small amplitude ($\sim 10^{-5}$ K). The top panel of Fig. 1.1 shows the map of the CMB anisotropies, the bottom panel presents their temperature power spectrum and the outstanding match of the ΛCDM model with the observed data. The CMB anisotropies evolved into the Universe we observe now. Figure 1.2 shows the *cosmic web* in the near Universe.

The theory of structure evolution describes how the present large-scale structure (LSS) of the Universe formed starting from the CMB anisotropies. In principle, we need to solve Einstein's equations (Eq. 1.1) in the case of a perturbed metric, where the perturbation fields are $\Psi(\boldsymbol{x},t)$ for the time component of the metric and $\Phi(\boldsymbol{x},t)$ for the space

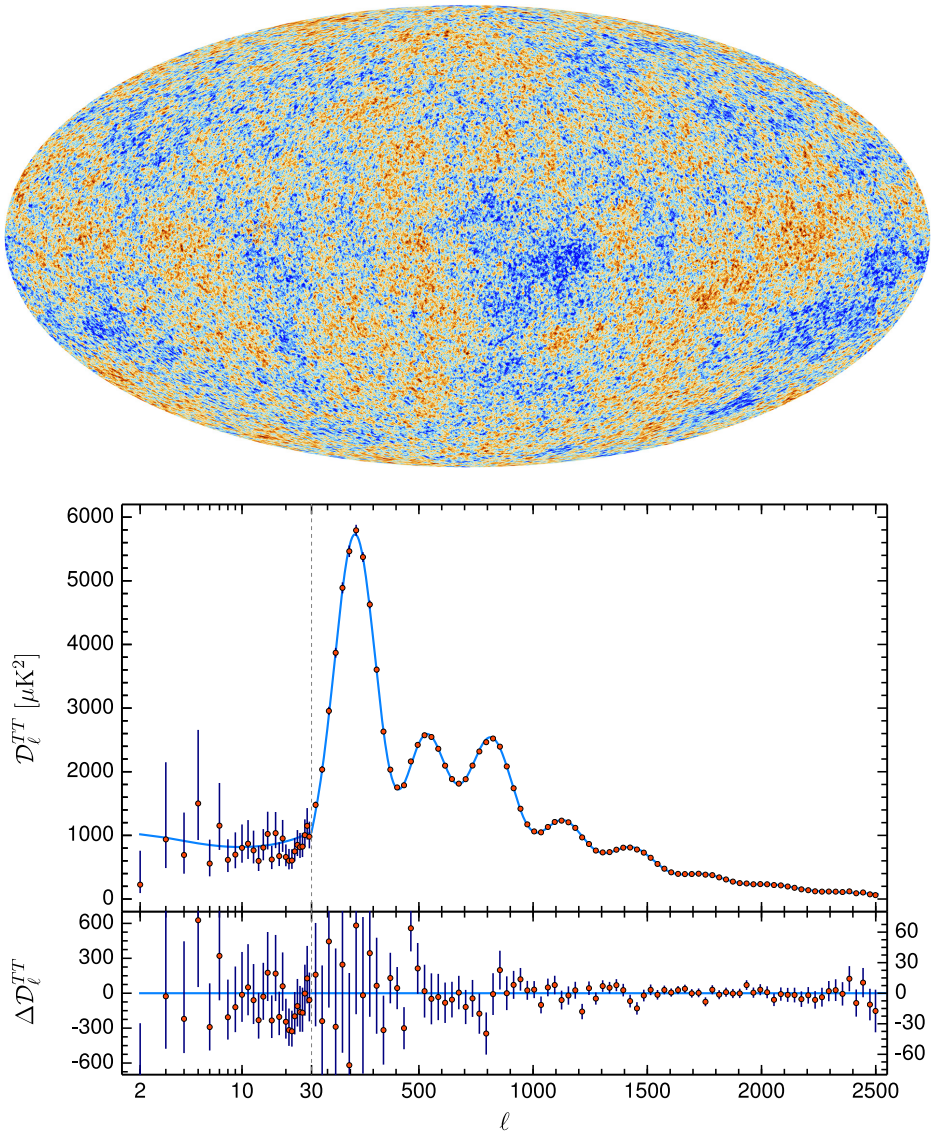

Figure 1.1: The cosmic microwave background radiation as it was observed by the ESA *Planck* satellite (top) and the power spectrum of its temperature anisotropies (bottom). In the bottom panel, the points correspond to Planck observations, while the blue solid line is the ΛCDM prediction. Figures credits to ESA and Planck Collaboration et al. (2020a).

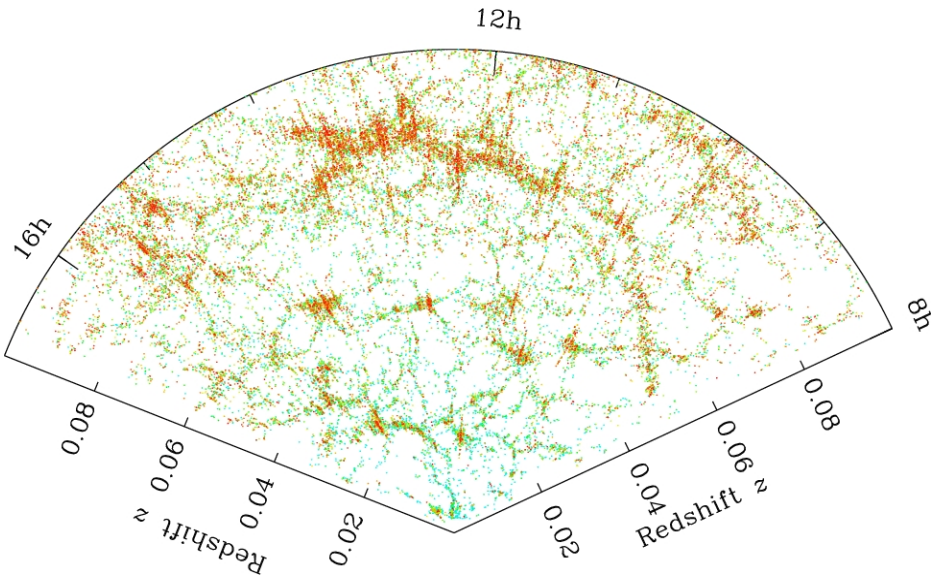

Figure 1.2: A slice through the distribution of the main galaxy sample in the northern part of the Sloan Digital Sky Survey (SDSS; York et al., 2000). Each dot depicts the position of a galaxy, with colour chosen to represent the actual colour of the galaxy. Figure credits to Michael Blanton and the SDSS Collaboration.

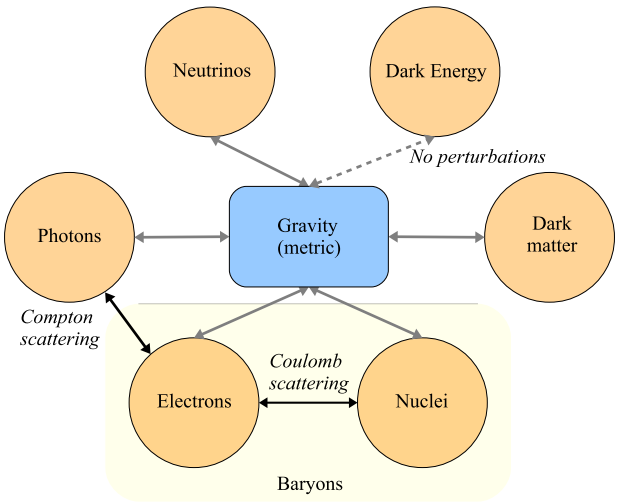

Figure 1.3: A schematic description of the interaction between the different components of the Universe. Figure credits to Dodelson & Schmidt (2020).

part, and an inhomogeneous energy-momentum tensor. Then, combining this with the *Boltzman's equation* we can derive the evolution of the perturbation for each component of the cosmic inventory. Boltzmann's equation gives the evolution of the *distribution function* f of a species in phase-space given its particle-particle interactions encoded in the *collision term*. Figure 1.3 schematically shows the possible interactions between the universe components. All of them interact through gravity. Additionally, electrons interact with nuclei and protons via *Coulomb scattering* and with photons via *Compton scattering*. It is important to note that in the case that dark energy is a cosmological constant it does not have any perturbations and contributes only to the background homogeneous part of the metric.

Ultimately, in LSS studies we observe the distribution of baryonic matter, which, in first approximation, is only determined by the dark matter distribution. For this reason, we are mainly interested in the evolution of the CDM perturbations and their distribution. We use a statistical description of random fields to analyse the matter distribution (Peebles, 1980) and we start defining the matter *density field*,

$$\delta(\boldsymbol{r}) = \frac{\rho(\boldsymbol{r}) - \bar{\rho}}{\bar{\rho}}\,, \qquad \text{with } \langle\rho(\boldsymbol{x})\rangle = \bar{\rho}\,, \tag{1.47}$$

where the symbol $\langle\cdot\rangle$ corresponds to the ensemble average. This average should be computed over different realisations of the Universe, but in practice we we use a large enough volume assuming ergodicity. The density field, as defined in Eq. (1.47), has $\langle\delta(\boldsymbol{r})\rangle = 0$. To describe this random field we use its *correlation functions*

$$\xi^{(n)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \langle\delta_{\boldsymbol{x}_1} \ldots \delta_{\boldsymbol{x}_n}\rangle\,, \tag{1.48}$$

where the superscript (n) refers to the n-th order correlation function. In a Universe that is isotropic and homogeneous, the correlation functions only depend on the relative distance between points. Therefore $\xi^{(n)}$ only depends on $n-1$ spatial coordinates. Then, the two-point correlation function of the density field reads as follows

$$\xi(r) = \langle\delta(\boldsymbol{x})\,\delta(\boldsymbol{x}+\boldsymbol{r})\rangle\,, \tag{1.49}$$

and depends only on the modulus of the distance of two points, r. The two-point correlation function represents the excess probability of finding two points at distance r and, if the random field is Gaussian, is the only non-null correlation function.

It can also be useful to work in Fourier space. The Fourier transform of the overdensity is

$$\delta(\boldsymbol{k}) = \int \mathrm{d}^3x\,\delta(\boldsymbol{x})\,e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\,, \tag{1.50}$$

while its inverse reads

$$\delta(\boldsymbol{x}) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}\,\delta(\boldsymbol{k})\,e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\,, \tag{1.51}$$

where $\boldsymbol{k}\cdot\boldsymbol{x}$ denotes the scalar product between the two vectors. Analogously to Eq. (1.48) we can define the correlation between the Fourier transform of the density field. The two-point function in Fourier space is the power spectrum

$$\langle\delta(\boldsymbol{k})\delta(\boldsymbol{k'})\rangle \equiv (2\pi)^3\,\delta^{\mathrm{K}}_{\boldsymbol{k}\boldsymbol{k'}}\,P(\boldsymbol{k})\,, \tag{1.52}$$

where δ^{K} is the Kronecker delta. The two-point correlation function as defined in Eq. (1.49) and the power spectrum form a Hankel pair

$$P(k) = 4\pi\int_0^\infty \mathrm{d}r\,\xi(r)\,r^2\,\frac{\sin kr}{kr}\,. \tag{1.53}$$

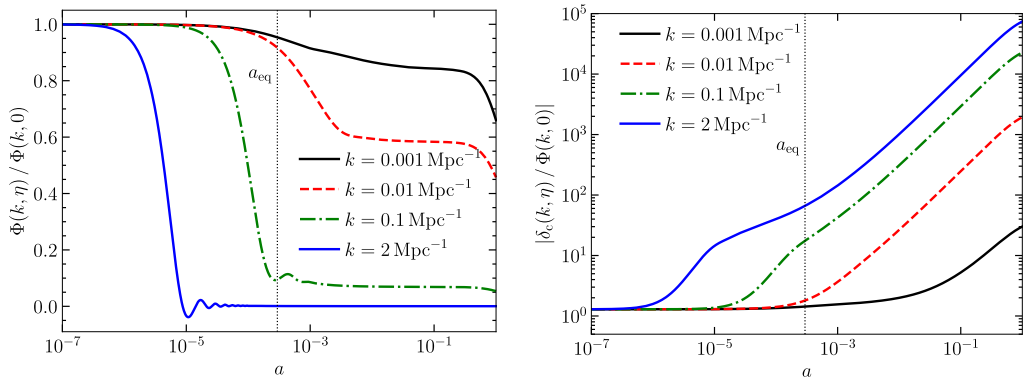

Figure 1.4: Evolution of the gravitational potential Φ and the dark matter density perturbation for modes of different wavenumber in the fiducial ΛCDM cosmology. The curves are normalised to the value of the potential at early times. *Left*: evolution of the gravitational potential perturbations. *Right*: evolution of the density perturbations. The amplitude of each mode starts to grow upon horizon entry. Well after $a_{\rm eq}$, all sub-horizon modes evolve identically, and scale as the growth factor $D_+(a)$, see Eq. (1.55). During matter domination, before Λ becomes relevant, $D_+(a) = a$. At the very latest times, we can see a slight suppression from this linear trend due to the onset of accelerated expansion. Figure credits to Dodelson & Schmidt (2020).

1.2.1 The large-scale structure evolution

The density contrast and its momenta either in configuration or Fourier space gives us a means to describe the matter field. Now, we would like to understand how this field, which is embedded in an expanding background, evolved with time.

When the perturbations are small ($\delta(\boldsymbol{x}) \ll 1$) we can use linear theory to describe the evolution and, in Fourier space, different k scales evolve independently. Figure 1.4 left panel shows the evolution of the gravitational perturbation at different scales as a function of the scale factor, while the right panel shows the evolution of the density contrast. From these plots, we can see that there are different evolution regimes for the perturbation. First, all the modes are outside of the horizon and the potential is constant. Later on, the modes start entering the horizon, from small to large scales. The scales that cross the horizon during the radiation domination era have a very sharp decay in comparison to the modes that enter the horizon after the epoch of equality. In late times, during the matter domination era, all the modes have entered the horizon and evolve identically remaining constant.

Given the primordial potential $\Phi_{\rm P}(\boldsymbol{k})$ we divide its evolution in a scale-dependent part and in a time-dependent part as follows

$$\Phi(\boldsymbol{k}, a) = \frac{3}{5}\,\Phi_{\rm P}(\boldsymbol{k})\,T(k)\,\frac{D_+(a)}{a} \qquad (a > a_{\rm late})\,, \tag{1.54}$$

where $a_{\rm late}$ is an epoch in the late matter domination era, $T(k)$ is the *transfer function*, and $D_+(a)$ is the *growth factor*. The transfer function encodes the evolution through the epoch of horizon crossing and the transition between radiation and matter domination eras and conventionally is normalised to be 1 for large scales. The growth factor describes the scale-independent evolution at late times. When the potential is constant in the matter-dominated era $D_+(a) \sim a$ and describes the growth of the matter density perturbation

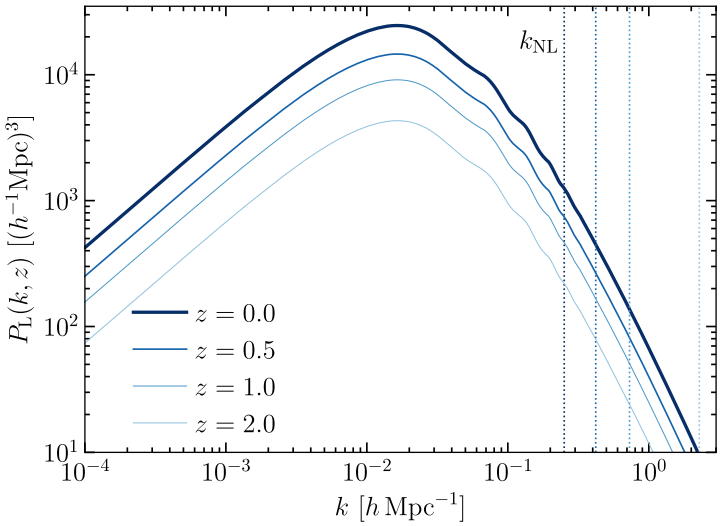

Figure 1.5: The linear matter power spectrum in the fiducial ΛCDM cosmology at different redshifts. Scales to the left of the vertical lines, which indicate $k_{\rm NL}(z)$ for each of the redshifts shown, are still evolving approximately linearly at each redshift. Figure credits to Dodelson & Schmidt (2020).

in time as depicted in the right panel of Fig. 1.4. We obtain this same result by relating the matter density and the potential through the Poisson equation in the large k and no radiation limit. Then, we can write the late time density evolution with respect to the primordial potential

$$\delta(\boldsymbol{k}, a) = \frac{2}{5}\,\frac{k^2\,c^2}{\Omega_{\rm m,0} H_0^2}\,\Phi_{\rm P}(\boldsymbol{k})\,T(k)\,D_+(a) \qquad (a > a_{\rm late}, k \gg a\,H)\,. \tag{1.55}$$

Equation (1.55) holds for any adiabatic perturbation. Finally, we can write the power spectrum of matter at late times in the case of linear evolution

$$P_{\rm L}(k,a) = \frac{8\pi^2}{25}\,\frac{\mathcal{A}_s}{\Omega_{\rm m,0}^2}\,D_+^2(a)\,T^2(k)\,\frac{k^{n_s}}{H_0^4\,k_{\rm p}^{n_s-1}}\,, \tag{1.56}$$

where the power spectrum of the primordial perturbation is

$$P_{\Phi_{\rm P}}(k) = 2\pi^2\,k^{-3}\,\mathcal{A}_s\,(k/k_{\rm p})^{n_s-1} \tag{1.57}$$

as a consequence of inflation (Baumann, 2011). In Eq. (1.56), n_s is the scalar spectral index, $k_{\rm p}$ is the pivot scale, and $\mathcal{A}_s$ is the scalar amplitude of the fluctuation. In the case of a galaxy survey σ_8 usually substitutes the amplitude $\mathcal{A}_s$. This parameter corresponds to the amplitude of the linear matter power spectrum at the present epoch and at the scale of $8\,h^{-1}$Mpc. Figure 1.5 shows the matter linear power spectrum as a function of the scale k and the redshift z. At large scales, where $T(k) = 1$, the power spectrum is proportional to k^{n_s}, while at small scales we observe a turnover. In Fig. 1.4 left panel we see that when a mode enters the horizon before the matter/radiation equality epoch its potential decays and its density (right panel) will start increasing again only after

matter/radiation equality. All the scales that enter the horizon before $a_{\rm eq}$ undergo a suppression leading to a decreasing power spectrum up to the scale $k_{\rm eq}$ that entered the horizon during matter/radiation equality. The value of this scale depends on $\Omega_{\rm m,0}$.

In the case of a linear perturbation, it is possible to write the equation that describes the evolution of the matter density contrast, δ,

$$\frac{{\rm d}^2\delta}{{\rm d}a^2} + \frac{{\rm d}\left(\ln a^3\, H(a)\right)}{{\rm d}a}\frac{{\rm d}\delta}{{\rm d}a} = \frac{3}{2}\,\frac{\Omega_{\rm m,0}\, H_0^2}{a^5\, H^2(a)}\,\delta\,. \tag{1.58}$$

In general, this differential equation has to be solved numerically. In the late Universe, where matter and the cosmological constant are dominant, we can write an integral solution for Eq. (1.58), which reads

$$D_+(a) = \frac{5\,\Omega_{\rm m,0}}{2}\,\frac{H(a)}{H_0}\int_0^a \frac{{\rm d}a'}{(a'\,H(a')/H_0)^3}\,. \tag{1.59}$$

Equation (1.59) is not a solution to Eq. (1.58) if dark energy is not a cosmological constant. In this case, Eq. (1.58) needs to be solved numerically. However, for the logarithmic derivative of the growth factor, which is the *growth rate* f, an empirical fit exists and takes the following form in GR

$$f(a) \equiv \frac{{\rm d}\ln D_+(a)}{{\rm d}\ln a} \simeq \left[\Omega_{\rm m}(a)\right]^{0.55}\,. \tag{1.60}$$

Figure 1.5 shows some additional features. First, at each redshift, the dashed vertical lines mark the nonlinear scale, $k_{\rm NL}$. For $k < k_{\rm NL}$ the linear approximation solution discussed above breaks as $\delta(\boldsymbol{x}) \sim 1$. A first approximation to analytically describe the nonlinear regime is the spherical collapse. Second, starting from $k \sim 0.1\,h{\rm Mpc}^{-1}$ there is an oscillation in the power spectrum. At early times, before the emission of the CMB radiation baryons and photons were coupled by the Compton interaction. Acoustic plasma waves travelled through the baryon-photon fluid and left a footprint in the baryon distribution, which subsequently affect the matter distribution. These oscillations we observe in the matter power spectrum are known as *baryon acoustic oscillations* (BAO) and have been detected in the clustering of galaxies (Eisenstein et al., 2005).

1.3 The concordance model of cosmology: ΛCDM

Starting from the theory that describes the homogeneous Universe and the formation of structure together with observation, the cosmological community has developed a concordance model. Nowadays the Universe is considered to be Euclidean, dominated by non-baryonic cold dark matter, and a cosmological constant (Planck Collaboration et al., 2020a). This standard cosmological model is usually referred to as flat ΛCDM and it only requires six parameters to describe the Universe with its content and its evolution. These primary six parameters are the baryon density parameter, $\Omega_{\rm b}h^2$, the cold dark matter density parameter, $\Omega_{\rm c}h^2$, the Hubble parameter, h, the scalar spectral index, n_s, the scalar power spectrum primordial amplitude, $\mathcal{A}_s$, and the re-ionisation optical depth, τ. We can compute all the other parameters introduced in the previous sections starting from the six primary parameters. An additional ingredient to the standard cosmological model is inflation, which is the most accredited mechanism to generate the initial conditions of the Universe.

We can also expand the ΛCDM model with some additional parameters. We can remove the assumption of the Universe flatness and add the curvature density as a free parameter of the model $\Omega_k h^2$ as well as the one on the total neutrino mass, m_ν. Additionally, we can drop the idea of a cosmological constant in favour of a dark energy component with a time-dependent equation of state (Linder, 2003)

$$w_{\rm DE}(z) = w_0 + w_a \frac{z}{1+z} \,. \tag{1.61}$$

Such a model is referred to as dynamical dark energy. The measurement of a possible dependence with time of the dark energy equation of state is one of the key objectives of all the next-generation cosmological surveys (e.g., LSST Science Collaboration et al., 2009; Laureijs et al., 2011; DESI Collaboration et al., 2016).

The standard cosmological model is extremely successful in its description of the Universe and an alternative model that is able to survive all the tests the ΛCDM model has overcome is yet to be developed. However, in the last years, the cosmological community is starting to see cracks in it. First, the nature of the two main components of the model, dark energy and dark matter, is still unknown and dark matter is still eluding a direct detection or description on the particle physics side. Moreover, there are tensions in the measurement of some cosmological parameters between early and late time observation. These tensions may be related to our ignorance of the physical processes that produce the observed data or related to new physics beyond the standard model. The most concerning discrepancies in the model are related to the measurements of H_0 (e.g., Verde et al., 2019) and $S_8 \equiv \sigma_8 \sqrt{\Omega_{\rm m}/0.3}$ (e.g., Abbott et al., 2022). The H_0 tension is particularly disconcerting as the measurements of this parameter from supernovae of type Ia (Riess et al., 2021) and the CMB (Planck Collaboration et al., 2020a) show a 4.2σ discrepancy.

1.4 Galaxy redshift surveys

In the previous section, I discussed the evolution of the overdensity of dark matter. However, what we observe is the distribution of the galaxies, which is a mapping of the underlying dark matter distribution. In first approximation, we expect the two density fields to be linearly related as follows

$$\delta_g(\boldsymbol{x}) \sim b_1 \, \delta_m(\boldsymbol{x}) \,, \tag{1.62}$$

where $\delta_g(\boldsymbol{x})$ and $\delta_m(\boldsymbol{x})$ are the density field respectively of galaxies and matter. In Eq. (1.62), b_1 is the *linear bias* and represents the response of the galaxy (or any other tracer) density field to the matter density field. The bias describes the fact that we expect galaxies to form in dark matter overdensities.

Given Eq. (1.62) and the definition of the correlation function and the power spectrum (Eqs. 1.49 and 1.52), the relation between the galaxy and matter correlation function or power spectrum reads

$$\xi_g(r) \sim b_1^2 \, \xi_m(r) \,, \tag{1.63}$$

$$P_g(k) \sim b_1^2 \, P_m(k) \,. \tag{1.64}$$

From the theory of structure formation, we know how the linear matter power spectrum, $P_m(k)$, evolves.

An additional complication in the analysis of the galaxy density distribution is that galaxies can have a peculiar velocity, which is a motion with respect to the background

evolution of the matter perturbation. Therefore, the redshift distance, s, of an object, which is

$$s \equiv c\, z \tag{1.65}$$

in velocity units, differs from the true distance $v = H_0\, r$ expressed in velocity units and defined by Hubble's law (Eq. 1.5). A galaxy appears displaced by the projection of its peculiar velocity, $\boldsymbol{u}$, along the line-of-sight $\widehat{\boldsymbol{r}}$ (Kaiser, 1986; Hamilton, 1998)

$$s = v + \boldsymbol{u} \cdot \widehat{\boldsymbol{r}}\,. \tag{1.66}$$

As a consequence of Eq.(1.66) the distribution of galaxies in redshift space, which is what we observe, is a distortion of the real distribution. This effect is called *redshift space distortions* (RSD). Kaiser (1986) showed that in linear theory RSD change the amplitude of the observed power spectrum,

$$P_g^s(k,\mu) = b_1^2 \left(1 + \frac{f}{b_1}\,\mu^2\right) P_m(k)\,, \tag{1.67}$$

where μ is the cosine of the angle between the line-of-sight and the object.

Despite all the complications, the measurement of the galaxy distribution still remains one of the best ways to map the matter density field and we can actually exploit RSD to indirectly measure the growth rate and test GR (Guzzo et al., 2008). The surveys that measure the galaxy angular positions and their redshift to map their three-dimensional distribution in the Universe are called *galaxy redshift surveys*. They are divided into two categories, spectroscopic and photometric redshift survey, and differ in the method used to measure the redshifts of the galaxies (see Chapt. 3). The first systematic redshift survey was the Center for Astrophysics redshift survey (CfA; Tonry & Davis, 1979), which was followed by the Sloan Digital Sky Survey (SDSS York et al., 2000) and the Two-degree-Field galaxy redshift survey (2dF; Colless et al., 2003).

1.4.1 Modern redshift surveys

After the pioneering work of the SDSS and 2dF galaxy redshift surveys many other projects mapped the galaxy distribution in the last two decades. In this section, I will introduce the three surveys I used during my thesis work.

The first survey is the VIMOS Public Extragalactic Redshift Survey (VIPERS; Guzzo et al., 2014). This survey is based on observation performed with the Visible MultiObject Spectrograph (VIMOS; Le Fèvre et al., 2003) mounted on the Very Large Telescope (VLT) of the European Southern Observatory (ESO) at Cerro Paranal in Chile. VIPERS consists in a spectroscopic sample of $\sim 90\,000$ galaxies with $i_{\rm AB} < 22.5$ and in the redshift range $0.5 < z < 1.5$. The spectroscopic targets of the survey were selected from the two fields W1 and W4 of the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS) Wide photometric catalogue.[1] In addition to the magnitude cut the targets were selected from the colour-colour plane $(r-i)$ vs $(u-g)$, where it was possible to identify a very efficient cut to select galaxies within the redshift range of interest. The two fields of VIPERS cover an area of $\sim 24\,\mathrm{deg}^2$ and the survey has a total volume of $\sim 5 \times 10^7\, h^{-3}\,\mathrm{Mpc}^3$. I used VIPERS data for the work presented in Chapts. 4 and 7.

The second catalogue is the extended Baryon Oscillation Spectroscopic Survey Data Release 16 quasar sample (eBOSS DR16Q; Lyke et al., 2020). The eBOSS survey is part of SDSS phase IV (SDSS-IV; Blanton et al., 2017) and is based on the observation of the Sloan

[1] https://www.cfht.hawaii.edu/Science/CFHTLS/

Foundation 2.5 m telescope at the Apache Point Observatory in New Mexico (Gunn et al., 2006). The eBOSS DR16Q sample contains a total of $\sim 350\,000$ quasars in the redshift range $0.8 < z < 2.2$. The sample is divided into two fields of view, the North and South Galactic cap, and covers an area of $\sim 4800\,\mathrm{deg}^2$ and a volume of $\sim 20\,h^{-3}\,\mathrm{Gpc}^3$. I used the eBOSS DR16Q sample in the work presented in Chapt. 6.

Finally, I worked with mock data that simulates *Euclid* observations. *Euclid* is a European Space Agency's medium-class mission (ESA), which was conceived to probe the nature of dark matter and dark energy by measuring the expansion of the Universe history and the growth of large-scale structures (Laureijs et al., 2011).[2] *Euclid* was successfully launched from Cape Canaveral, Florida, on July 1st 2023 on board of the SpaceX Falcon 9 launcher. It is now in the second Sun-Earth Lagrangian point at $\sim 1.5 \times 10^6$ km from Earth and has started observations. *Euclid* is a 1.2 m Korsch telescope with two mounted instruments, the Near-Infrared Spectrograph and Photometer (NISP; Maciaszek et al., 2022) and the VISual instrument (VIS; Cropper et al., 2016). NISP has three broadband near-infrared filters (Euclid Collaboration: Schirmer et al., 2022) and a set of grisms for slitless spectroscopy; VIS is a single optical broadband filter with high spatial resolution. The large field of view of *Euclid* ($\sim 0.5\,\mathrm{deg}^2$) was specifically designed to observe one-third of the sky ($\sim 15\,000\,\mathrm{deg}^2$) over the six years of its operations. *Euclid* will combine an imaging and a spectroscopic survey, with which the Euclid Consortium will respectively perform weak lensing and galaxy clustering analyses. The weak lensing survey will cover the redshift range $0.2 < z < 0.8$ and contain $\sim 1.5 \times 10^9$ galaxies, while the galaxy clustering analyses will interest $\sim 50 \times 10^6$ Hα emitters within $0.9 < z < 1.8$. In Chapt. 5 I will discuss my work with *Euclid* mock data.

1.4.2 Analysis pipeline

The extraction of cosmological information from the raw observation of a redshift survey is a complex process. In this section, I will schematically describe this pipeline as the main topic of this thesis is the management and analysis of redshift survey data at different stages of data processing.

In the case of traditional spectroscopic redshift survey the spectroscopic targets are selected beforehand from a *parent* photometric catalogue, e.g., the spectroscopic targets of VIPERS were selected from the CFHTLS Wide photometric catalogue. Then, we can define the sample *completeness* and its *purity* with respect to the parent sample. Slitless spectroscopic surveys, e.g., *Euclid*, miss this pre-selection step and the sample completeness and purity have a less straightforward definition (see Sect. 5.1). In the case of a slitless survey, in order to improve these metrics we can apply a target post-selection, which is applied to the already observed data.

After their acquisition, the data undergo a first reduction, e.g., one-dimensional spectra are extracted from the two-dimensional observation and spurious objects, such as stars, are identified, then the redshifts of the objects are measured. Potentially we now have all the information needed to measure the galaxy field summary statistics. However, when analysing real data a few additional steps are required. First, we need to identify any systematic effects related to observation, such as the redshift rate failure, the fibre collision in the case of spectroscopic surveys, or imaging inhomogeneities. Second, we need to determine the *window function* of the survey, which represents its footprint on the sky. The window function has an angular component, which is determined by the peculiar shape of the survey on the sky surface, and a radial component, which depends

[2] https://www.esa.int/Science_Exploration/Space_Science/Euclid

on the radial selection of the survey.

Starting from the survey footprint the *random* catalogue and *mock* catalogues are built. The random catalogue traces the mean density of the observed galaxy sample in the case of no clustering. The random catalogue is required to estimate the field summary statistics (e.g., Landy & Szalay, 1993; Yamamoto et al., 2006). The mocks are simulated catalogues that represent different realisations of the Universe and are used to determine the covariance of the summary statistics. Finally, the cosmological parameters of interest are measured with a Bayesian analysis.

The first half of this thesis (Chapts. 4 and 5) is related to photometric redshift measurements and sample selection, while in the second half (Chapts. 6 and 7) I will discuss optimal methods to extract cosmological information from the galaxy density field.

CHAPTER 2

Machine learning

Machine learning (ML) is a branch of the broader field of *artificial intelligence*. We call artificial intelligence any software that automates routine labour starting from the simplest repetitive task to the understanding of speech and figures. Computers can easily solve and perform problems that we describe with a list of mathematical rules and usually prove to be challenging tasks for human beings. On the other hand, computers struggle to solve tasks that are difficult to describe formally, but that are performed daily by humans, such as recognising faces, animals, and words. To solve these tasks a computer should learn from experience rather than use a set of mathematical rules or hard-coded knowledge.

Machine learning and *deep learning* are the branch of artificial intelligence in which a computer gathers its own experience by directly extracting patterns from raw (or semi-raw) data. Figure 2.1 describes the hierarchy inside the field of artificial intelligence and what characterises each of its branches. As described above, the simplest artificial intelligence algorithm is a rule-based system, where we feed the inputs to and hand-designed program. In the case of classic machine learning, the algorithm is 'free' to find the mapping between a *representation* of the data and the output. The representation of the data contains *features*, which summarise pieces of information.

An example of a classic machine learning algorithm is logistic regression, which is a simple model that usually gives binary answers. Logistic regression is widely used in medicine to predict, e.g., mortality in patients or the probability of developing specific diseases (e.g., Boyd et al., 1987; Biondo et al., 2000). These algorithms take as input a representation of the data previously determined and synthesised by the doctor, who directly analyses the patient. Then, we can say that these features are hand-designed and the algorithm does not have any control over their definition. The representation of the data can heavily influence the performance of an algorithm: for example, for humans, it is easier to count using Arabic numbers rather than Roman numbers.

There are tasks in which it is difficult to understand the best representation of the data. In cosmology, for example, we usually describe the galaxy field with its power spectrum, but we know that the field is non-Gaussian on small scales (and also on large scales, in the case of primordial non-Gaussianities). Thus, the power spectrum representation of the data is sub-optimal. A solution to this representation problem is not only to have the algorithm find the mapping from a representation to output, but also to have it find the representation itself. This is what is usually called *representation learning*. When the representation learning algorithm is able to express features in terms of simpler representations we talk about deep learning. An example is a model that analyses images. Given, e.g., the image of a house, a deep learning algorithm will represent it in terms of the simpler shapes by which it is composed, which in turn can be represented by straight

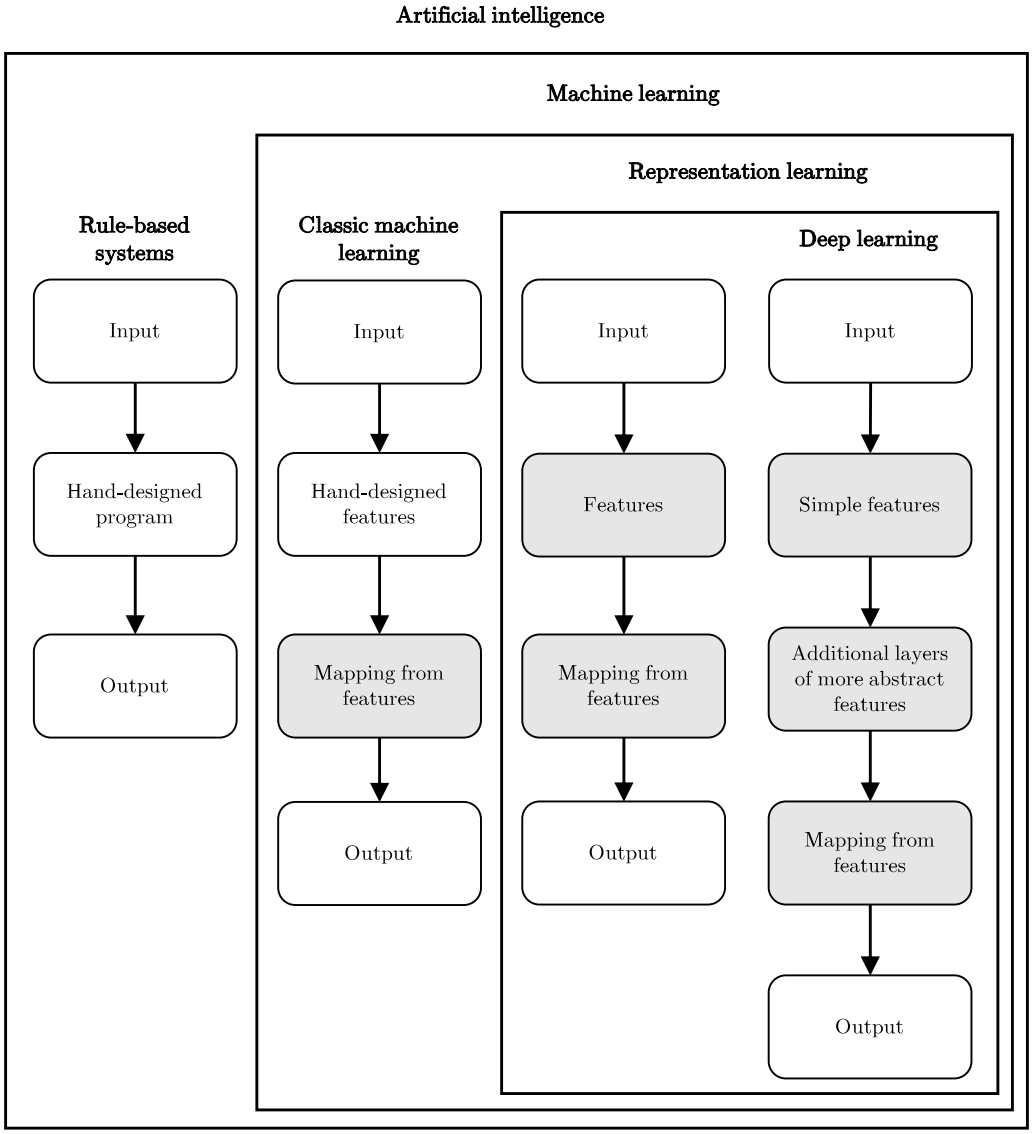

Figure 2.1: Flowcharts showing how the different parts of an artificial intelligence system relate to each other within different artificial intelligence disciplines. The shaded boxes indicate components that are capable of learning from the data. Figure from Goodfellow et al. (2016).

or curved lines. The most famous example of a deep learning model is the neural network (NN). Neural networks will be the most used algorithm in this thesis work (see Chapts. 4, 5, and 7) and I will discuss them in more detail in the next section.

Finally, machine learning algorithms can be *supervised* or *unsupervised*. Supervised learning works with labelled data, where the algorithm learns to make predictions based on the provided examples. The goal of a supervised algorithm is to find a mapping between the input and the output based on the labelled data. During the training, the algorithm adjusts its internal parameters to minimise the *loss function*, which depends on the difference between the prediction and the label. Supervised learning is usually employed in regression and classification problems. We evaluate a supervised algorithm's performance from its ability to accurately predict target values of previously unseen data. On the other hand, unsupervised learning deals with unlabelled data and aims to discover patterns or structures within data without explicit guidance. Unsupervised algorithms are used in clustering, dimensionality reduction, and anomaly detection tasks. As they do not deal with labelled data what is minimised during the training process is not the loss function, but, depending on the task, other functions are minimised. For example, in clustering algorithms what is minimised is the distance between the features or metrics of cluster quality.

2.1 Neural networks

A neural network is a computational model inspired by the human brain, consisting of interconnected nodes, also known as *neurons*, organised in *layers*, with the ability to learn patterns and make predictions from data through iterative training processes (LeCun et al., 2015).

In a feed-forward neural network, the information flows in only one direction, from one layer to the following, and there are no connections between neurons of the same layer. Each layer is characterised by the *activation function*, $h^{(i)}$, its neurons apply to their input features. These activation functions can be distinct for each layer or uniform throughout the entire network, but they must be nonlinear functions. Given the input data, $\boldsymbol{x}$, the output of the network is

$$\boldsymbol{y} \equiv h(\boldsymbol{x};\boldsymbol{\vartheta}) = h^{(n)}\left(h^{(n-1)}\left(\dots\left(h^{(1)}\left(\boldsymbol{x};\boldsymbol{\vartheta}^{(1)}\right);\dots\right);\boldsymbol{\vartheta}^{(n-1)}\right);\boldsymbol{\vartheta}^{(n)}\right), \tag{2.1}$$

where $\boldsymbol{\vartheta}^{(i)}$ are the free parameters of the i-th layer. The length of the chain of Eq. (2.1) is equivalent to the number of layers n and represents the depth of the network. We call the first layer *input layer* and the last one is the *output layer*. All the other layers, those with which the user has no direct interaction, are the *hidden layers*.

As information progresses from one layer to the next, it undergoes a linear transformation. We can reorganise the parameter vector $\boldsymbol{\vartheta}^{(i)}$ of a layer into an $n \times m$ matrix $\boldsymbol{W}^{(i)}$, where n is the number of neurons in the layer and m is the number of neurons in the previous layer, and an n-component vector $\boldsymbol{b}^{(i)}$ known as *bias*. The input on the i-th layer is

$$\boldsymbol{x}^{(i)} = \boldsymbol{W}^{(i)} \cdot h^{(i-1)}\left(\boldsymbol{x}^{(i-1)}\right) + \boldsymbol{b}^{(i)}, \tag{2.2}$$

where $h^{(i-1)}\left(\boldsymbol{x}^{(i-1)}\right)$ is the m-component output of the previous layer. The components of $\boldsymbol{W}^{(i)}$ and $\boldsymbol{b}^{(i)}$ are referred as *weights*. To summarise, the features first undergo a linear transformation (see Eq. 2.2) and then a nonlinear one through the neuron activation

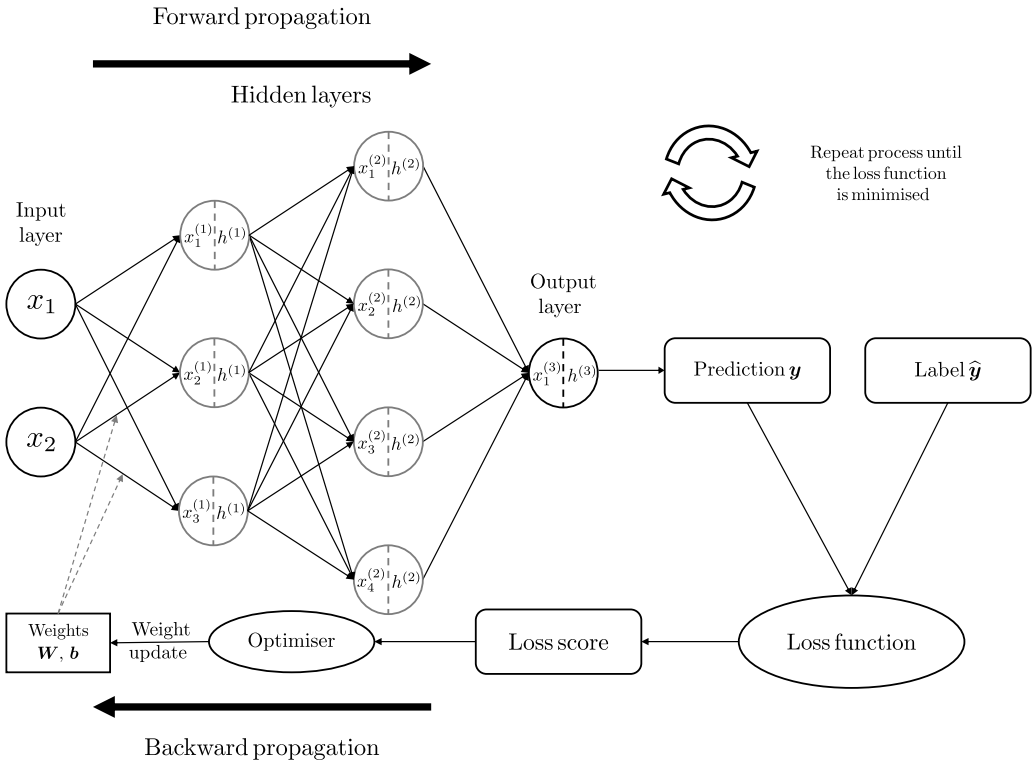

Figure 2.2: Schematic representation of the training process. First, the inputs are forward propagated to the output; then, the model prediction is compared to the label through the loss function. Finally, in the backward propagation, the gradients of the loss function with respect to the outputs of the network are computed. These gradients are used by the optimiser to update the model weights. This process is a training iteration and is repeated until the loss function is minimised.

function. This simple scheme is repeated for each layer and thanks to it neural networks have the potential of fitting any nonlinear function (LeCun et al., 2015).

2.1.1 Training

During the training process, the network weights are iteratively adjusted to minimise the difference between the predicted and real outputs. This process enables the model to learn and generalise patterns from the training data. More rigorously, the objective of the training of a neural network is to approximate its output $\boldsymbol{y} = h(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b})$ to the real relation between the input $\boldsymbol{x}$ and the label $\widehat{\boldsymbol{y}} = h^*(\boldsymbol{x})$ by optimising the weights $\boldsymbol{W}$ and $\boldsymbol{b}$. Figure 2.2 schematically presents the steps that take place during the training process. In this section, I will discuss in detail all these steps.

When a neural network is initialised its weights are just random numbers and the network output will be completely uncorrelated to the real value of the label. The training of a neural network can be divided into three main steps: the *forward propagation*, the *loss computation*, and the *backward propagation*. The sequence of these three steps is a *training iteration*. During the forward propagation input data pass through the network

as described in the previous section (see Eqs. 2.1 and 2.2) and predictions are calculated using the current weights.

The next step is the loss computation. The loss function compares the prediction $\boldsymbol{y}$ with the label $\widehat{\boldsymbol{y}}$ and outputs a *loss score*. The loss function quantifies in the loss score the distance between the prediction and the label. Different problems require different loss functions, e.g., classification and regression tasks need different loss functions. The most diffuse loss function for classification is the *cross-entropy* loss function, which, in the case of binary classification, reads as follows

$$L^{\text{BCE}}(\widehat{y}, p) = -\left(\widehat{y}\,\ln(p) + (1-\widehat{y})\,\ln(1-p)\right)\,, \tag{2.3}$$

where $\widehat{y}$ is the label, which can either be 0 or 1, and p is the predicted probability of belonging to class 1. The cross-entropy penalises the model more heavily for making confident incorrect predictions and encourages it to be confident in the correct predictions. Many other loss functions exist for classification, but they usually are modified versions of this cross-entropy loss. For regression tasks, the most diffuse loss functions are the mean root squared error, the mean squared error, and the mean absolute error. In my work, I mainly use the mean squared error loss function or loss functions derived from it. The mean squared error loss function reads as follows

$$L^{\text{MSE}}(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y}_i)^2\,, \tag{2.4}$$

where $\widehat{y}_i$ and y_i are the label and the prediction of the i-th input and m is the number of samples observed during a training iteration. Equation (2.4) is valid for a model that outputs one value for each input. In the case of regression over multiple parameters, we usually take either the sum or the mean of the mean squared error of each parameter as the loss score.

Backward propagation is the last and arguably the most important step of the training iteration. It is an optimisation algorithm used to adjust the model weights based on the computed gradients of the loss function with respect to these parameters. The objective is to minimise the loss, improving the model's ability to make accurate predictions. Backward propagation starts by calculating the gradient of the loss with respect to the output of the neural network. The computed gradients from the backward pass are used to update the model weights. This is part of the gradient descent optimisation algorithm and the algorithm that performs it is also called the *optimiser*. The weights are updated based on the gradients and shifted in the direction that minimises the loss. After this first weight update, the gradients are also propagated backward through the layers of the network and they are used to calculate the local gradients with respect to the layer inputs and outputs. These local gradients are used to update the weights of each layer individually. The general formula for the weight update is

$$w_{\text{new}} = w_{\text{old}} - l_r\,\partial_w L\,, \tag{2.5}$$

where w_{new} is the updated weight, w_{old} is the current weight, l_r is the learning rate, which determines the step size in the weight update, and $\partial_w L$ is the gradient of the loss function with respect to the corresponding weight w_{old}. The update of Eq. (2.5) can be modified by more complex optimiser algorithms, which are used in the gradient descent phase; however, for the second update of the weights the standard update formula is used. One of the most common optimisers is the Adaptive moment estimator (Adam;

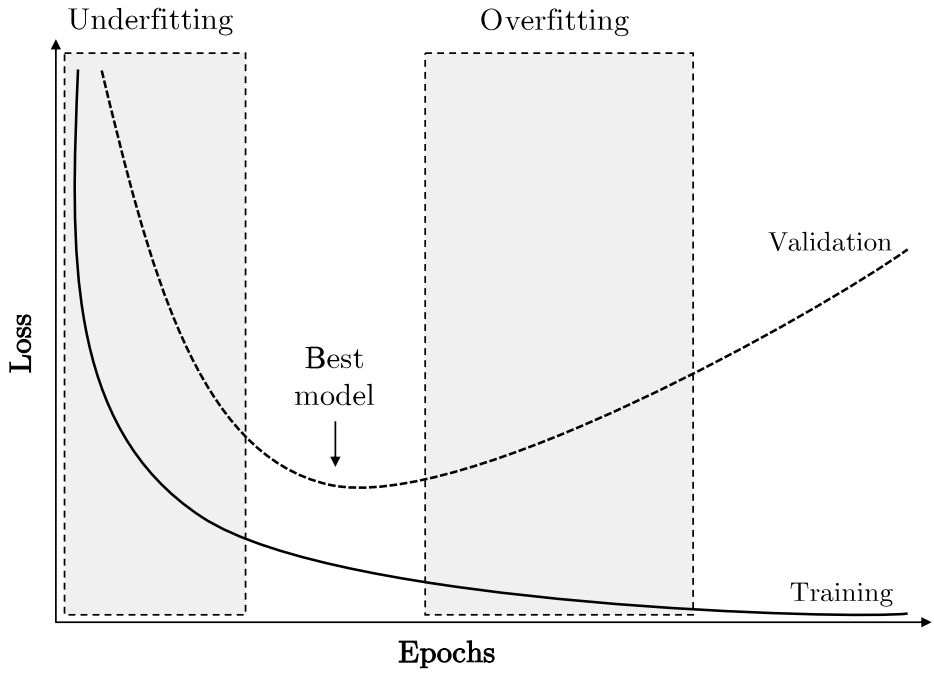

Figure 2.3: Example of the loss score as a function of the training epoch. The solid line represents the loss of the training set, while the dashed curve is the loss of the validation set. In the first phase of the training the model is underfitting the data, but as training goes on the model starts to overfit the training set. It is important to identify the model that minimises the validation loss and it is possible to stop the training when this model has been selected.

Kingma & Ba, 2014), which adjusts the learning rate for each parameter based on both the first and second moments of the gradients. In all my work I used the Adam optimiser.

With the layer-wise weight update, the training iteration is concluded. This process should be repeated until the loss function is minimised. During a training iteration, the model does not necessarily see the whole training data set. We say that the training has completed an *epoch* when the entire training set has passed through the network. An epoch can be composed of more than one training iteration. This happens when the *batch size* is smaller than the number of data points in the training sample. The batch size refers to the number of training examples utilised in one training iteration. Larger batch sizes may provide more accurate gradient estimates but require more memory. Smaller batch sizes introduce more noise but may lead to faster convergence. The number of epochs and the batch size are *hyper-parameters* of the model.

As discussed above, to train a network we use a *training set* of data points, however, the performance of a network is determined by its ability to make correct predictions for unseen data. We say that a data point is unseen if it is not part of the training set and has not had any role in the optimisation of the network weights. We can use unseen data not only to determine the performance of a model after the training, but also to monitor the network performance during the training itself. In the former case, we call the set of unseen data the *test set*, while in the latter case, we talk about *validation set*. The validation set has a critical role in the training of machine learning models and

acts as an essential benchmark to evaluate the model performance on unseen data (see Fig. 2.3). As models train, they face the risk of overfitting, capturing noise and details from the training set that do not generalise well. Conversely, underfitting arises when a model is too simplistic and fails to grasp the underlying patterns in the data. The use of the validation set can prevent these issues. By evaluating the model on data it has not encountered during training, we can discern whether the network strikes the right balance between complexity and generalisation. Additionally, we can also exploit the validation set to determine when we can stop the training of the network, this practice is called *early stopping*, and to select the best weights of the model. Early stopping consists of monitoring the model performance on the validation set and halting training when further iterations produce diminishing returns or risk overfitting. Moreover, the validation set helps in selecting the best weights for the network. We usually say that the best model is the one that minimises the loss evaluated over the validation set rather than the training set, doing so we select a model that not only fits the training data well but also exhibits robustness and effectiveness on new and unseen examples.

2.1.2 Data structures and neural network architectures

The neural network described in Sect. 2.1 is usually referred to as a *dense* neural network. The name remarks the fact that all the neurons of a layer are connected to the neurons of the following layer through the linear transformation of Eq. (2.2). To summarise the layers are densely connected. As mentioned in the previous sections, this structure has the potential of fitting any nonlinear function $\widehat{\boldsymbol{y}} = h^*(\boldsymbol{x})$. However, this statement, broadly known as *universal approximation theorem* (e.g., Hornik et al., 1989), is based on a number of assumptions that cannot be met in real-life models, e.g., arbitrary width of the layer or arbitrary depth of the network. To actually reach a good approximation of the function $h^*(\boldsymbol{x})$ it is important to identify the appropriate hyper-parameters of the model. In Sect. 2.1.1 I mentioned the batch size and the number of epochs for the training as hyper-parameters, but the loss function, the optimiser, and the neurons' activation functions are hyper-parameters as well. Broadly speaking any choice we make when building the network is a hyper-parameter that could be optimised for the problem at hand.

Arguably, the most important hyper-parameter is the *architecture* of the network. By network architecture, we refer to the overall design and structure of the neural network. The arrangement, connectivity, and number of components, such as layers and neurons, are part of the network architecture and determine how the information flows through the network. In particular, some architectures are specifically designed to manage different types of data structures. In the case of data with a grid-like topology, such as pixelised images or time series, an optimal architecture is the *convolutional neural network* (CNN; LeCun, 1989; Goodfellow et al., 2016). A network is called convolutional if at least one of its layers employs convolution instead of the general matrix multiplication described in Eq. (2.2).

Figure 2.4 represents a two-dimensional convolution. Rigorously, this operation is a *cross-correlation* and reads as follows

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m, j+n)\, K(m,n)\,, \tag{2.6}$$

where I is the input two-dimensional grid and K is the $m \times n$ two-dimensional convolution *kernel*, also known as filter. Many machine learning libraries (e.g., PyTorch; Paszke et al., 2017) implement the cross-correlation of Eq. (2.6) and call it convolution. Convolution has three main advantages over a dense neural network: it has *sparse interaction*,

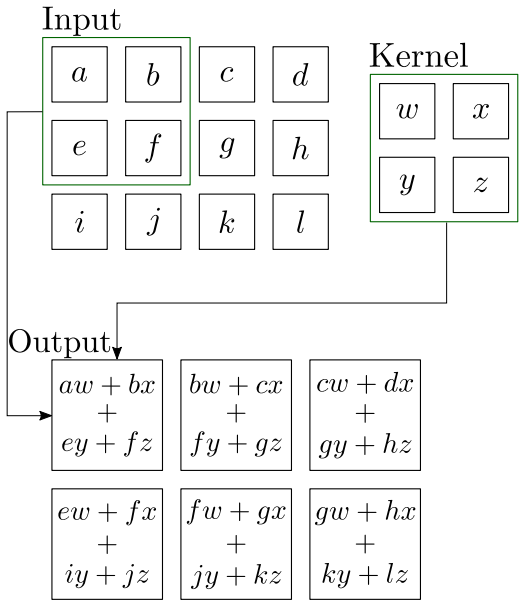

Figure 2.4: A schematic representation of a 2-dimensional convolution. For this example, the output is restricted to only positions where the kernel lies entirely within the image. The green boxes with arrows indicate how by applying the kernel to the upper-left region of the input produces the upper-left element of the output. Figure from Goodfellow et al. (2016).

parameter sharing, and an *equivariant representation*. As mentioned above in dense neural networks a neuron is connected to all the neurons of the previous and following layer, so every output unit interacts with every input unit. However, if the kernel is smaller than the image the CNN is processing, the interactions between input and output units are sparse. In Fig. 2.4, since we apply a 2×2 filter to a 3×4 input, the outputs only depend on four of the input units. This is a sparse interaction. This does not necessarily mean that there is no interaction between input neurons far from one another, e.g., a and l. If the convolutional neural network is deep enough the deeper outputs can indirectly interact with a larger portion of the input image, if not the whole image.

The convolution operation also enforces parameter sharing, which corresponds to the fact that it uses the same parameter for more than one operation in the model. In Eq. (2.2) the weight matrix W has $n \times m$ different parameters, where m and n are the numbers of neurons in the input and output layer respectively. As a consequence of this, there are $n \times m$ weights to be stored and $n \times m$ operations to be performed. On the other hand, the convolution applies more than once the same weights. In Fig. 2.4 we see that the same 4 elements of the kernel appear in each output unit. This greatly reduces the memory requirements and the number of operations the model performs. Finally, convolution is also equivariant to translation. It means that if a specific feature in the input is shifted the corresponding representation in the output is shifted in the same way. This means that the network is able to identify the same pattern in different regions of an image.

In general, we build a convolutional neural network as a sequence of convolutional layers whose output is flattened and processed by a dense neural network into the output neurons. Each convolutional layer enforces Eq. (2.6) multiple times with different filters. Therefore, with each convolution the input is compressed and its information content is divided into multiple outputs. In this process, each filter learns a specific feature of the input. We can also combine convolutional layers with normalisation or pooling operations, but I will discuss each one of these solutions in the description of the specific CNNs I designed for this thesis work (see Chapt. 7). Finally, the dense layers post-process the convolutional layers output. This output is a $(N+1)$-dimensional tensor, where N is the original dimension of the input (e.g., 2 for a one-colour image), and the additional dimension is introduced by the use of more than one filter in each convolutional layer. Potentially, we could use only one dense layer to map the convolution output to the final output of the network. However, as I just described, the convolution can output a very large data vector. In these cases, it is more efficient to compress the information not with one layer, but with a sequence of dense layers.

Another data type that can be interesting in LSS studies is the *graph*. Graphs are used to represent interaction-based data, e.g., friend and citation networks or molecular structures. Graphs can be an alternative way to represent the cosmic web and we can exploit their versatility for any type of unstructured data. Mathematically, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a set of *nodes* $\mathcal{V}$ and *edges* $\mathcal{E}$. We denote a node with $v_i \in \mathcal{V}$ and the edge from node v_i to node v_j as $e_{ij} = (v_i, v_j) \in \mathcal{E}$. Edges normally have a direction. In the definition above e_{ij} goes from v_i to v_j, while e_{ji} goes from v_j to v_i. Additionally, we can define the node $v \in \mathcal{V}$ *neighbourhood* as $N(v) = \{u \in \mathcal{V} | (u, v) \in \mathcal{E}\}$, which corresponds to the set of nodes that have an edge going to v. We collect all the edge information in the *adjacency matrix* $\mathbf{A}$. This is a $n \times n$ matrix, where n is the number of nodes in the graph, with $A_{ij} = 1$ if $e_{ij} \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. As edges are directed, the adjacency matrix is not symmetric a priori. It becomes symmetric in the case of an *undirected* graph. In this case, if two nodes are connected there is a pair of edges with opposite directions between them. Finally, both nodes and edges in a graph may have attributes. The node

attributes are collected in $\mathbf{X} \in \mathbb{R}^{n \times d}$, where d is the number of attributes the nodes have. Analogously, we define an edge attribute matrix $\mathbf{X}^e \in \mathbb{R}^{m \times f}$, where m is the number of edges in the graph and f the number of their attributes.

Graphs are extremely flexible in comparison to grid-structured data and can easily represent irregular data sets. This comes at the cost of complexity in the application of deep learning to graphs (Hamilton, 2020). The networks designed to manage graphs are called *graph neural networks* (GNNs). Graph neural networks are designed to manage inputs with varying dimensions as graphs representing the same type of object, e.g., a molecule, can have a different number of nodes and edges. Under the name of GNNs goes a large number of deep learning models that solve many different tasks (Wu et al., 2019). Graph neural networks can perform node classification, node regression, or predict missing node attributes. We can also use GNNs for relation prediction or graph classification and regression. These last two tasks consist of the extraction of graph-level information and are the most similar to standard deep-learning methods.

What makes all these graph neural networks very efficient is the ability to share information between node neighbourhoods. In comparison to the grid convolution defined for CNNs (see Eq. 2.6) the information sharing in GNNs is based on the concept of *message passing*. Many GNN models implement and exploit message passing in different ways, but this mechanism is always characterised by a series of defined steps. The three main phases of message passing are *message generation*, *message aggregation* and *update*. During message generation nodes send messages to their neighbours. These messages are generated by combining the sender node's attributes with edge-specific information. After receiving the messages from the neighbourhood a node aggregates them before updating its representation by combining its current attribute with the aggregated message. The most general way to describe the message passing operation that transforms the i-th node attributes $\boldsymbol{x}_i$ into the updated attributes $\boldsymbol{x}'_i$ is as follows (Gilmer et al., 2017)

$$\boldsymbol{x}'_i = \gamma\left(\boldsymbol{x}_i, \square_{j \in N(i)} h\left(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}^e_{ji}\right)\right) , \tag{2.7}$$

where $\gamma(\cdot)$ is a differentiable update function, $\square_{j \in N(i)}$ is a permutation-invariant aggregation function, e.g., a summation or an average over the neighbourhood messages, and $h(\cdot)$ is the differentiable message function. The j-th neighbour attributes are $\boldsymbol{x}_j$ and $\boldsymbol{x}^e_{ji}$ are the attributes of edge e_{ji}, which sometimes undergo a transformation before the message creation (e.g., Gong & Cheng, 2018). We update the attributes of all the nodes in the graph sharing the same message, aggregation, and update functions. Therefore, message passing itself is a form of parameter sharing. Note that the update and message functions can also be dense neural networks. This choice makes the message-passing operation extremely general, but it also makes a GNN complex to manage memory-wise.

Similarly to convolutional neural networks, where there can be a sequence of convolutional layers, in graph neural networks, we can build a sequence of message-passing layers. Each time we repeat a message passing operation the information is propagated further away from the node it originated from. Finally, when the graph neural network performs a graph regression or classification we aggregate the information from all the nodes and usually pass it through a sequence of dense layers (as we do for CNNs) to obtain the final output.

In summary, each task requires an attentive design of the deep learning model used to tackle it. In this section, I specifically discussed convolutional and graph neural networks as I used them during my thesis work. However, other neural network architectures exist that can be used to manage different data types or tasks, e.g., recursive neural networks for sequence-like data or generative adversarial networks to generate new synthetic data sets from existing ones.

Part I

Redshift and sample selection

CHAPTER 3

Galaxy distances and redshifts in cosmology

Measuring distances in astrophysics is no trivial task. We can not use the same method for all the distance scales. Therefore, over the years, astronomers and astrophysicists developed different methods to measure the distances of increasingly distant objects. These methods overlap on some distance scales so that their precision can be tested. Therefore, the precision of each step of this 'distance ladder' depends on the precision of the previous one until the first step is reached. The first step must be model-independent in order to make the whole ladder coherent.

This first method is parallax. Parallax is the measure of the maximum angular annual displacement of nearby stars due to the revolution of Earth around the Sun. Knowing a star's parallax and the Earth-Sun distance we can measure the distance of the star from the Sun. We can measure the parallax of stars as far as $1\,\mathrm{kpc}$, which are galactic objects. However, now, thanks to the new astrometric data acquired with the ESA satellite *Gaia*, it is possible to reach distances up to $10\,\mathrm{kpc}$ using the parallax method.

The following step is the so called spectroscopic parallax. This method measures the distance of a star confronting its apparent magnitude, which is what we measure, with its absolute magnitude extrapolated from its temperature using the Hertzsprung-Russel (HR) diagram. The temperature of the star is measured with spectroscopy and the HR diagram is built from stars whose distances were measured with parallax. Spectroscopic parallax reaches stars as far as $10\,\mathrm{kpc}$, which are still galactic objects.

Next, there are some methods based on variable stars, the most famous of which are the variable Cepheids.[1] From parallax and spectroscopic parallax measurements we were able to understand the relation between the absolute magnitude and the period of luminosity variation for Cepheid stars. Therefore, by measuring the variation period of a distant variable star and its apparent magnitude we can estimate its distance. With variable star methods, we can measure the distance of every galaxy in which we can resolve single objects (in particular variable stars). Variable stars are used to measure distance up to $\sim 10\,\mathrm{Mpc}$, which is the distance of the Virgo cluster, one of the nearest galaxy clusters.

At further distances, we can not resolve single objects in a galaxy except for masers and supernovae. Masers are monochromatic point sources in the microwave wavelengths. They are produced by specific quantum transitions in non-thermal gas populations and they can occur in gas accretion disks around the black holes in the centre of galaxies. These masers are produced by transitions of water molecules and are also known as water masers. Knowing the rest-frame wavelength of the transition, assuming a Keplerian rotation of the maser around the central black hole, and measuring the angular distance of the maser from the centre of its orbit, its velocity, and its acceleration, we can estimate the physical distance of the maser from the black hole and have a

[1] The variable Cepheids method was used by Hubble to measure the distance of the nearest galaxies.

direct measurement of the distance of the maser from the Sun. The most distant maser is at $\sim 150\,\mathrm{Mpc}$. However, masers are rare, up to date there are only six masers with a distance measurement obtained with this method. For this reason, we mainly use them to make an additional calibration of other extra-galactic methods.

The other type of objects that we can resolve within other galaxies are supernovae. Supernovae are explosions during which the luminosity of a star can exceed the luminosity of a whole galaxy. A supernova is a very complicated phenomenon. However, supernovae type Ia, which are the explosion of white dwarfs that reach the Chandrasekhar mass, have a very peculiar characteristic: their peak absolute luminosity is constant and correlates with the peak width. Therefore, if a supernova Ia explodes in a far galaxy we can observe it and, knowing its absolute luminosity, we can measure its distance.

However, also supernovae are rare phenomenons, thus for the majority of the furthest galaxies the only information we receive is the electromagnetic spectral energy distribution (SED) of the whole galaxy which, on first approximation, is the sum of the spectra of the galaxy stars. Due to the expansion of the Universe the galaxy SEDs are stretched toward longer wavelengths and appear redshifted. As shown in Sect. 1.1.3, a wave packet wavelength is red-shifted and its flux density is dimmed by a factor $(1+z)$. Moreover, from Eq. (1.38) we are able to calculate distances from redshifts given a set of cosmological parameters. Thus, in observational cosmology, redshift measurements are distance measurements once we assume a cosmological model.

Before we start to describe in detail the methods used to measure cosmological redshifts we need to mention that baryon acoustic oscillations can be used to measure cosmological distances as well. From structure formation theory we know that the physical dimension of the BAO is fixed by the sound horizon at recombination and its angular dimension can be measured using the angular correlation function for a sample of galaxies measured in a survey (see Sect. 1.2.1). Having this information provides the relationship between redshift and angular diameter distance from which we can constrain the parameters of the cosmological model.

3.1 Spectroscopic redshift

There are two different methods used to measure redshifts on cosmological scales, both with advantages and flaws. They both aim to identify characteristic features in the galaxy's spectral energy distributions and measure how much they were redshifted. Let us start our description from the more straightforward one which is spectroscopy.

Spectroscopy provides a measure of the SED with high wavelength resolution. Galaxy spectra are the sum of the spectra of their content. Galaxies have different spectra depending if their stellar population is young or old and if they are rich or poor in interstellar medium. However, all galaxy spectra have in common two features: the Lyman break and the Balmer break. Any photon with a wavelength shorter than $912\,\text{Å}$, Lyman continuum, will be absorbed by neutral hydrogen gas, both in the galaxy itself and in the intergalactic medium. Not only will these photons be absorbed, but also any with a wavelength corresponding to the line in the Lyman series. Consequently, we do not receive light with a wavelength shorter than the Lyman-α line at $1216\,\text{Å}$. Hence, a step is produced in the galaxy SED at $1216\,\text{Å}$, this feature is known as *Lyman break* and it is marked by the black solid line on the left in Fig. 3.1.

Analogously, the Balmer break can be explained. This spectral feature is related to the hydrogen Balmer series from $3646\,\text{Å}$ to $4000\,\text{Å}$ and is seen in stellar spectra. Photons with wavelengths less than $3646\,\text{Å}$ have sufficient energy to excite the Balmer transition

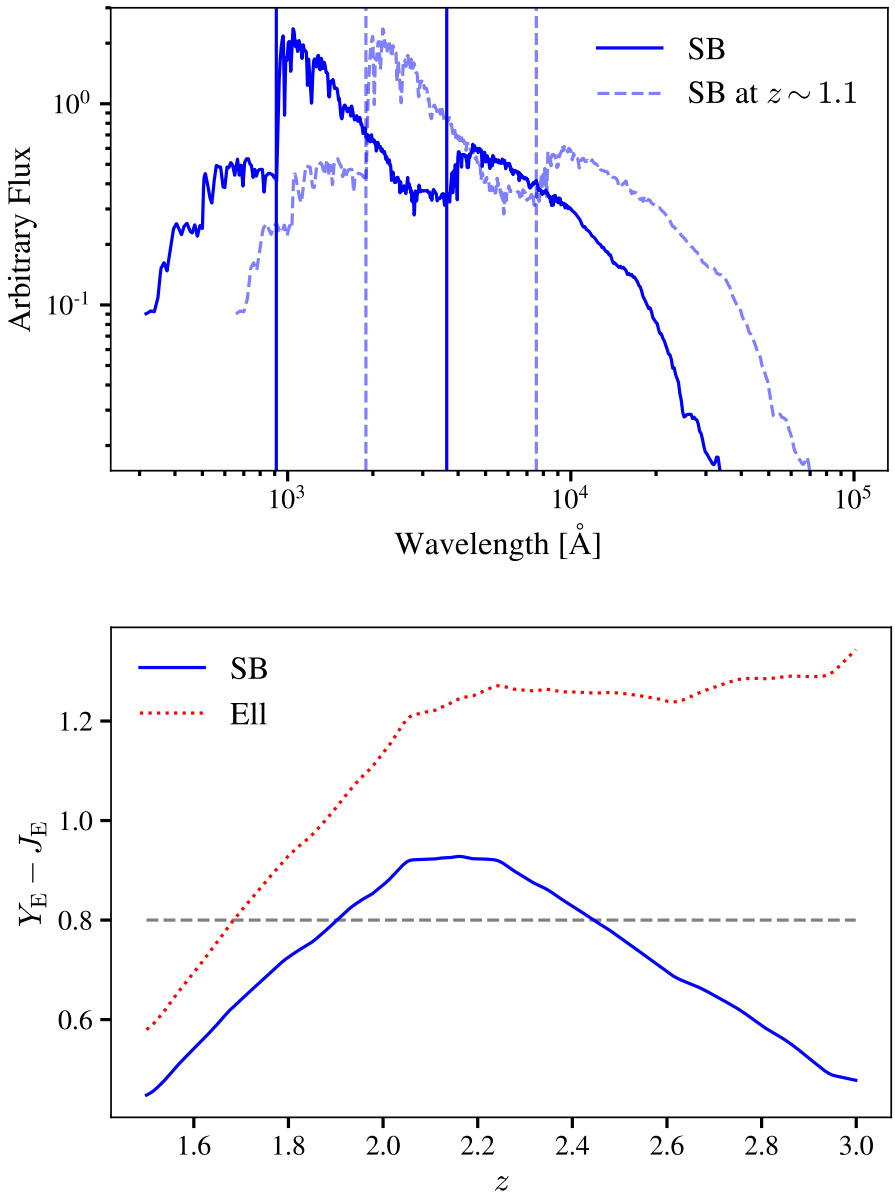

Figure 3.1: Spectroscopic and photometric redshifts. *Top*: The SEDs at $z = 0$ of a Starburst galaxy (SB). The solid vertical lines indicate the Lyman and the Balmer breaks in the galaxy rest frame. The dashed line is the Starburst galaxy SED shifted at $z \sim 1.1$, while the dashed vertical lines indicate the Lyman and the Balmer breaks shifted at that same redshift. The shown SED is from the COSMOS templates (Ilbert et al., 2009). *Bottom*: Observed $Y_E - J_E$ colour as a function of redshift for the starburst galaxy plotted above and an elliptical galaxy (Ell). The dashed horizontal line indicates $Y_E - J_E = 0.8$ and shows the photo-z degeneracy.

and are thus absorbed by hydrogen atoms in stars. The step we observe in the SED at 4000 Å is therefore called *Balmer break*. It is the excited hydrogen to produce this feature, therefore the Balmer break is more visible in galaxies with a higher star-formation rate. In Fig. 3.1 the Balmer break is marked by the solid line on the right.

Observing the spectrum of a far galaxy enables us to measure with high precision how much these features, with other characteristic emission or absorption lines, were redshifted, and infer the cosmological redshift z of the galaxy (see Fig. 3.1). With this method, redshifts are measured with high precision with an error lower than $5 \cdot 10^{-3}(1+z)$ for resolution $R > 200$ (e.g. Guzzo & Vipers Team, 2017). One of the first spectroscopic redshift surveys with a high number of objects ($\sim 10^6$) is the Sloan Digital Sky Survey (York et al., 2000), which enabled cosmologists to study the three-dimensional structure of the Universe, the properties of galaxies and their scaling in redshift and much more (see Sect. 1.4.1).

However, measuring spectroscopic redshifts is time and resource-consuming. Observing the spectra of distant faint objects, which are affected by cosmological dimming, requires large telescopes and long-time exposure in order to be able to decompose the light signal into the spectrum. The faint object signal-to-noise ratio is often too low for redshift measurements. In addition, for a solid spectroscopic redshift measurement, at least two spectral features are needed, which means a wide wavelength coverage is essential. Therefore, even if spectroscopy is the best method to probe the local and near Universe it becomes less suitable for this task the farther we want to study.

3.2 Photometric redshift

An alternative to spectroscopic redshifts is the so called photometric redshift, known as photo-z. This method was first proposed by Baum (1957). Photometric redshift is based on the idea that we can constrain the SED shape with wide band flux measurements and infer the redshift of the object from its *observed colour*, which is the difference in magnitudes of two bands, that is related to SED broad features such as the Lyman and Balmer breaks. The breaks are steep changes in the SED, therefore we should be able to detect gradients between observed fluxes in adjacent filters, which are the colours, and identify the position of a break.

Figure 3.1 bottom panel shows how the colour of a galaxy depends on its redshift. The first thing we notice from Fig. 3.1 is that the colour has a maximum at a given redshift that depends on the galaxy type. When designing a photometric redshift survey, the filters must be chosen in order to observe key features of the redshift and objects of interest. Second, we see that there is a degeneracy for the redshift solution in the colour space. This degeneracy, known as photo-z degeneracy, can be broken by combining several colours. Therefore, also a photo-z survey must have a wavelength coverage as broad as possible with multiple filters.

In principle, if we follow these prescriptions, we can derive the redshift of every source in an imaging survey. The price we must pay for this surveying completeness is the redshift precision, which decreases by one or two orders of magnitude compared to spectroscopic measurements. Photo-z popularity has increased in the last two decades. If well calibrated with spectroscopy, the photometric redshifts enable us to make statistical analyses of larger samples than spectroscopic redshifts do. Moreover, it makes it possible to infer redshifts for very faint objects. This means that we simultaneously have a more complete observation of the region than spectroscopy probes and the possibility to study regions with higher redshifts than spectroscopy does. Photo-z preci-

sion is enough to study galaxy evolution, formation and properties with cosmic time, to search primordial galaxies (e.g., Dunlop et al., 2012), to identify galaxy clusters (e.g., Finoguenov et al., 2007) and limited precision analyses of galaxy environment (Etherington & Thomas, 2015; Malavasi et al., 2016). In addition, photo-z has recently become a tool to probe large-scale clustering (e.g., Abbott et al., 2022), measure the galaxy bias and cosmological parameters and study properties of dark energy (e.g., Abbott et al., 2018).

All photo-z techniques have in common the aim to build a map between the colour (or flux) space and the redshift one (Salvato et al., 2019; Newman & Gruen, 2022). When the map is ready we can obtain by comparison the redshift of a source and the redshift probability distribution function, hereafter redshift distribution. The photo-z techniques can be divided into two macro groups: the physical methods and the data-driven methods. The physical, or template-fitting, methods start from a set of theoretical or empirical SED templates. Theoretical templates are built from stellar emission models, empirical ones are based on observed spectra sometimes extended over broader wavelength ranges. Then, all the physical processes light undergoes travelling from the source to the observer are taken into account to build the colour-redshift map. For example, two processes the accounting of which greatly improves the photo-z measurements, are nebular emission lines and dust absorption and extinction. Most of all dust extinction needs to be modelled because it reddens light and normally it is most efficient in the ultraviolet (UV) band. The galaxy rest frame UV part of the SED, where the Lyman break lies, is what optical and NIR data observe for galaxies with $z > 1$. Extinction reddens the light signal we observe and may cause an overestimation of the galaxy redshift if it is not taken into account. Physical methods usually model dust as a free parameter, using both interstellar attenuation laws (e.g., Calzetti et al., 2000) and intergalactic ones (e.g., Madau, 1995).

The most popular data-driven methods are machine learning algorithms. Starting from a sample of data an ML algorithm learns the colour-redshift map during the training. ML algorithms can be supervised or unsupervised learning methods (see Chapt. 2). Supervised learning needs labelled data, therefore galaxy redshifts have to be known as well as photometry, during the training, while unsupervised learning needs only photometry in this phase. Supervised learning algorithms aim to approximate the function between the multi-dimensional photometry space and the redshift space starting from the training data. At the end of the training, the algorithm has built a function that maps each point in flux space to a redshift, ideally, every galaxy has a different redshift from the others. Therefore, to ensure a good interpolation of the mapping function the sampling data set must be representative of the properties of the sample for which prediction will be made, otherwise, accuracy will be lost. Moreover, supervised ML methods are by nature limited to low redshift and bright objects because the redshift values used during the training are obtained from spectroscopic redshift measurements. Two of the most common supervised learning algorithms are random forest and neural networks (see Sect. 2.1). On the other hand, unsupervised learning methods need only photometry for their training. Rather than a detailed approximation of the mapping function, as supervised methods do, they aim to roughly describe this function assigning different redshift values to galaxies that have been grouped. An unsupervised algorithm constructs the flux-redshift map in two steps. Firstly the training galaxies are divided into groups based on their properties (e.g., k-means; MacQueen, 1967), usually their colours, then, when the groups are fixed, another training sample of galaxies with known redshift is used to estimate the redshift of each group. Therefore, an unsupervised learning method assigns a redshift value to each group of galaxies, not to each galaxy, and it is as good as the groups it builds are compact in redshift space. Possibly the most popular

unsupervised ML algorithm for photo-z measurements is the self-organising map (SOM; Masters et al., 2015; Wilson et al., 2020).

CHAPTER 4

Augmenting photometric redshift estimates using spectroscopic nearest neighbours

*The present chapter is based on the paper 'Augmenting photometric redshift estimates using spectroscopic nearest neighbours' by Federico Tosone, **Marina S. Cagliari**, Luigi Guzzo, Benjamin R. Granett, and Andrea Crespi, published in Astronomy & Astrophysics in April 2023 (Tosone et al., 2023).*

4.1 Introduction

Knowledge of galaxy distances is of uttermost importance for cosmology, as to reconstruct the underlying three-dimensional dark matter distribution that encapsulates key information about the evolution and matter content of the Universe. On cosmological scales, the most efficient method to estimate distances is through their cosmological redshift, which directly connects to the standard definitions of distance. Sufficiently precise redshift measurements allow us to test the world model through the redshift-distance relation, coupled with standard rulers and standard candles (e.g., Riess et al., 1998; Perlmutter et al., 1998).

Over the past 25 years, galaxy clustering measurements from large *redshift surveys* have been able to quantify the universal expansion and growth histories, pinpointing the value of cosmological parameters to high precision (e.g. Tegmark et al., 2006; Colless et al., 2003; Blake et al., 2011; de la Torre et al., 2017; Alam et al., 2017; Pezzotta et al., 2017; Bautista et al., 2021). Even larger redshift surveys are now ongoing (DESI; DESI Collaboration et al., 2016), or soon to start (*Euclid*; Laureijs et al., 2011), with the goal of further refining these measurements to exquisite precision and find clues on the poorly understood ingredients of the remarkably successful standard model of cosmology.

The redshift is measured from the shift in the position of emission and absorption features identified in galaxy spectra, typically through cross-correlation techniques with reference templates, which capture the full information available (e.g., Tonry & Davis, 1979). Despite the considerable advances of multi-object spectrographs over the past 40 years, collecting spectra for large samples of galaxies, however, remains an expensive task. A cheaper, lower-precision alternative is offered by photometric estimates, i.e., measurements based on multi-band imaging, in which integrated low-resolution spectral information is collected at once, for large numbers of objects over large areas. The price to be paid is that of larger measurement errors, together with a number of catastrophic failures, which limit the scientific usage of such photometric redshifts to specific applications (e.g., Newman & Gruen, 2022). Still, when a sufficient number of photometric bands is available, (Benitez et al., 2014; Laigle et al., 2016; Alarcon et al., 2021), or even information about the ensemble mean spectrum can be obtained (Cagliari et al.,

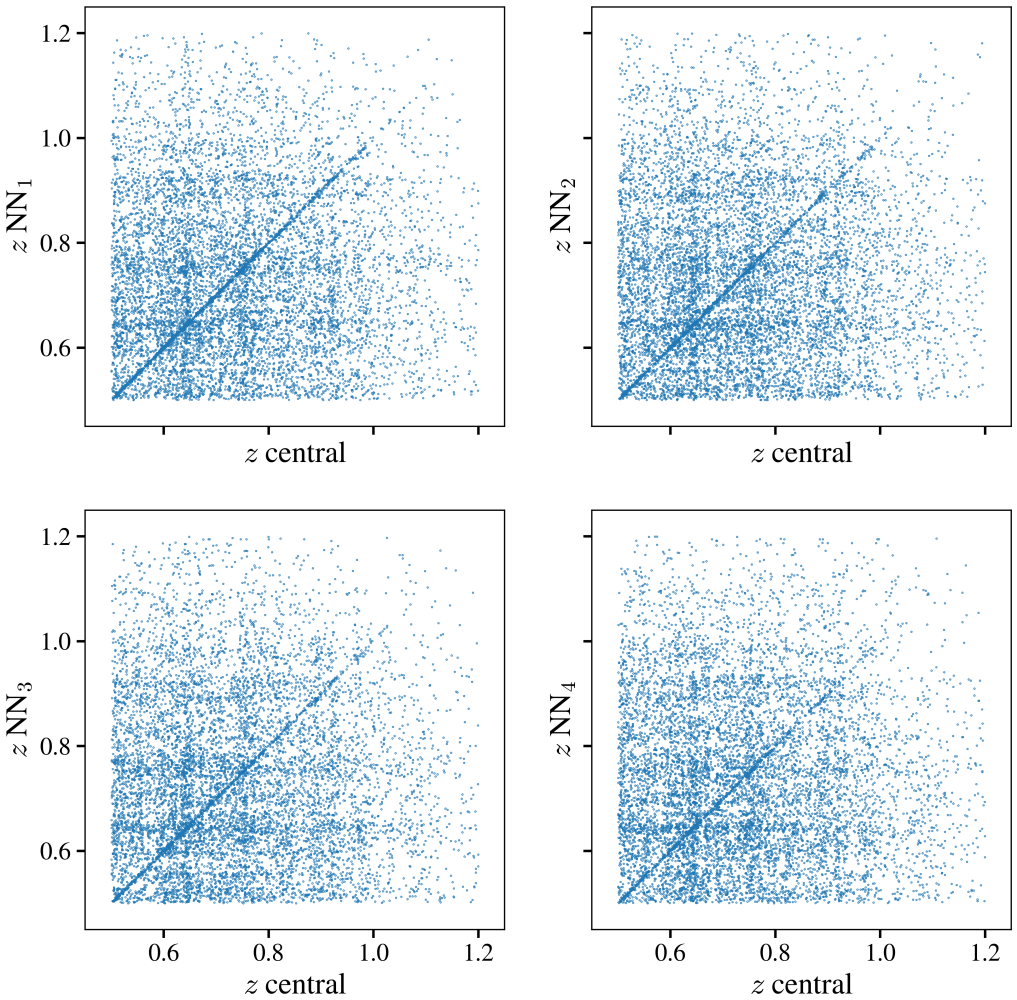

Figure 4.1: Correlation between a galaxy's own redshift and that of its n^{th} nearest angular neighbour ($n = \{1, 2, 3, 4\}$), as seen in the VIPERS redshift survey data, which cover the range $0.5 < z < 1.2$. Clearly, while a tight correlation exists for a number of objects, many other angular pairs just correspond to chance superpositions.

2022), these samples become highly valuable in many respects. Photo-zs are traditionally estimated by fitting template spectral energy distributions to the measured photometric fluxes (e.g., Bolzonella et al., 2000; Arnouts et al., 2002; Maraston, 2005; Ilbert et al., 2006). Detailed reviews can be found in Salvato et al. (2019), Brescia et al. (2021), and Newman & Gruen (2022).

Since the pioneering work of Collister & Lahav (2004, see also Lahav 1994), who first used artificial neural networks (ANN) to obtain photo-z estimates, machine learning algorithms have seen many further applications in this context. These include *random forests* (Carliles et al., 2010), *self-organising maps* (Masters et al., 2015), and advanced ANNs (Sadeh et al., 2016). A notable recent application uses the full images of galaxies through convolutional neural networks (Pasquet et al., 2019; Henghes et al., 2022). All these methods provide photo-z estimates by using information that is strictly local, i.e., the flux of each object measured in a number of photometric bands, independently of correlations with the other galaxies in the sample.

In the specific case when a photometric survey includes spectroscopic redshifts for a representative sub-sample spread over the same area, these represent extra information, which can be exploited to obtain improved estimates of the missing redshifts. Since galaxies are spatially clustered, angular neighbours on the sky preserve a degree of redshift correlation, depending on the depth of the catalogue. The deeper the catalogue, the weaker the correlation, due to projection over a deeper baseline. Still, an angular correlation remains, as can be seen explicitly in Fig. 4.1, in the data of the VIMOS Public Extragalactic Redshift Survey (Guzzo et al., 2014).

Such a correlation was exploited, for example, to improve knowledge of the sample overall redshift distribution (Newman, 2008), a fundamental quantity for many cosmological investigations, as, e.g., weak lensing tomography. With VIPERS, instead, it was used in the estimate of the galaxy density field, to fill the gaps due to missing redshifts (Cucciati et al., 2014). Even more finely, Aragon-Calvo et al. (2015) used the fact that galaxies are typically confined within cosmic web structures to get a dramatic improvement in the estimate of photo-zs for $\sim$ 200 million Sloan Digital Sky Survey galaxies, starting from only $\sim$ 1 million spectroscopically measured redshifts.

Our goal with the work presented here has been to optimally retrieve such non-local information from the neighbouring objects of a given galaxy building upon a specific class of ML architectures, graph neural networks. The key property of this class is the ability to combine information from unstructured data, based on our priors of the task at hand (Bronstein et al., 2017). The end goal is to obtain an improved estimate of the galaxy redshift.

As shown by Fig. 4.1, the existing correlation between angular neighbours is strongly diluted by the sea of chance superpositions along the line of sight. Thus, the problem can be more appropriately recast into quantifying the probability that a given angular neighbour (with known redshift) is a physical companion for a given galaxy and thus closely correlated in redshift as well. Our GNN model, dubbed *NezNet*, combines the intrinsic features of a *target* galaxy and a *neighbour*, i.e., their multi-band fluxes, the spectroscopic redshift of the neighbour and their relative angular distance, to output the probability for the two galaxies to be spatially correlated. We train and test NezNet using the spectroscopic sample of VIPERS. We show that discarding those targets for which no real physical neighbour is identified with significant probability, improves the quality of the associated photo-z catalogue obtained through classic SED-fitting, increasing precision and accuracy and reducing the fraction of catastrophic outliers. Moreover, when real neighbours are identified, the redshift of the highest-probability one represents an estimate for the redshift of the target that is typically more precise than that obtained

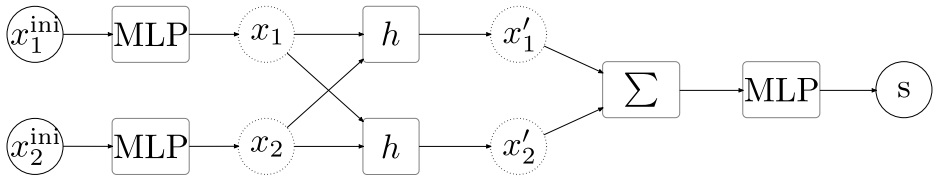

Figure 4.2: Schematic architecture of NezNet: the input features are first processed by a dense network; afterwards, message passing between the two layers through Eq. (4.1) is applied, to take into account both relative differences and global values of the features as well. Before the final dense layer, the features are summed and then reprocessed with an MLP to output the score probability of two galaxies being actual neighbours.

through the classical SED fitting.

The idea of using GNNs to draw extra redshift information from neighbouring galaxies is not totally new. Beck & Sadowski (2019) present preliminary results of an approach based on using only the photometry of a neighbourhood of galaxies, obtaining a 10% improvement on the median absolute deviation of the photo-zs estimated via a single object-based ML algorithm. We believe that the main shortcoming of methods based on apparent neighbours lies in the large fraction of chance superpositions, evident in Fig. 4.1. Here, we reformulate the problem as a detection task that identifies the physical neighbours of the surrounding spectroscopic objects, including also the neighbour's spectroscopic information, obtaining in this way a significant improvement.

The chapter is organised as follows. In Sect. 4.2 we give a brief description of how GNNs work and specify the architecture of our model. In Sect. 4.3 we describe the properties of VIPERS data and the way we prepare the training set, in particular how we define real or apparent neighbouring objects. Section 4.4 describes how the model is applied to the data and the metrics that we use to quantify the performance of the results. Finally, in Sect. 4.5 we present and discuss our results, and conclude in Sect. 4.6.

4.2 Model

A neural network model can be summarised as a set of nonlinear functions applied to a set of inputs which undergo a linear mapping. Each mapping has many parameters that are optimised through a training process, which allows the network model to approximate a wide variety of almost arbitrary functions (LeCun et al., 2015). In its simplest form, a neural network model corresponds to a multi-layer perceptron (MLP), also known as a dense neural network (Murtagh, 1991). If one is dealing with images, neural architectures such as CNN are more suited, as they take into account our a priori knowledge about the data structure (O'Shea & Nash, 2015).

This reasoning can be pushed further by introducing neural networks for graph representations (Zhou et al., 2018). In this work, we make use of one key aspect of GNN, i.e., message passing (Gilmer et al., 2017). To fix ideas, the problem we want to address is the following: we need to find the spectroscopic galaxies with the highest probability of being close to a galaxy for which only photometric information is available. This can be recast as a classification task for each pair of galaxies, in which our aim is to distinguish between apparent and real neighbours when projected on the plane of the sky.

Intuitively, a model to distinguish between apparent and real neighbours should be based on the relative difference between galaxy features. Such a neural network can be

designed by including a layer of the form

$$\boldsymbol{x}_i' = \sum_{j \in \mathcal{N}(i)} h(\boldsymbol{x}_i, \boldsymbol{x}_i - \boldsymbol{x}_j)\,, \tag{4.1}$$

where $\boldsymbol{x}_i$ refers to the array of input features of the node i-th, $\mathcal{N}(i)$ is the neighbourhood of the same node, $\sum$ is the aggregation function which sums the outcomes from each pair of nodes. The function h is an MLP that explicitly combines the value of the input feature at the node and the relative difference of that feature with respect to the neighbour. It is worth noting that such a GNN is both permutation equivariant and permutation invariant, so that it is not affected by changing the order of the nodes, i.e. the input galaxies.

The complete architecture of our model is illustrated in Fig. 4.2. Each node is a galaxy, whose inputs (e.g. the photometric measurements) are pre-processed through a MLP, before undergoing the message passing of Eq. (4.1). In our work, we restrict ourselves to the case of pairs of galaxies, so that the neighbourhood $\mathcal{N}(j)$ includes one galaxy only, and the aggregation function simply sums the features $\boldsymbol{x}_1' + \boldsymbol{x}_2'$. This model can be seen as a trivial version of EdgeConv (Wang et al., 2018), where the adjacency matrix is a 2×2 matrix, with 0 entries for diagonal elements and 1 for the off-diagonal elements. Finally, the summed features undergo a last dense layer with a scalar output. All the activation functions are rectified linear units, with the exception of the final layer where we use a sigmoid, as to represent a probability for our classification task.

We dub this classification model Nearest-z Network (NezNet). NezNet provides the probability for a pair of galaxies to be real neighbours. The loss function adopted to train NezNet is a standard binary cross-entropy

$$\mathcal{L} = \frac{1}{n} \sum_{i}^{n} \left[y_i \log p_i + (1 - y_i) \log (1 - p_i) \right], \tag{4.2}$$

where p_i is the output probability of NezNet for each pair of galaxies, while $y_i = 0, 1$ is the corresponding training label, and the sum is averaged over the mini-batch. To design our model we made use of the Spektral library (Grattarola & Alippi, 2020), where the EdgeConv layer is conveniently already implemented.[1]

4.3 Data

We train and test our approach on the final data release of VIPERS (Guzzo et al., 2014; Scodeggio et al., 2018), for which the redshift correlation between angular neighbours has been in Fig. 4.1. The survey used the VIMOS multi-object spectrograph at the ESO Very Large Telescope to target galaxies brighter than $i_{\rm AB} = 22.5$ in the Canada-France-Hawaii Telescope Legacy Survey Wide (CFHTLS-Wide) catalogue, with an additional $(r - i)$ vs $(u - g)$ colour pre-selection to remove objects at $z < 0.5$. The resulting sample covers the redshift range $0.5 \lesssim z \lesssim 1.2$, with an effective sky coverage of $16.3\,\mathrm{deg}^2$, split over the W1 and W4 fields of CFHTLS-Wide. We used only galaxies with secure redshift measurements, as identified by their quality flag, corresponding to a 96.1% confidence level (see Scodeggio et al. 2018).

For each galaxy in the catalogue the following information is considered:

- the spectroscopic redshift measurement $z_{\rm spec}$,

[1] https://graphneural.network

- the 6 magnitudes u, g, r, i, z (not to be confused with redshift) and K_s,
- right ascension α (RA), in radians,
- declination δ (Dec), in radians.

The angular separation on the sky between two objects with RA α_1 and α_2, and Dec δ_1 and δ_2, is given by the haversine formula

$$\Delta\Theta = \arccos\left(\sin\delta_1 \sin\delta_2 + \cos\delta_1 \cos\delta_2 \cos\left(\alpha_1 - \alpha_2\right)\right). \tag{4.3}$$

We select the parent photometric sample by applying the same VIPERS colour and magnitude cuts defined above, to be fully coherent with the spectroscopic data.

4.4 Application

We set up a training set from the VIPERS W1 galaxy catalogue: we randomly select about 3×10^4 target galaxies, whose spectroscopic redshift during training is ignored. For each of them, we identify the first n_NN angular nearest neighbours as defined by Eq. (4.3), which we dub *spectroscopic galaxies*, since their spectroscopic redshift information is used in our model. Each of these spectroscopic neighbours is associated with the same target galaxy, but the pairs can be considered independent from one another in our model. Each angular pair is assigned label 1 if it is a real physical pair, otherwise, it is assigned a 0. The training set is thus made of pairs of galaxies.

A target galaxy of a pair can also be the nearest neighbour of another target galaxy, in another pair. We make this choice in order to maximise the number of training examples available in W1. Our final tests on the W4 catalogue show that this does not lead to any over-fitting of VIPERS data, as the model generalises well. We note that this setting assumes a ratio of spectroscopic to photometric objects of $1 : 1$. In the Conclusions (Sect. 4.6) we also confirm these results in the more realistic case where the number of spectroscopic redshifts used for training is a fraction of the number of photometric objects.

The definition of a real neighbour is arbitrary; it is reasonable to consider that two angular neighbours form a physical pair if their spectroscopic separation is smaller than a given threshold

$$\Delta z\,(1 + z_\mathrm{spec}). \tag{4.4}$$

This means that in setting up the training data there are two hyper-parameters, the number of nearest neighbours n_NN to be considered, and the spectroscopic separation Δz. As we will show, these two hyper-parameters can affect the results significantly, and it is thus relevant to set them up wisely, depending on the specific survey.

In NezNet, for each galaxy in the pairs, the input features of the nodes are the photometry, the spectroscopy and the angular position, listed in Sect. 4.3. For the target galaxy, we always set $z_\mathrm{spec} = 0$, so to have the model consider it as a missing feature, while providing its value for the neighbouring galaxy. Magnitudes are normalised to the range $[0, 1]$, as computed over the whole VIPERS dataset. The angular inputs are provided in terms of relative distance with respect to the target galaxy, so that $\Delta\Theta = 0$ for the latter, while for the neighbour it corresponds to Eq. (4.3). By adopting this choice we guarantee that the model has translational invariance.

Another tested option (see Sect. 4.6), is to use as input variables the relative distance in the two sky coordinates RA and Dec, rather than the angular separation of the two galaxies. This choice is due to the fact that the surface distribution of the sample is not

rotationally invariant on the sky, due to the technical set-up of the slits in the VIMOS focal plane, with the spectral dispersion oriented along the declination direction. As spectra must not overlap on the detector, targets need to be separated in Dec much more than in RA. As a result, the minimum separation is ~ 1.9 arcmin in Dec and 5 arcsec in RA. More details can be found in Bottini et al. (2005) and Pezzotta et al. (2017, cf. their Sect. 4.1). As a matter of fact, our experiments show that providing the model with the angular separation $\Delta\Theta$ introduces a bias in the redshift metrics, which is not observed when the relative separations along RA and Dec are given. In general, however, we find that the separation information does not significantly improve the classifier and, for this reason, we do not use it in our final model. Rather, spatial information comes only from the number of nearest neighbours considered.

The other hyper-parameters of the model, i.e. batch size, number of neurons and learning rate, have a much lesser impact than Δz and $n_{\rm NN}$ and have been set to fiducial values: a batch size of 32, a learning rate of 0.001, and a total number of parameters of the order of a few thousand. We find little difference in the output metrics of the redshift estimates when increasing the complexity of the model, or changing the batch size and the learning rate around these fiducial values.

NezNet gives in output the probability for two galaxies to be real neighbours. As each target galaxy corresponds to $n_{\rm NN}$ independent pairs, we can select the neighbour with the highest probability among them. If this probability is below the classification threshold set to define a positive case, we conclude that in the catalogue there is no physical neighbour for that target galaxy. This implies that the latter is to high probability an outlier in terms of its properties, when compared to its neighbours. Removing such objects from the final catalogue significantly improves the metrics when comparing photo-z and spectroscopic measurements. In particular, the reduction in the number of catastrophic redshifts confirms our assumption. Finding a true neighbour, instead, reinforces the confidence in the photo-z. At the same time, in this case, the spectroscopic redshift of the neighbour is typically an even better estimate of the target redshift, compared to the SED-estimated photo-z. These tests are discussed in the following section.

The quantitative comparison between NezNet results, spectroscopic measurements $z^{(i)}_{\rm spec}$ and SED-fitting estimated photo-zs is performed using the metrics defined in Salvato et al. (2019). These are the precision, (i.e., the dispersion of the estimated values),

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(\frac{z^{(i)}_{\rm spec} - z^{(i)}}{1 + z^{(i)}_{\rm spec}}\right)^2}\,, \tag{4.5}$$

the bias

$$b = \frac{1}{N}\sum_{i}^{N}(z^{(i)}_{\rm spec} - z^{(i)})\,, \tag{4.6}$$

and the absolute bias

$$|b| = \frac{1}{N}\sum_{i}^{N}|z^{(i)}_{\rm spec} - z^{(i)}|\,, \tag{4.7}$$

quantifying systematic deviations. Finally, the outliers are defined as those objects for which

$$|z^{(i)}_{\rm spec} - z^{(i)}| \geq 0.15(1 + z^{(i)}_{\rm spec})\,. \tag{4.8}$$

All the results presented in the following section have been obtained by applying the trained NezNet to a test catalogue built in a similar fashion to W1, randomly selecting about 2×10^4 galaxies from the twin W4 field of VIPERS.

Finally, in the following discussion about our classifier, we will use the notion of true positive rate (TPR), which is the fraction of correctly predicted positive examples with respect to all the real positive examples, defined as

$$\mathrm{TPR} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}\,, \tag{4.9}$$

where N_{TP} stands for true positives and N_{FN} stands for false negatives. Similarly, we can define the false positive rate (FPR), which is the fraction of negative examples classified as positives with respect to all the real negative examples, which reads

$$\mathrm{FPR} = \frac{N_{\mathrm{FP}}}{N_{\mathrm{FP}} + N_{\mathrm{TN}}}\,, \tag{4.10}$$

where N_{FP} stands for false positives and N_{TN} stands for true negatives.

4.5 Results

As explained in the previous section, NezNet can be used to simply clean a photo-z sample by discarding low-probability neighbours or to provide an alternative redshift estimate derived from the highest probability neighbour. This is demonstrated on the test catalogue in Fig. 4.3, for a model trained using the hyper-parameters $\Delta z = 0.08$ and $n_{\mathrm{NN}} = 30$. In addition to the VIPERS spectroscopic redshifts, this comparison includes also the original photo-zs estimated by Moutard et al. (2016) using standard SED fitting. For these and all following results, angular information (i.e., the separation of the two objects on the sky) was not used as an input variable. The reason for this was already mentioned in the previous section and is discussed again in more detail below.

From Fig. 4.3, we see that by simply dismissing the outliers as identified by NezNet, all the metrics show significant improvements (top-right panel). Also, when the *best neighbour* redshifts are adopted for the target galaxies (bottom panel), we obtain metrics that are comparable or even better than those of the *cleaned* photo-z sample. It is worth noting that in this case the plot shows a characteristic checkerboard pattern due to the reflection of the spectroscopic redshift striping, as spectroscopic redshifts are now assigned to target photometric objects.

Figure 4.3 also shows the limits of the method. Comparing the left panel with the other two, we can notice that NezNet tends to cut off the high redshift tail of the distribution. This is easily understood considering the magnitude-limited ($i_{\mathrm{AB}} < 22.5$) character of the sample used here, which becomes very sparse at $z \gtrsim 1$, where only rare luminous galaxies are present. This means that the model becomes intrinsically less efficient, due to the lower number of real physical neighbours available both for the training and for inference, as also evident from the density of points at high redshift in Fig. 4.1. Devising a different loss function to up-weight the few physical pairs in this regime could perhaps improve the classification task, but an intrinsic limit to the method clearly exists when the density of the sample decreases.

Figure 4.4 shows the same set of plots, but using in the training a larger value for the spectroscopic separation, i.e., $\Delta z = 0.15$. As expected, allowing for a larger separation in the definition of real angular neighbours discards a smaller fraction of data. Conversely, there is in general a lower precision and a small increase in the fraction of outliers.

In principle, using a stricter Δz could remove even more outliers, retaining only pairs that are closer in redshift and leading to a smaller, but more precise sub-sample. We explore this dependence in Fig. 4.5. Overall, this method is always able to clean bad

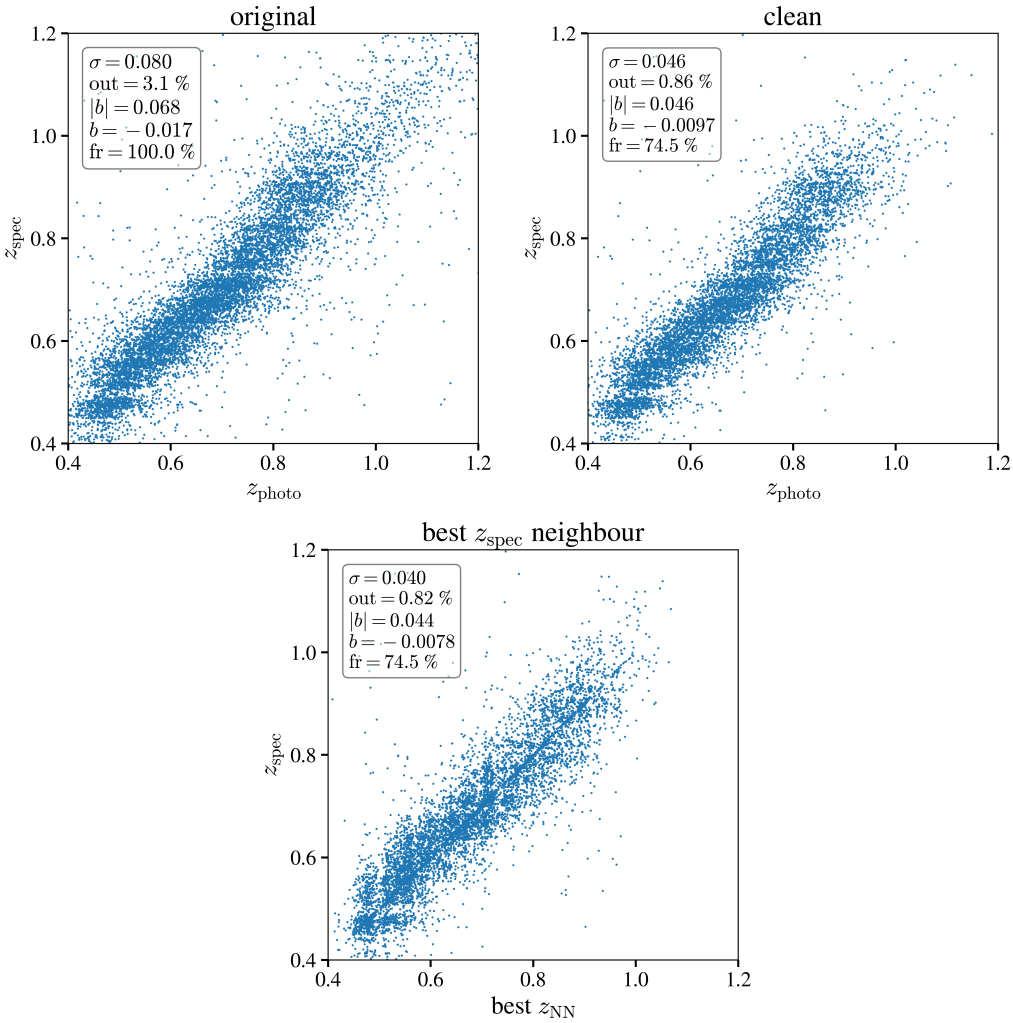

Figure 4.3: The top-left panel shows the distribution of photometric vs spectroscopic estimates in the original data. In the top-right panel, we show the same distribution after removing from the catalogue the galaxies with low score probability (*fr* stands for the fraction of retained data). Finally, the bottom panel shows estimates of redshift by assigning to the target galaxy the spectroscopic redshift of the neighbour with the highest detection probability. The model was trained with $n_{\rm NN} = 30$ and $\Delta z = 0.08$.

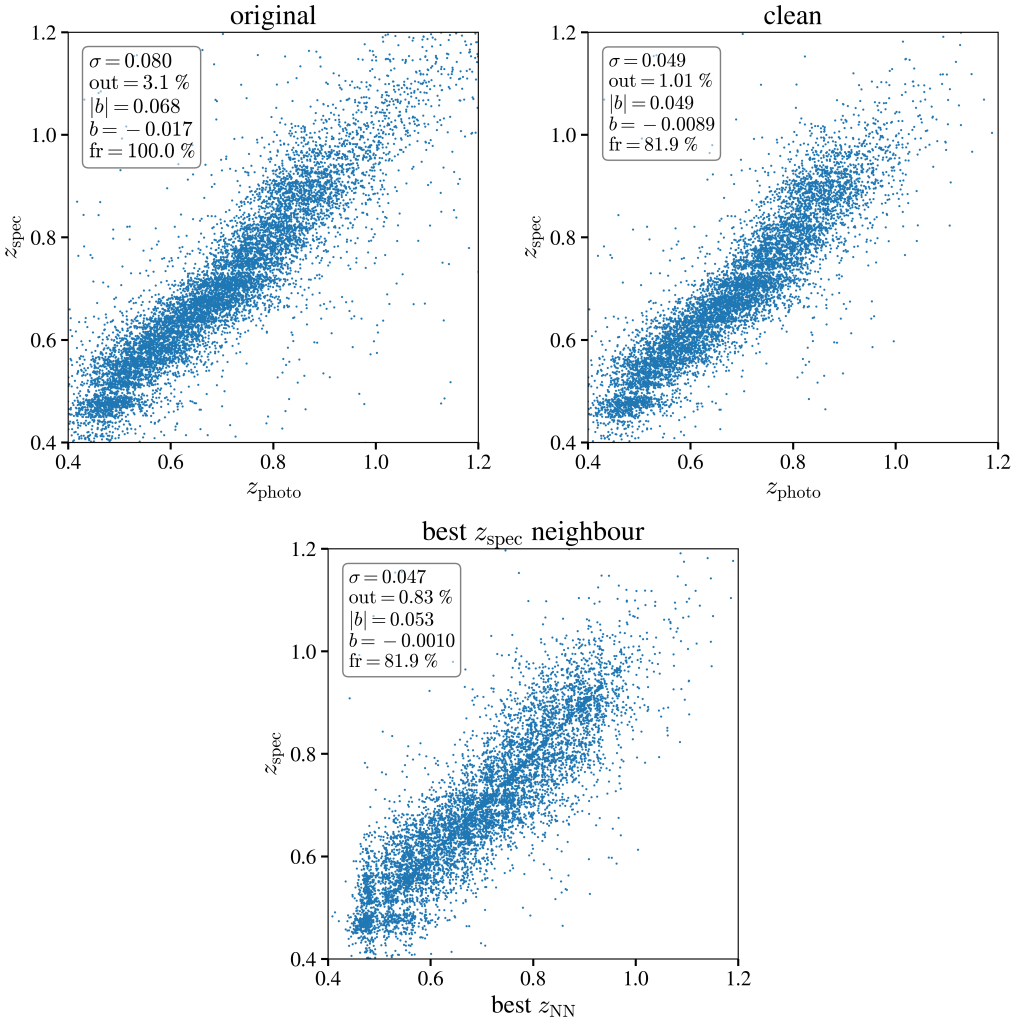

Figure 4.4: Same as Fig. 4.3, but the model was trained with the higher $\Delta z = 0.15$, while $n_{\mathrm{NN}} = 30$ is the same as before. As we can see, increasing the error which defines a neighbour retains more data points, but the precision decreases slightly.

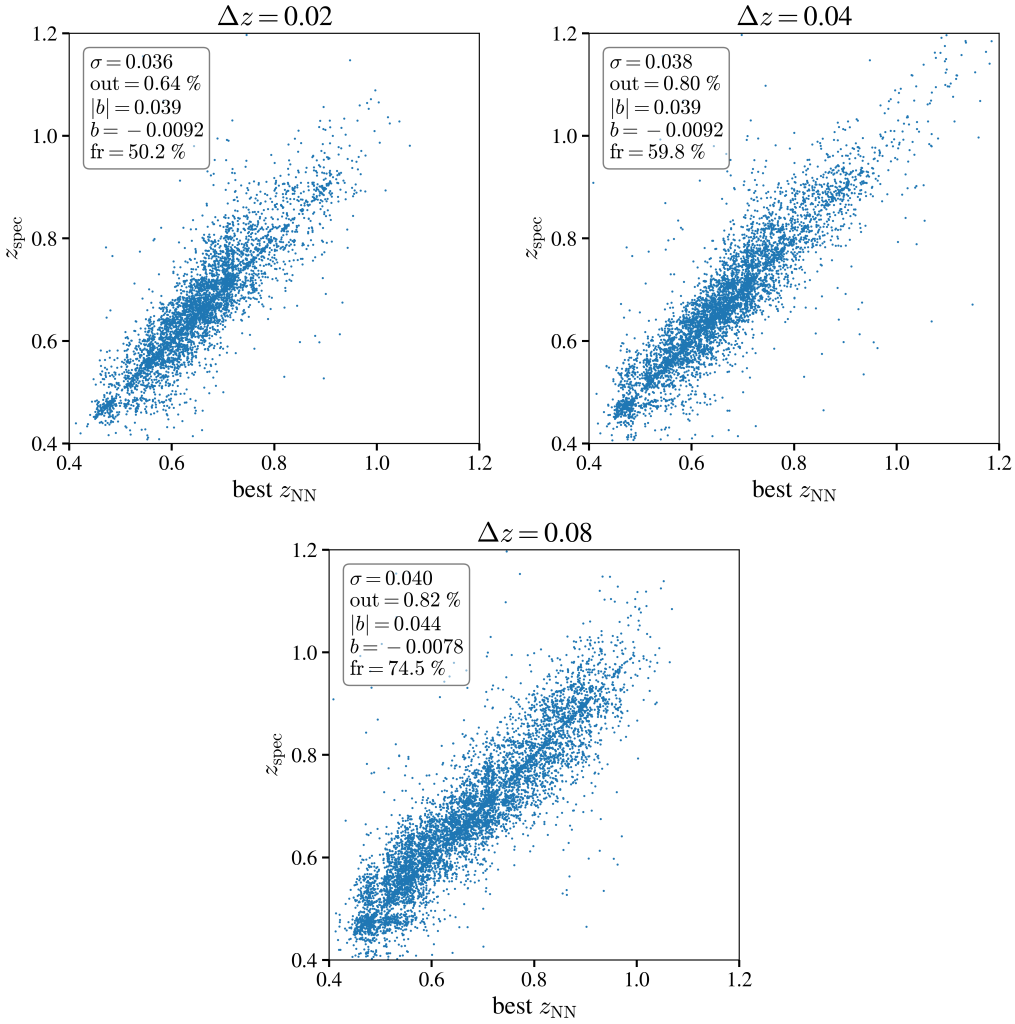

Figure 4.5: Redshift estimates derived from the best nearest neighbour, for various Δz, at fixed $n_{\rm NN} = 30$. Increasing the spectroscopic separation to define physical neighbours, while diminishing the quality of the metrics, increases the fraction of data not dismissed from the catalogue.

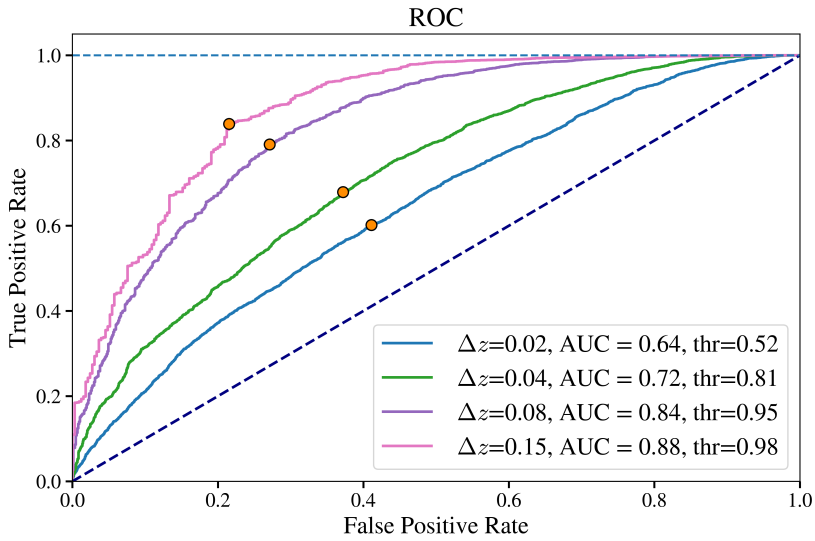

Figure 4.6: The ROC curve varying the redshift threshold Δz, at fixed $n_{\rm NN} = 30$. The performance of our classifier (AUC) improves as we use a less strict definition of what we define as a true neighbour. The probability that an angular neighbour is a physical neighbour increases at larger Δz, which is also reflected by the high detection threshold (thr).

estimates off the sample, but at the price of discarding many data points. The minor improvement in precision probably does not justify the use of $\Delta z < 0.08$ in the case of VIPERS, because more than half of the sample is excluded.

It is apparent that the hyper-parameter Δz is very relevant for the quality of the classifier. This is made clear by the receiver operator characteristic (ROC) curve in Fig. 4.6, which shows the TPR (Eq. 4.9) against the FPR (Eq. 4.10), and has been computed from the target galaxies in the test catalogue by considering their neighbour with the highest probability. In general, the area under the curve (AUC) is higher for the better classifier. Increasing Δz increases the AUC, which would tend to unity for very large values of this parameter, as all galaxies would then be considered real neighbours. However, our ultimate goal is not to increase the performance of the classifier per se, but to improve the metrics of our redshift estimates. These show that $\Delta z \gtrsim 0.08$ represents the best choice for VIPERS.

The other hyper-parameter of NezNet, i.e., $n_{\rm NN}$, the number of nearest neighbours considered in the training, has a lesser impact on the classifier. We show this in Fig. 4.7, where each ROC curve corresponds to a model trained with a different $n_{\rm NN}$, but all with the same Δz. Changing drastically $n_{\rm NN}$ does not correspond to comparable changes in the AUC. However, $n_{\rm NN}$ has a large impact on the redshift estimates, as Fig. 4.8 makes apparent. Considering a larger number of angular neighbours increases the probability of finding a physical pair, as can be seen from the metrics in Fig. 4.7. We also experimented with raising the value of $n_{\rm NN}$ up to 50, but found no further gain with respect to using $n_{\rm NN} = 30$. Already above $n_{\rm NN} = 10$ the redshift metrics start to saturate to the optimal values.

We also computed, as a further test, the gradients of the predictions with respect to

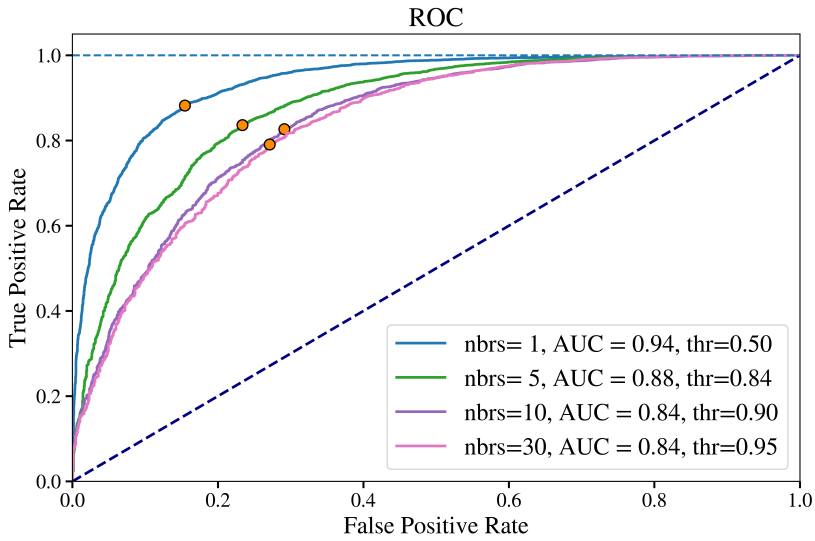

Figure 4.7: The ROC curve varying the number of nearest neighbours $n_{\mathrm{NN}} = 30$, at fixed $\Delta z = 0.08$. Increasing the number of neighbours that are given in input to the training seems to make the training more difficult. However, this test of the classifier does not reflect the quality of the final redshift estimate, as Fig. 4.8 shows.

their input variables, to detect the most relevant ones, as shown in Fig. 4.9. It is interesting to see that the neighbour's redshift is a relevant input, as one would expect, and some of the photometric bands are even more relevant. This confirms the intuition that the photometric information of the neighbours does indeed provide additional information about the relative distance from the target. In this plot, we also show results for the case when the angular separation is considered as one of the input variables. These show that the angular separation $\Delta\Theta$ between the target and the neighbour does affect the predictions. This manifests itself as a bias in the redshift estimates, as visible in Fig. 4.10: in this case NezNet systematically favours neighbours that are closer to us than the target, increasing the value of the bias b (Eq. 4.6). We also tested what happens if the angular separation information is rather given in terms of the relative difference in the angular coordinates RA and Dec of the two galaxies. In this case, the bias disappears and the results are comparable to the standard case in which no angle information is provided. However, we see that in this case the two parameters have smaller gradients than when $\Delta\Theta$ alone is considered, which suggests they are in fact not contributing to the predicting power of the model. For these reasons, in our final results, the angular separation is not considered as input variable.

One of the novelties of NezNet is the message passing between node features. This is where GNNs differ from a standard ANN, where all input variables of both galaxies would be provided directly to dense layers. We also experimented with a simpler graph model, closely resembling the architecture of NezNet, but without message passing. The input features are processed independently by MLP layers for each node (we tried using either just one or several layers). The new architecture is as in Fig. 4.2, with the exception of **h** function blocks which are now substituted with new MLP blocks, without applying any message passing. The $\boldsymbol{x}'_i$ features are summed by the aggregation function, and the

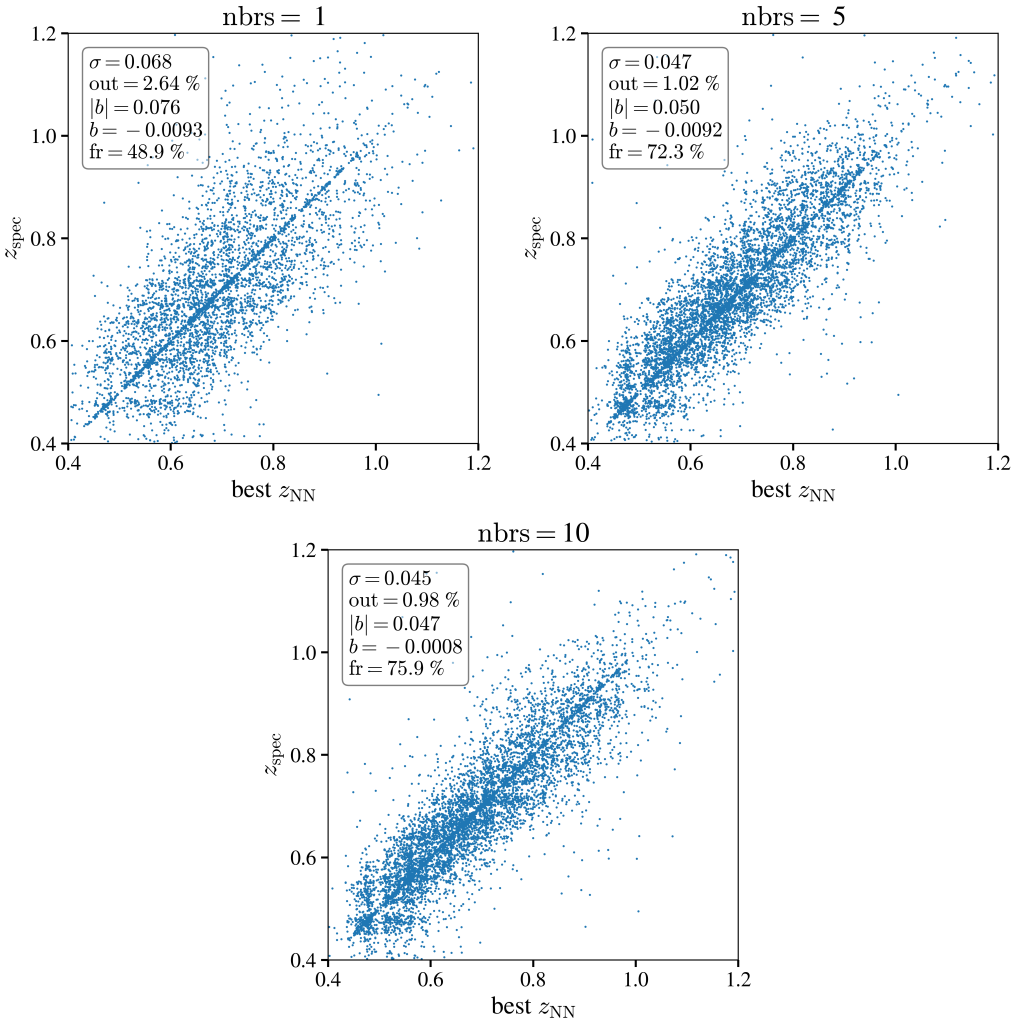

Figure 4.8: Redshift estimates based on the best nearest neighbour, for various n_{NN}, at fixed $\Delta z = 0.08$. Increasing the number of nearest neighbours for each target improves the performance of NezNet in estimating redshifts, as it increases the probability that physical pairs are considered. .

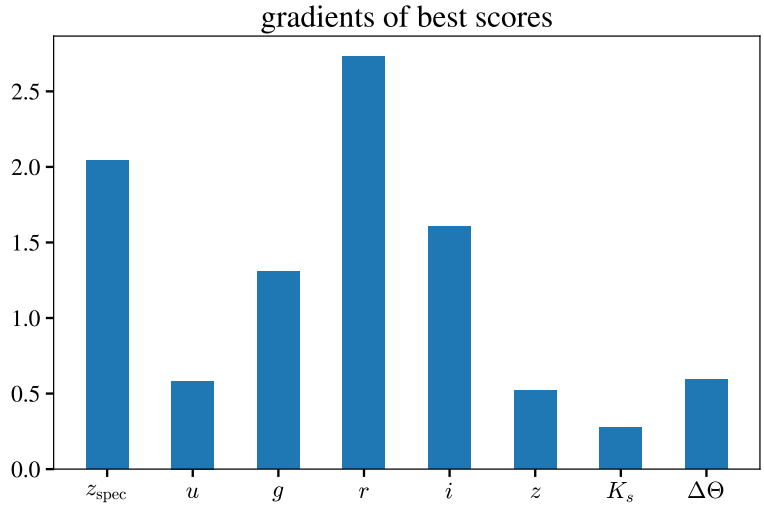

Figure 4.9: Average absolute values of the gradients of NezNet with respect to the input features of the neighbours. For each target, we only considered the neighbour with the highest probability. Despite the angular separation $\Delta\Theta$ can be a relevant input, we do not use it in our final results, because of the bias it introduces, apparent in Fig. 4.10.

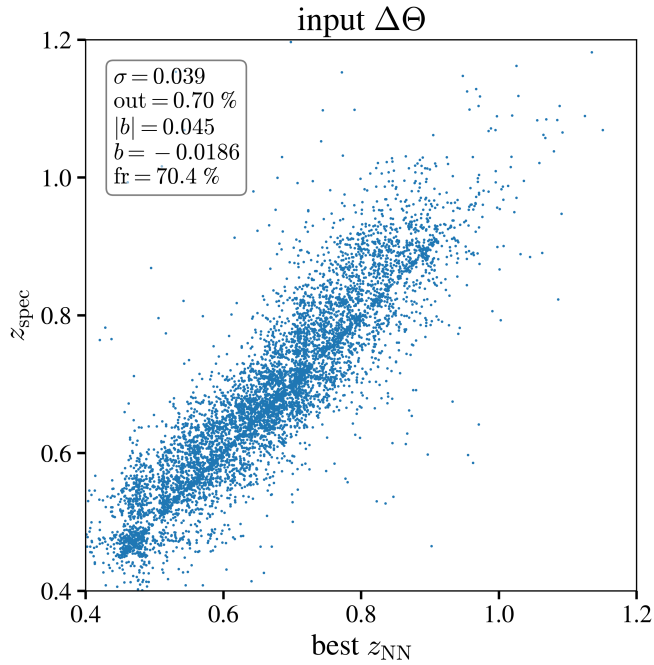

Figure 4.10: Results of redshift estimates for the target galaxies, in the case where the angular separation Eq. (4.3) is an explicit input of the model. We can see that many galaxies have slightly smaller values than the real spectroscopic value, resulting in a large bias b. Currently, we do not have an explanation of this observed effect.

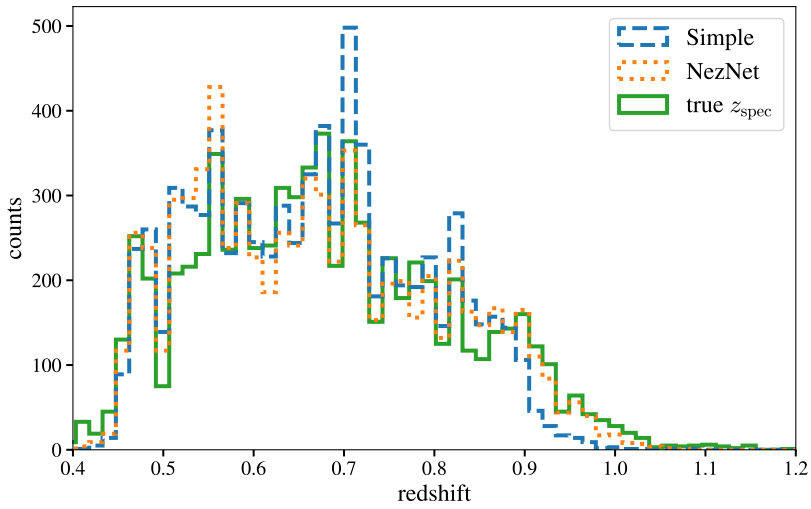

Figure 4.11: Comparison of the redshift distribution for the predictions of NezNet, and a simpler graph model without message passing. While the latter performs reasonably well in general, it tends to cut the tail of the distribution.

summed features are mapped to the output probability through final dense layers with sigmoid activation output, just like in the model with message passing. This kind of model, which maintains the permutation invariance property of a graph, is often referred to as a *deep set* (Zaheer et al., 2017). We find that this simple model still works remarkably well and is in general comparable to NezNet. However, it systematically cuts off the high-redshift tail of the catalogue (Fig. 4.11), despite the overall metrics remaining good.

4.6 Conclusions

We have presented a new ML model, dubbed NezNet, which for a pair of galaxies takes in input their measured fluxes in a number of bands, together with the redshift of one of the two. NezNet is capable of learning probabilistically whether their redshift distance is below a given threshold Δz, which is set as a hyper-parameter of the model. The angular separation between the galaxies is implicit in the training set, as for every target galaxy we select its first $n_{\rm NN}$ angular neighbours (another hyper-parameter), but it can be an explicit input variable of the model. The backbone of the model is a GNN, a class of neural networks based on message passing and the aggregation of features (Fig. 4.2). This message passing is explicitly performed as a relative difference between features (Eq. 4.1).

NezNet outputs the score probability for a pair of galaxies to be real neighbours, an information that can be used in two ways. On one side, if none of the $n_{\rm NN}$ nearest neighbours is identified as a physical one, the target galaxy can be considered an outlier in terms of its properties. This may suggest it is an interloper, i.e., a foreground or background object with respect to the volume sampled by the spectroscopic sample we are using for the comparison. As such, it should be discarded from any sample that aims at covering the same redshift range of the spectroscopic catalogue, e.g., via photometrically estimated redshifts. We have proved this to be true using the VIPERS catalogue. On the

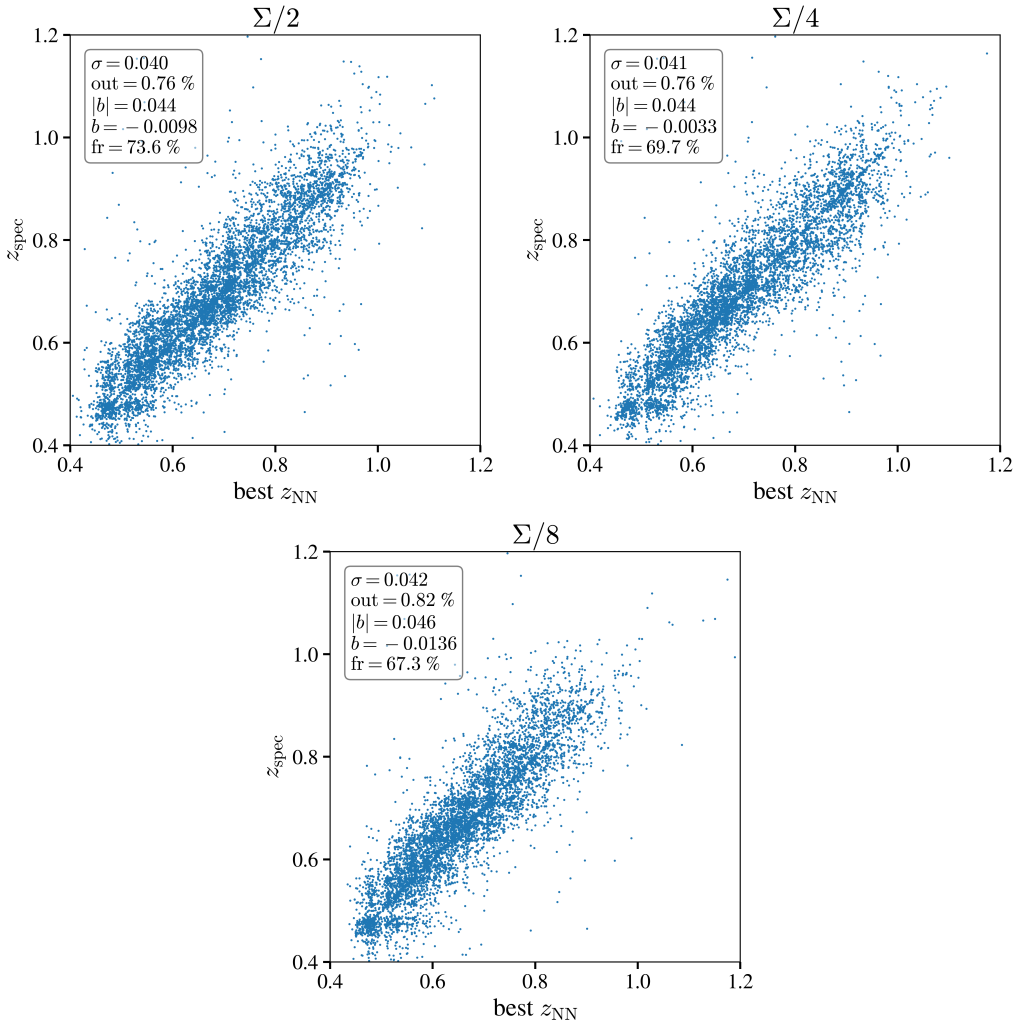

Figure 4.12: Redshift estimates based on the best nearest neighbour, obtained by uniformly subsampling the W1 catalogue, at fixed $n_{\mathrm{NN}} = 30$ and $\Delta z = 0.08$. The titles of the panels refer to the surface density of spectroscopic objects of W1 used for training, with Σ referring to the complete W1 sample. Apart from minor fluctuations in the redshift statistics, we see that NezNet maintains a performance similar to the case without subsampling. The only noticeable trend is the fraction of central galaxies for which a physical pair is found, which decreases for lower densities. This could be due to the decreasing number of training data available. The percentage of real physical neighbours for a central galaxy, which decreases only slightly when going from Σ to $\Sigma/8$, remaining around $40\ \%$, explains why NezNet is still effective.

other side, if a physical neighbour is identified, the target galaxy can be assigned the spectroscopic redshift of the highest scoring galaxy among the $n_{\rm NN}$ angular neighbours, providing in this way an independent estimate of its redshift.

These results are summarised in Fig. 4.3 and Fig. 4.4: by simply discarding outliers as detected by NezNet, all the metrics of the sample improve considerably. Moreover, NezNet's redshift estimates are comparable or superior in precision to SED-based photometric redshifts, depending on the values chosen for the hyper-parameters. Increasing Δz increases the goodness of the classifier (Fig. 4.6), as well as the fraction of retained data (Fig. 4.5). Changing $n_{\rm NN}$ has a smaller impact on the classifier (Fig. 4.7), although it significantly affects the redshift quality metrics, since a large enough $n_{\rm NN}$ improves the probability of detecting a real neighbour; a value $n_{\rm NN} \sim 30$ is optimal in the case of VIPERS (Fig. 4.8).

It is often the case that the fraction of the parent photometric sample without a spectroscopic measurement has a higher density than the spectroscopic sample. Indeed VIPERS has a spectroscopic surface density of $\Sigma \sim 6 \times 10^3\,\mathrm{deg}^{-2}$, to compare against the photometric surface density $\Sigma_{\rm ph} \sim 45 \times 10^3\,\mathrm{deg}^{-2}$. For this reason, we tested NezNet by varying the surface density of the spectroscopic sample used during training. We achieve this by repeating the training procedure on a uniformly subsampled catalogue extracted from W1. The test is performed on W4 without any subsampling, so that we test for the effectiveness of NezNet trained on a lower-density catalogue. Figure 4.12 shows that NezNet keeps its effectiveness even when using a subsample of one eight of the original spectroscopic density Σ, similar to the VIPERS ratio of spectroscopic to photometric objects.

This suggests that NezNet could have an interesting potential also in the context of future experiments, such as *Euclid* or the NASA *Nancy Grace Roman* mission (Akeson et al., 2019). Indeed, such slitless spectroscopic surveys will naturally deliver overlapping photometric and spectroscopic data, which can be combined using NezNet to improve photometric redshift estimates.

It is worth stressing that some details of the results presented here depend on the specific features of VIPERS and its parent CFHTLS photometric sample. Some of these may have been advantageous, but others could have penalised the success of the method. For example, the slit-placement constraints in VIPERS limit the ability to target close pairs of galaxies, which introduces a 'shadow' in the layout of a VIMOS pointing (see Fig. 6 of Guzzo et al., 2014), and forces a lower limit in the separation of observable galaxy pairs (see Sect. 4.4). This means that, in fact, in the present analysis the training sample of NezNet was not ideal, as surely many of the missed angular pairs were also physical pairs. This increases our confidence in the obtained results, as it shows that also for samples characterised by small-scale incompleteness, as typical of surveys built using fibre or multi-slit spectrographs, the method still delivers very useful results. In the case of the VIPERS data, an interesting exercise in this respect would be to use as a training sample the data from the VLT-VIMOS Deep Survey (VVDS) (Le Fèvre et al., 2005), which used the same spectrograph, but with repeated passes over the same area of $0.5\,\mathrm{deg}^2$ that substantially mitigate the proximity bias. We leave this exercise for future work.

CHAPTER 5

Euclid: Testing photometric selection of emission-line galaxy targets

The present chapter is based on the paper in preparation 'Euclid: Testing photometric selection of emission-line galaxy targets' by ***Marina S. Cagliari****, Benjamin R. Granett, Luigi Guzzo, Matthieu Bertermin, Micol Bolzonella, Sylvain de la Torre, Pierluigi Monaco, Michele Moresco, Will J. Percival, Claudia Scarlata, Yun Wang, Meriam Ezziati, Olivier Ilber, Vincent Le Brun et al., the paper will be submitted to Astronomy & Astrophysics on behalf of the Euclid Collaboration.*

5.1 Introduction

The ESA *Euclid* mission will carry out an imaging and spectroscopic survey over one-third of the sky (Laureijs et al., 2011). The imaging channel will enable measurements of cosmic shear providing a tomographic view of the matter distribution, while the spectroscopic redshift survey will map the large-scale structure in three dimensions. Jointly, the two probes will yield unprecedented constraints on the cosmological model (Euclid Collaboration: Blanchard et al., 2020).

The *Euclid* near-infrared spectrograph and photometer (Maciaszek et al., 2022) has three broadband filters for imaging, $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$ (Euclid Collaboration: Schirmer et al., 2022) and a set of grisms for spectroscopy, while the visual instrument (Cropper et al., 2016) images through a single broad pass band, $I_{\rm E}$, spanning the range $[530, 920]$ nm, with high spatial resolution of 0.1 arcsec/pixel. Jointly, these two instruments will carry out the Euclid Wide and Deep Surveys (Euclid Collaboration: Scaramella et al., 2022). The NISP instrument operates as a slitless spectrograph, to record the dispersed light of all sources in the field of view to a nominal emission-line flux limit of 2×10^{-16} erg s^{-1} cm^{-2}, which corresponds to a 3.5σ detection of a 0.5 arcsec diameter source in the Wide survey as designed. The use of slitless spectroscopy makes the spectroscopic survey highly efficient, since individual sources do not need to be targeted; however, reliable redshift measurements will only be secured for a fraction of the galaxies that are detected photometrically. The Wide Survey will detect the most luminous Hα emitters over the redshift range $0.9 < z < 1.8$, with typical broadband flux corresponding to $H_{\rm E} \lesssim 24$; however, it will be sensitive to continuum emission only from the most luminous galaxies and, so, the redshift estimation will be based primarily on the detection of emission lines (Euclid Collaboration: Gabarra et al., 2023). The Wide Survey will be complemented by the Deep Survey, which will reach 2 magnitudes deeper in flux over an area of $50\,\mathrm{deg}^2$ split over three separate fields. In the Deep Survey blue grism ($[926, 1366]$ nm) observations will complement those with the standard red grism

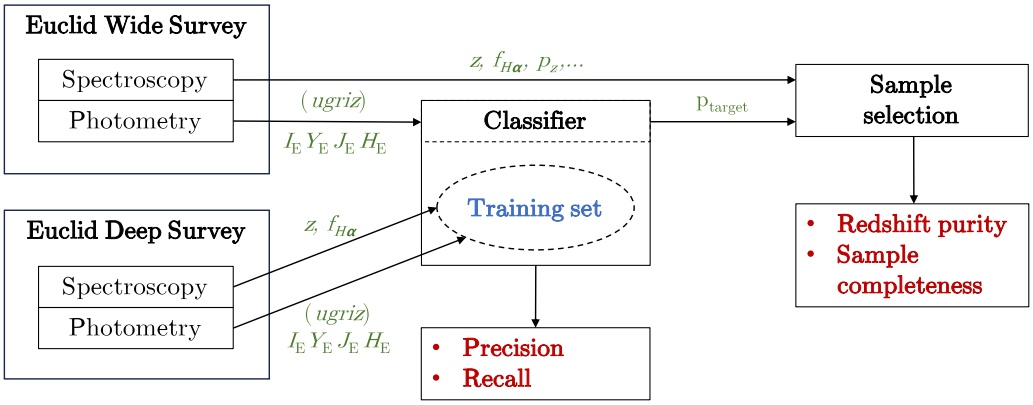

Figure 5.1: Schematic description of the spectroscopic sample selection pipeline. The flowchart shows where a photometric target selection would be inserted in the spectroscopic selection pipeline. The photometric classifier performance is quantified by its precision and recall (defined in Sect. 5.2.1), while the final spectroscopic sample is characterised by the redshift purity and sample completeness.

([1206, 1892] nm). Both the grisms have a dispersion of 13 $\AA$/pixel. With greater sensitivity and an extended wavelength range, the Deep Survey will be used to construct a reference galaxy sample with secure spectroscopic redshift measurements, to characterise the selection function and redshift error distribution of the Wide Survey.

The design of the *Euclid* spectroscopic survey poses a particular challenge for sample selection: bright emission-line galaxies for which the redshift can be measured make up a small fraction of all photometrically detected sources and this sample is not known beforehand. We can illustrate our expectations of the *Euclid* spectroscopic sample using the Flagship2 mock galaxy catalogue, which was calibrated against the Hα luminosity function model 3 of Pozzetti et al. (2016). The mock catalogue contains approximately 2×10^5 galaxies/deg^2 to the magnitude limit $H_E < 24$. Out of this sample, only 2% are in the redshift range $0.9 < z < 1.8$ and have Hα emission-line flux greater than 2×10^{-16} erg s^{-1} cm^{-2}. The majority of the photometrically detected sources with $H_E < 24$ will leave no signal on the spectrograph, being either too faint in continuum emission, or not having a detectable emission line in the wavelength range of the red grism. When targeting galaxies at the low signal-to-noise limit, spurious noise features can be mistaken for emission lines leading to wrong redshift measurements. Current end-to-end tests of the data reduction pipeline suggest that the spurious detection rate is even higher than the naive prediction based on Gaussian noise statistics due to artefacts from spectral contamination. If not appropriately treated, such wrong redshifts in the galaxy catalogue degrade the cosmological constraints derived from the two-point correlation function or power spectrum galaxy clustering statistics (Addison et al., 2019).

In principle, when selecting the sample for analysis all available information should be used to minimise the fraction of spurious measurements, while at the same time, maximising the number density of the sample, or other figure of merit. However, the benefits from including additional constraints in the sample selection criteria must be carefully weighed against potential systematic biases. In the case of *Euclid*, including additional information from ground-based photometry modifies the selection function of the survey and could couple the sample with unwanted systematic effects that arise

from observations made through the Earth's atmosphere (see, e.g., Ross et al., 2011, for a quantitative discussion of the impact of angular systematics on the measured clustering). The trade off of adding ground-based information will clearly also depend on the scientific analysis being considered. With slitless spectroscopy, since every galaxy in the field is in any case observed, we shall have the important advantage of being able to test a posteriori the impact of any chosen selection on the measured clustering, and evaluate the robustness of the results.

Our aim with this work is to investigate photometric classification criteria that are sensitive to both redshift and emission line flux, in order to identify the sources that are likely to give successful spectroscopic redshift measurements in the Wide Survey. This strategy is similar to the methods used in ground-based spectroscopic surveys that make use of magnitude and colour selections to build the target sample for spectroscopy. For example, colour selections were applied to build the SDSS Luminous Red Galaxy sample (Eisenstein et al., 2001) and VIPERS (Guzzo et al., 2014). A sample of emission line galaxies was targeted by eBOSS using a colour selection (Comparat et al., 2016), and a similar approach was adopted for the emission line galaxy sample targeted by the dark energy spectroscopic instrument (DESI; Raichoor et al., 2023).

As a generalisation of the conventional colour cuts that are made in a two-dimensional colour-colour plane, we apply machine learning-based classification algorithms. These algorithms are well suited to optimising classification tasks in a high-dimensional parameter space. Thus, we expect them to outperform simple selection rules.

An option that is immediately available for such a use are photometric redshifts. *Euclid* will construct an unprecedented photometric redshift catalogue from the combination of ground-based and *Euclid* photometric bands. However, as we will discuss, photometric redshifts alone do not solve the problem. Even if photometric redshifts allow us to select a sample of galaxies at the target redshift range, additional criteria on galaxy physical properties, such as the star formation rate, will still be needed to identify the population with bright emission lines (see Sect. 5.4.4).

A schematic representation of the *Euclid* spectroscopic sample selection pipeline is shown in Fig. 5.1. A redshift measurement will be performed for all sources detected in photometry, and will be accompanied by an assessment of its confidence level, as well as the measurements of spectral features including emission line fluxes. Sources that do not have a significant detection in spectroscopy should be assigned a low measurement confidence. Additionally, *Euclid* will produce photometric catalogues based on the $I_{\rm E}$, $Y_{\rm E}$, $J_{\rm E}$ and $H_{\rm E}$-band images, which will be augmented with ground-based measurements (u, g, r, i, z) needed particularly for photometric redshift estimation (Stanford et al., 2021).

The photometric classification that we discuss enters as a second input to spectroscopic sample selection. The classifier can be trained on the Deep Field catalogues, which is expected to give robust redshift measurements for the emission line target galaxies in the Wide survey. The classifier will be applied to the photometric data of the Euclid Wide Survey, and its results combined with the spectroscopic measurements to build the final selected sample. This can be characterised in terms of its *redshift purity* and *sample completeness*. Any photometric criteria will necessarily reduce the number density of the sample; however, if emission line galaxy targets can be identified from the photometry, this will increase the fraction of correctly-measured redshifts and improve the purity.

We use the terms sample completeness and redshift purity to characterise the quality of the *Euclid* spectroscopic samples. We define completeness with respect to the Hα

emission line galaxy sample that exists in the Universe, which we call the true targets.[1] These are defined by a set of intrinsic properties, including angular position, redshift, size and flux, that do not depend on the measurement process. Once the observations are made, we construct the sample catalogue which contains the set of measured properties, signal-to-noise estimates and quality flags for the detected sources. The completeness tells us the fraction of the true targets that have a correct redshift measurement and makes it into the sample for analysis,

$$C = \frac{N_{\text{True Targets \& Sample \& Correct-}z}}{N_{\text{True Targets}}} . \tag{5.1}$$

On the other hand, the redshift purity tells us the fraction of the sample that has a correct redshift measurement,

$$P = \frac{N_{\text{Sample \& Correct-}z}}{N_{\text{Sample}}} . \tag{5.2}$$

The redshift purity only makes reference to the sample selected for analysis and does not depend on other intrinsic properties of the galaxies besides redshift.[2]

In this chapter, we focus on the photometric classification, which is one step of the selection process illustrated in Fig. 5.1. We consider the potential gain from the photometric classification in terms of its precision and recall (defined in Sect. 5.2.1), which will impact the final purity and completeness of the spectroscopic redshift sample. The photometric selection reduces the size of the sample in the numerator of completeness (Eq. 5.1) and thus leads to a lower value of completeness. However, it acts on both the numerator and denominator of purity (Eq. 5.2), and so is a way to potentially boost the purity. The propagation of the photometric classification to the spectroscopic sample selection and the computation of purity and sample completeness requires full end-to-end simulations of the *Euclid* reduction pipeline. In Sect. 5.5, we will present results from preliminary simulations based on the *Euclid* spectroscopic pipeline, leaving a more detailed investigation to follow-up work.

The chapter is organised as follows. In Sect. 5.2 we present the different algorithms we tested, and introduce the metrics we used to quantify the classifier performance. In Sect. 5.3 we discuss the mock catalogues, the noise model we apply to the photometry, and give the target definition. The results of the different analyses are presented in Sect. 5.4 and discussed in Sect. 5.4.4. In Sect. 5.5 we discuss how the photometric selection affects the spectroscopic sample. We conclude in Sect. 5.6.

5.2 Classification algorithms

A classifier is an algorithm that outputs the probability of an object of being an element of a given class, or group. For the purpose of this work, which is to identify target galaxies from their photometric properties, we use a binary classifier. In this case, the algorithm simply outputs the probability p of the object being a target, and $1 - p$ the probability of it being a non-target. A galaxy enters the target sample if $p > p_{\text{thresh}}$, where p_{thresh} is a threshold probability value. How the threshold is chosen is discussed in Sect. 5.2.1.

[1] This definition differs from that typically used in ground-based multi-object spectroscopic surveys that define completeness with respect to a known target sample constructed from photometric catalogues. Since the detection in *Euclid* spectroscopy will depend primarily on the signal-to-noise ratio of the emission lines, the sample with spectroscopic redshifts will not be representative of a simple photometric selection.

[2] We do not consider the sample purity, which can include other criteria such as flux, since our main objective is to select galaxies with good redshift measurements for the galaxy clustering analysis.

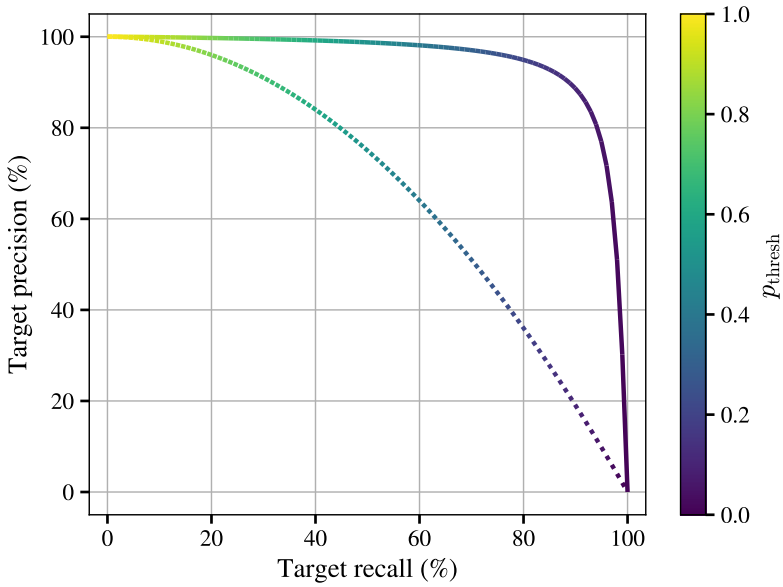

Figure 5.2: Relationship between precision and recall of a classifier. The lines are colour-coded as a function of the classification probability threshold. The solid and dotted lines show the behaviour of two classifiers for illustration. The classifier represented by the solid line performs better than the dotted line since it gives higher precision and recall.

In this work, we tested six different machine learning classifiers. The first three are self-organising maps (SOMs), dense neural networks (NNs) and support vector machine classifiers (SVCs). The other three are voting classifiers based on decision trees: the random forest (RF), the adaptive boosting classifier, or AdaBoost (ADA), and the extremely randomised tree classifier, or extra-tree classifier (ETC). These specific algorithms were chosen for our tests as they are known to perform well in classification tasks and are able to identify non-linear boundaries between classes.

5.2.1 Classification metrics

To compare the results from different classifiers, we adopt three metrics defined from their *confusion matrix*. The elements of the confusion matrix of a binary classifier are the counts of true positives ($N_{\rm TP}$), true negatives ($N_{\rm TN}$), false positives ($N_{\rm FP}$), and false negatives ($N_{\rm FN}$). Our chosen metrics are the *precision*, *recall*, and *false positive rate* (FPR), defined respectively as

$$\text{precision} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FP}}\,, \tag{5.3}$$

$$\text{recall} = \frac{N_{\rm TP}}{N_{\rm TP} + N_{\rm FN}}\,, \tag{5.4}$$

$$\text{FPR} = \frac{N_{\rm FP}}{N_{\rm FP} + N_{\rm TN}}\,. \tag{5.5}$$

The precision is the fraction of the selected sample that are true targets, i.e., it quantifies the level of contamination due to wrongly classified sources. The recall, also known as

true positive rate, is the fraction of true targets that are identified correctly (as $N_{\rm TP} + N_{\rm FN}$ corresponds to the total number of targets). The false positive rate, or *fall-out*, is the fraction of non-targets that are mislabelled as targets and enter the selected sample as interlopers. The complement of the false positive rate is the *true negative rate*,

$$\mathrm{TNR} = \frac{N_{\rm TN}}{N_{\rm FP} + N_{\rm TN}} = 1 - \mathrm{FPR}\,, \tag{5.6}$$

which characterises the fraction of non-targets that are correctly removed from the sample.

These metrics change as functions of the probability threshold chosen for the classifier, i.e., the probability value $p_{\rm thresh}$ above which an object is classified as a target. This is a hyper-parameter of the model, which we set to maximise a chosen metric. In a binary classification, a training set is said to be 'balanced' when it is evenly split between targets and non-targets, and $p_{\rm thresh} \sim 0.5$. When the training set contains a much larger number of targets than non-targets, or vice versa, it is called *unbalanced*, and we refer to this case as an *unbalanced classification*. In general, in unbalanced classifications the optimal probability threshold is very different from 0.5. Precision and recall can be computed as a function of $p_{\rm thresh}$ and plotted against each other, as shown in the example of Fig. 5.2. Such a plot is very informative for the photometric selection task that is the scope of our work. In Fig. 5.2 we present two possible behaviours of this curve. The solid line is an almost ideal classifier that has high precision also when the recall is high, while the dotted curve corresponds to a classifier with worse performance. Since the photometric criteria make up only one step of the spectroscopic sample selection process (see Fig. 5.1), we want to keep the recall of the photometric classification as high as possible. In other words, we want to get a resulting sample as complete as possible, discarding the minimum number of true targets. Thus, we choose a specific value for the recall and, from this relation, derive the corresponding precision yielded by the algorithm. We use the precision at 95% recall as our benchmark value. A similar plot can be produced in terms of redshift purity and sample completeness. The shape of this curve will depend on the chosen probability threshold, and consequently the recall, of the photometric classification. In Sect. 5.5 we justify the choice of the 95% recall value and present results for the redshift purity and sample completeness.

Finally, we use the false positive rate as the main metric to compare algorithms trained with different input features (see Sect. 5.4.4). The false positive rate helps to visualise the fraction of misidentified objects in terms of redshift or emission-line flux and shows the source of the contaminants.

5.2.2 Self-organising map

Self-organising maps (Kohonen, 1982, 1990) use unsupervised learning to project a high-dimensional feature space onto a lower-dimensional one, usually a two-dimensional space, as the name map suggests. We build a 55×55 map trained for 60 epochs, where an epoch corresponds to an iteration of the algorithm during which the entire training set is processed. To train the self-organising map, in addition to the photometric features used as inputs for all the other methods, we also add the target label (see Sect. 5.3). Then, when projecting new data onto the self-organising map the target labels are removed. These steps make the implementation of the self-organising map presented here more similar to a supervised learning algorithm. We also introduce a weight, $w_{\rm SOM}$, of the photometric features, which enables us to control the importance of the label in the training. This is a hyper-parameter of the self-organising map model. Finally, the probability of an

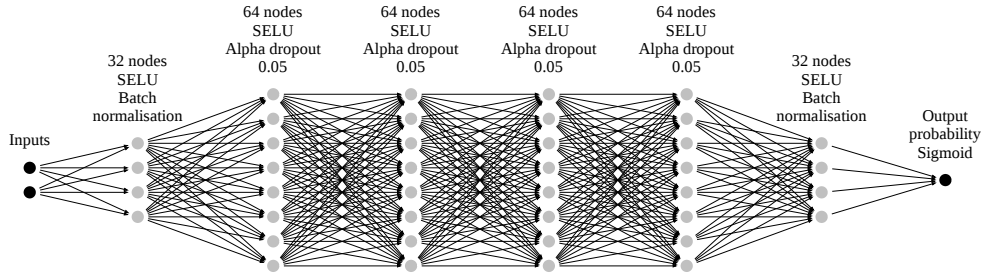

Figure 5.3: A schematic representation of the neural network architecture used for classification. Values pass from the input to the output along the connected edges; each node represents a linear combination of the inputs and the application of a non-linear activation function. The value at the output represents the binary classification probability between 0 and 1. The number of input neurons varies for the different configurations (4 for *Euclid*-only, and 8 after adding ground-based photometry, see Sect. 5.4). For visualisation, the number of neurons in each hidden layer has been divided by 4.

object of being a target is defined by the target fraction in the cell it has been projected onto. The self-organising maps were implemented using `SOMPY` (Moosavi et al., 2014).

5.2.3 Neural network

Neural networks are by far the most popular supervised learning algorithms. They can be described as a sequence of layers; when the inputs are processed by a layer they first undergo a linear transformation and then a nonlinear function is applied to them. During the learning process, the neural network updates the coefficients, usually called weights, of the linear transformation of each layer in order to fit the target function $y = f(x)$ that relates the inputs x, to the labels y. This structure enables neural networks to potentially fit any function of the input features (LeCun et al., 2015).

Our neural network architecture was optimised for the problem at hand. Figure 5.3 shows a schematic representation of the neural network. The input layer is followed by a first block that consists of a dense layer with 32 neurons and a batch normalisation layer (Ioffe & Szegedy, 2015). Then, a second block which consists of a dense layer with 64 neurons and an alpha dropout layer (Klambauer et al., 2017) with rate 0.05 is repeated four times. Finally, the first block is repeated before the output layer, which consists of 1 neuron. The activation function of all the layers except for the output is a scaled exponential linear unit (SELU; Klambauer et al., 2017). The last layer, as it has to output a probability, has a sigmoid activation function. Since the ratio between positive and negative examples is very low, we opted for a sigmoid focal cross-entropy loss function (Lin et al., 2017),

$$\mathrm{FL}(p) = -\alpha\,(1-p)^{\gamma}\ln(p)\,, \tag{5.7}$$

where α and γ are two hyper-parameters of the model. We use $\alpha = 0.6$ and $\gamma = 4$. We implemented the neural network in the TensorFlow2 framework (Abadi et al., 2015).

5.2.4 Support vector machine classifier

Support vector classifiers (Boser et al., 1992) partition the feature space by applying a kernel transformation to map curved boundaries into planes and finding the maximum-margin hyperplane that separates the classes. It is important to note that for our training we weight differently the target and non-target examples. This weighting is necessary in the case of imbalanced classes. Alternatively, one could select a balanced subsample of the original training set. However, such a solution would greatly reduce the size of the training sample. Our approach uses the support vector classifier implementation of `scikit-learn` (Pedregosa et al., 2011), which has an inbuilt functionality to balance the sample via weighting.

We adopt the `scikit-learn` default kernel, which is the radial basis function kernel (RBF),

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|^2\right) , \tag{5.8}$$

where $\|\mathbf{x} - \mathbf{x}'\|^2$ is the Euclidean squared distance, and γ is the hyper-parameter that controls the dimension of the region of influence of the training point.

5.2.5 Decision tree-based classifiers

The last three classifiers are voting or ensemble classifiers. In general, a voting classifier is an algorithm that combines the output of different base classifiers through a vote, which can be weighted or not. In this work we used classifiers based on the same base algorithm, the decision tree. These classifiers differ in how they split the data set to train the trees, how they build the trees, and how they combine together their probability outputs.

A decision tree is a supervised machine learning model that approximates a function with a series of simple decision rules (see Hastie et al., 2001, Chap. 9). Decision trees have the advantages that they can be easily visualised, have high explainability, and require very little data preparation; however, they can easily over-fit the training sample making their output and final structure dependent on the training set. These issues can be reduced by combining the results of different trees (e.g., Bauer & Kohavi, 1999).

The first of these voting classifiers are random forests. Random forests (Breiman, 2001) are an ensemble of decision trees each one trained with a subsample of the training set. This subsample is a bootstrap sample, which means its elements are randomly selected with replacement from the complete training set. The final output of the random forest for classification tasks is a majority voting between all the decision trees of the forest. Random forests very efficiently reduce the overfitting of single-decision trees. To take into account the class imbalance of the sample we weigh the two class examples by the inverse of their frequency. The weights are computed for each bootstrap subsample.

The second ensemble classifier is a discrete adaptive boosting classifier (Freund & Schapire, 1997). Differently from the random forest, adaptive boosting classifiers can use different base classifiers. In this work, we limited the analysis to adaptive boosting classifiers based on decision trees with weighted data to balance the sample examples. Adaptive booster classifiers combine the results of subsequently trained base learners with a weighted majority vote. At each step of the training, a new learner is built from the training set, which is re-weighted to reduce the importance of data that have been correctly classified in the previous steps.

Finally, the last algorithm we use is the extra-tree classifier. Extra-tree classifiers are ensemble classifiers based on decision trees (Geurts et al., 2006). An extra-tree classifier is composed of a group of decision trees, which are trained with bootstrap subsample

of the training set, as in random forest training. The difference between a random forest and an extra-tree classifier lies in how the decision rules of the trees are selected. In random forests, the splits of the tree nodes are deterministic and depend on the selection algorithm; in extra-tree classifiers, instead, they are randomly drawn and the final rule is chosen as the best-performing one among them. This helps in reducing even more the variance of the method. All three voting classifiers are implemented in `scikit-learn`, and the function to weigh the data to balance them is part of their built-in functionalities.

5.3 Benchmark data

5.3.1 Mock galaxy catalogues

We use two catalogues to benchmark the selection algorithms: the EL-COSMOS catalogue and the Euclid Flagship2 mock galaxy catalogue. These catalogues include broad-band photometry, emission-line fluxes and morphological properties. We make use of the *Euclid* photometric bands from VIS, $I_{\rm E}$, and NISP, $Y_{\rm E}$, $J_{\rm E}$, and $H_{\rm E}$ (Euclid Collaboration: Schirmer et al., 2022), with depths listed in Table 5.1. Additionally, photometric data from multiple ground-based surveys will be included in *Euclid* analyses to extend the wavelength coverage to the optical with u, g, r, i, and z bands and obtain reliable photometric redshifts that are key for *Euclid* weak lensing science. These include the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al., 2009), the Dark Energy Survey (DES; Flaugher, 2005), and the Ultraviolet Near Infrared Optical Northern Survey (UNIONS).[3] In order to benchmark the photometric selection in this work, we use the Vera C. Rubin filter system, $ugriz$, and UNIONS survey depths, which are listed in Table 5.1. Hereafter, we refer to the photometry of the four *Euclid* filters as *Euclid* photometry, and to the photometric data from the five LSST filters as ground-based photometry. The photometry does not include the effect of Milky Way extinction.

The resolution of *Euclid* NISP spectroscopic observations is not sufficient to separate Hα from its neighbouring [N II] λ6549 and [N II] λ6584 companions. As such, *Euclid* will measure the combined flux of this triplet of emission lines, which we shall use here and indicate for brevity as

$$f_{\rm H\alpha+[N\,II]} = f_{\rm H\alpha} + f_{\rm [N\,II]\lambda6549} + f_{\rm [N\,II]\lambda6584}\,. \tag{5.9}$$

We refer to the triplet as the Hα complex.

We also investigate the benefit of adding morphological information to the target classification. The two mock galaxy catalogues we use here include morphological model parameters including disk ellipticity, bulge scale, disk scale and bulge-to-disk ratio; however, since these properties will not be, in general, directly measured from the data, we used them to derive the observable half-light radius, $r_{\rm half}$ and axial ratio, e. To do this, we ran `GALSIM` (Rowe et al., 2015) using the morphological parameters for each mock galaxy to generate a simulated image of the galaxy as it would be observed by VIS, from which we estimated the half-light radius and axial ratio. We carried out this procedure only for the Flagship2 catalogue.

We use only galaxies in the mock catalogue, without accounting for the possibility that stars or active galactic nuclei may be misclassified in real data and enter the sample. Contamination from faint stars, in particular, can potentially reduce the purity of the galaxy sample. The severity of such contamination depends on the performance of the

[3] https://www.skysurvey.cc/aboutus/.

star-galaxy classification, which is a separate step of the Euclid data analysis and whose impact is beyond the scope of this work.

EL-COSMOS

The EL-COSMOS catalogue is an extension of the COSMOS 2020 photometric catalogue (Weaver et al., 2021). The COSMOS catalogue is a multi-band data set assembled in the *Hubble* Space Telescope COSMOS field over the past fifteen years (Scoville et al., 2007). The catalogue was extended as described in Saito et al. (2020) with synthetic photometry and emission-line fluxes assigned by spectral energy fits. To assign the fluxes of the emission lines the authors combined spectral energy fits of the stellar continuum, which correlates with the intrinsic emission line fluxes, with a careful modelling of dust attenuation as a function of redshift. We use an update to the emission-line catalogue produced for the Euclid Consortium (Euclid Collaboration, in prep.). It contains about 2×10^5 galaxies and 2000 active galactic nuclei. This catalogue also contains stars observed in the COSMOS field, which, as explained, we do not consider.

Euclid Flagship

The Euclid Flagship2 mock galaxy catalogue (Euclid Collaboration, in prep.) is based on the Flagship2 N-body simulation, the large reference simulations built by the Euclid Consortium. Galaxies were added to the simulation using an extended halo occupation distribution model. The Flagship2 galaxy mock catalogue represents an improvement with respect to the previous version in terms of modelling of the galaxy properties. The catalogue contains photometric and spectroscopic information, morphological parameters, along with lensing properties. The morphological parameters are correlated with the galaxy properties to reproduce observed trends in galaxy size. For our work, we selected a subsample of $\sim 2 \times 10^5$ objects to contain a number of galaxies comparable to EL-COSMOS. We note that Flagship2 does not contain active galactic nuclei, while EL-COSMOS contains about 2000 of them.

An additional step must be taken to compute the total flux of the Hα complex for Flagship2 mock galaxies. The catalogue gives the flux of Hα and of the [N II] $\lambda 6584$ line only. Assuming a relative 1:3 ratio for the [N II] doublet, we estimate the total flux as

$$f_{\mathrm{H}\alpha+[\mathrm{N\,II}]} = f_{\mathrm{H}\alpha} + \frac{4}{3} f_{[\mathrm{N\,II}]\lambda 6584}\,. \tag{5.10}$$

The emission line fluxes in Flagship2 were calibrated against the Hα luminosity function model 3 of Pozzetti et al. (2016). We use the line and broadband fluxes with internal dust attenuation applied. From Flagship2 we use both *Euclid* and ground-based photometric data, as well as the morphological parameters derived as discussed earlier.

The Flagship2 catalogue also provides photometric redshift estimates obtained with state-of-the-art algorithms using both *Euclid* and ground-based photometry (Euclid Collaboration: Desprez et al., 2020). In order to allow the computations of photo-zs for billions of *Euclid* sources, a two-stage approach has been adopted. First, `Phosphoros`, a template-fitting code (Paltani et al. in preparation), is used to compute the redshift probability distribution functions on a sample of galaxies selected from reference fields that benefit from very deep observations in a large number of photometric bands (e.g., COSMOS; Weaver et al., 2021). The k-nearest neighbour photometric redshift algorithm (Tanaka et al., 2018) is then used to estimate the posterior distributions of redshift for sources in the Euclid Wide Survey. This procedure was replicated in the Flagship2 mock

Band	$m_{\mathrm{lim},10\sigma}$
u	23.5
g	24.4
r	24.1
i	23.5
z	23.3
I_{E}	24.6
Y_{E}	23.0
J_{E}	23.0
H_{E}	23.0

Table 5.1: Point source magnitude limits at depth $(\mathrm{S/N})_{\mathrm{lim}} = 10$ for $ugriz$, and for I_{E}, Y_{E}, J_{E}, and H_{E} in AB magnitude.

galaxy catalogue. In this work, we use the first mode of the posterior redshift distribution as the photo-z estimate. We use the photo-z to select galaxies within the redshift range of interest and compare the metrics with the results from the trained classifiers.

5.3.2 Noise model

The errors on the broadband photometric measurements were simulated assuming background limited observations (Euclid Collaboration: Pocino et al., 2021) such that the standard deviation on the measurement is

$$\sigma_f = \frac{f_{\mathrm{lim}}}{\mathrm{S/N}_{\mathrm{lim}}}\,, \tag{5.11}$$

where f_{lim} is the flux at the specified signal-to-noise limit $\mathrm{S/N}_{\mathrm{lim}}$. In Table 5.1 we show the AB magnitude limits, m_{lim}, corresponding to f_{lim}. for $(\mathrm{S/N})_{\mathrm{lim}} = 10$.[4] The observed fluxes were then extracted from a Gaussian distribution with the true galaxy flux, f, as mean, and variance given by σ_f. In order to be able to reproduce the results, we constructed observed catalogues for both EL-COSMOS and Flagship2, which contain realisations of the flux errors produced following the recipe described above.

The driving idea in the application of our selection procedure to the real *Euclid* data, is that the training set will be constructed from the higher signal-to-noise data of the Euclid Deep Fields, which will have high completeness and purity at the depth of the Wide Survey. In order to build a training set that matches the noise properties in the Wide Survey, the photometry from the Deep Fields will have to be either measured in Wide-like stacks, or degraded appropriately as to match the noise level of the Wide Survey.

5.3.3 Sample selection and pre-processing

For our analysis, we selected from the EL-COSMOS and Flagship2 catalogues two subsamples limited to $H_{\mathrm{E}} < 24$ (which corresponds to a 4σ point-source detection limit). In addition, as mentioned earlier, the resulting Flagship2 catalogue was further sparsely sampled in order to match the same number of objects of EL-COSMOS. Each catalogue

[4]The magnitude limits for UNIONS in Table 5.1 were computed from the 5σ limits available at https://www.skysurvey.cc/survey/.

was then split into three subsets, for training, validation, and testing, containing respectively 75%, 15%, and 10% of the total parent catalogue. In fact, the validation set is needed only for the training of the neural network; for the other algorithms, we could use 90% of the total sample as the training set. However, for the sake of a fair comparison, we opted to use the same training and test sets for all methods, by discarding the validation set objects when not needed.

The galaxies we aim to select with the photometric selection have, on top of the $H_{\rm E} < 24$ cut,

$$\begin{cases} 0.9 < z < 1.8 \\ f_{\rm H\alpha+[N\,II]} > 2 \times 10^{-16}\,{\rm erg\,s^{-1}\,cm^{-2}} \end{cases} , \tag{5.12}$$

where z and $f_{\rm H\alpha+[N\,II]}$ are the true redshift and emission line flux of the galaxies. The objects satisfying this selection are what we call *target* galaxies. In terms of the classifier training, we assign a label 1 to the target galaxies and a label 0 to the remaining objects, hereafter non-targets. It should be noted that the $\rm H\alpha + [N\,II]$ flux criterion in the target definition is specified to select galaxies with bright emission lines that are likely to give successful spectroscopic redshift measurements. We will see that this target definition does not impose a sharp flux cut in the measured sample; galaxies just below the flux limit still have a high probability of being selected and of giving a correct redshift measurement. Moreover, these galaxies will also contribute to the redshift purity metric.

The percentage of galaxies entering the target sample within the full $H_{\rm E} < 24$ catalogues is very low: $\sim 8\%$ for EL-COSMOS and $\sim 3\%$ for Flagship2. The difference between the two catalogues is consistent with the current uncertainty in the $\rm H\alpha$ luminosity function at $z > 1$. The low target fractions of the two catalogues make the classification task extremely unbalanced. The solutions adopted for each classifier were discussed in Sect. 5.2 and span from weighting schemes to specific loss functions.

Finally, all input training parameters are pre-processed via standard scaling,

$$X = \frac{x - \bar{x}}{\sigma_x} , \tag{5.13}$$

where $\bar{x}$ is the mean value of input feature x over the training sample, and σ_x its standard deviation. After this normalisation, the sample has zero mean and unit standard deviation, which makes the training of the algorithms more efficient, typically leading to better results.

5.4 Results and discussion

5.4.1 Benchmark selections

Before discussing the performance of the machine learning classifiers we present the results from simple classifiers based on magnitude and colour with *Euclid* photometry. These tests provide a benchmark for the machine learning algorithms. We focus on the $(I_{\rm E} - H_{\rm E})$ versus $H_{\rm E}$ plane, which shows the largest displacement between targets and non-targets (see Figs. 5.4 and A.1). The distributions are seen to be most separated in $H_{\rm E}$ magnitude. Indeed, the use of $H_{\rm E}$ is expected to be particularly suited to capture information on the $\rm H\alpha$ flux, as it covers the $[1.5, 2.0]\,\mu{\rm m}$ band, which encompasses the $\rm H\alpha$ complex for $1.3 \lesssim z \lesssim 2$. In addition, the $(I_{\rm E} - H_{\rm E})$ colour is sensitive to redshift, since it spans the $4000\,$Å break at $z > 1$.

We thus begin by applying a cut in $H_{\rm E}$ to select the target sample. Table 5.2 gives the resulting recall and precision metrics. For the Flagship2 catalogue, all targets have

	EL-COSMOS		Flagship2	
$H_{\rm E}$ cut	Recall	Precision	Recall	Precision
22.84	95	13.8	-	-
22.06	-	-	95	8.9
21.0	20.2	11.3	47.6	10.0
22.0	63.1	16.3	93.3	9.1
23.0	97.1	12.7	100	4.8
24.0	100	7.8	100	2.6
Colour cut	95	14.3	95	9.9

Table 5.2: Recall (%) and precision (%) for different $H_{\rm E}$ cuts for EL-COSMOS and Flagship2. The first two rows correspond to the $H_{\rm E}$ cut that gives 95% recall respectively for EL-COSMOS and Flagship2.

		SOM	NN	SVC	RF	ADA	ETC
Euclid	EL-COSMOS	13.9	17.5	17.3	16.4	12.9	16.7
	Flagship2	12.7	16.0	18.0	15.5	10.4	16.9
Euclid +	Flagship2 morphology	9.6	17.6	16.8	15.3	11.0	14.7
	EL-COSMOS ground	20.7	34.3	34.3	31.5	29.1	28.0
	Flagship2 ground	26.1	47.9	43.5	39.3	39.7	35.6

Table 5.3: Precision values (%) at 95% recall for the different classifiers. The two top rows give the results for training using *Euclid* photometry only, while morphological data and ground-based photometry, respectively, are used in the bottom rows. The relative uncertainty on all values is $\sim 6\%$, estimated from multiple realisations of the training and test sets.

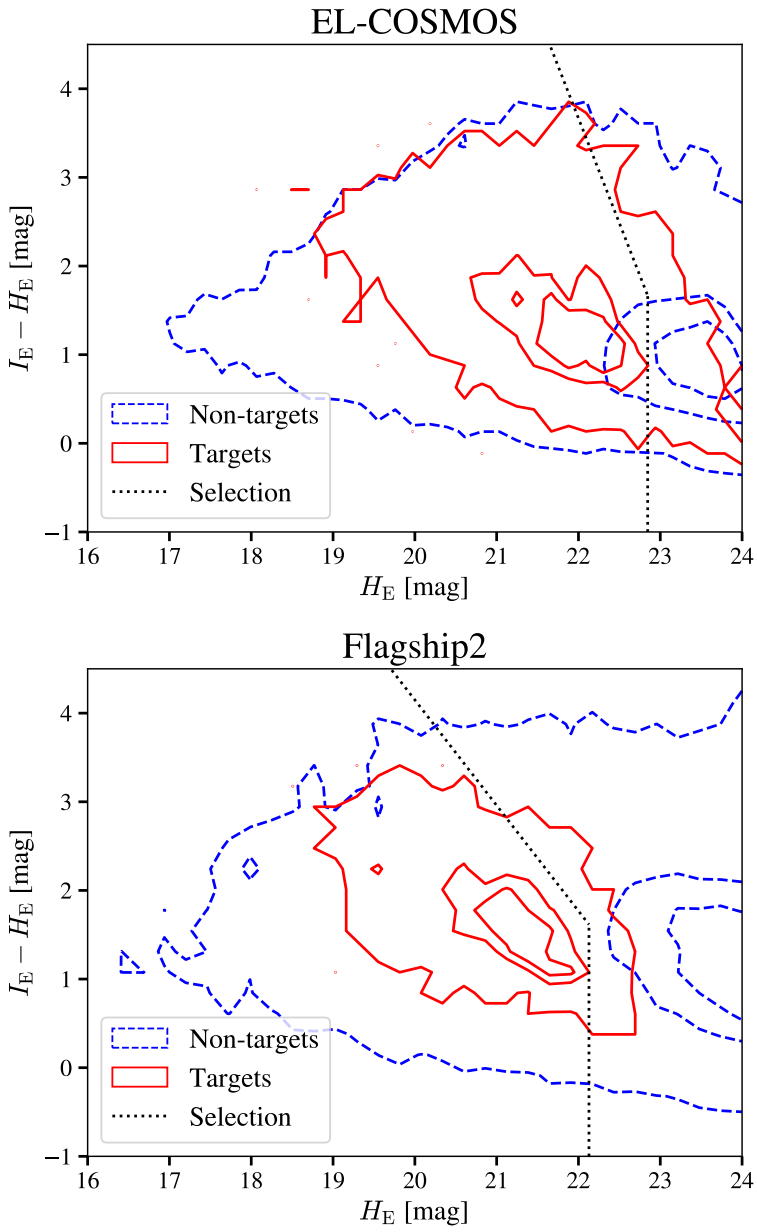

Figure 5.4: Optimised colour selection in the $(I_E - H_E)$ versus H_E colour-magnitude plane for EL-COSMOS and Flagship2. The blue dashed lines correspond to the non-target distribution and the solid red lines to the target distribution. The contours contain 99%, 50%, and 25% of the samples. The dotted black segments represent an optimised colour cut in this plane corresponding to recall $\sim 95\%$.

$H_{\rm E} < 23$ giving 100% recall at that limit, while for EL-COSMOS, 100% recall is reached at $H_{\rm E} < 24$.

Next, we consider a selection in the $(I_{\rm E} - H_{\rm E})$ versus $H_{\rm E}$ plane. The colour-magnitude selection reads as follows,

$$(I_{\rm E} - H_{\rm E}) < a\,(H_{\rm E} - b) \quad \text{AND} \quad H_{\rm E} < H_{\rm E}^{\rm cut}\,. \tag{5.14}$$

We searched for a selection with the form of Eq. (5.14) that maximises the purity while giving recall $\sim 95\%$. The best colour cut for EL-COSMOS has slope $a = -2.36$, pivot $b = 23.60$, and $H_{\rm E}^{\rm cut} = 22.85$. For Flagship2 the slope is $a = -1.90$, $b = 14.74$, and $H_{\rm E}^{\rm cut} = 22.13$.

Figure 5.4 shows the targets (solid red) and non-target (dashed blue) distributions in the colour-magnitude plane of interest for EL-COSMOS (top panel) and Flagship2 (bottom panel). The dotted black line corresponds to the colour-magnitude cut. The two panels show the difference in the target distributions of EL-COSMOS and Flagship2. Flagship2 does not have any targets with $H_{\rm E} > 23$, in contrast, EL-COSMOS targets reach the magnitude limit of the sample. For this reason, we allow the $H_{\rm E}$ cut to adapt to the training data. We report the precision of these cuts in the bottom line of Table 5.2.

From Table 5.2, we see that the selection gives a higher precision for EL-COSMOS than Flagship2. This can be understood since the fraction of targets is lower in Flagship2 than in EL-COSMOS. We next show the results from machine learning classifiers, which make full use of the high-dimensional parameter space to optimise the selection.

5.4.2 Using *Euclid* data only

We first discuss the results obtained training the classifiers using only *Euclid* photometry, comparing the two catalogues EL-COSMOS and Flagship2. The input features for each object are the same for both catalogues, namely its $H_{\rm E}$ magnitude and near-infrared colours, $(I_{\rm E} - Y_{\rm E})$, $(Y_{\rm E} - J_{\rm E})$, $(J_{\rm E} - H_{\rm E})$.

Figure 5.5 shows the precision-recall curve produced by the six different classifiers using respectively EL-COSMOS (top panel) and Flagship2 (bottom panel). We remark that, for an ideal classifier, the plot would show a close-to-flat precision around unity (see Fig. 5.2), followed by a sharp drop at the highest possible recall value. To provide a reference baseline, in Fig. 5.5 we also present (dotted magenta line) the curve one obtains when simply selecting $H_{\rm E} < H_{\rm E}^{\rm limit}$ magnitude-limited samples. The curve has been computed by smoothly varying $H_{\rm E}^{\rm limit}$ between 20.0 and 24.0 (see Table 5.2). The vertical black line corresponds to 95% recall, which we chose as the reference value for comparing the algorithms (see Sects. 5.2.1 and 5.5), as reported in Table 5.3.

Comparing the two panels, the first evident difference is the larger variance in performance over the whole recall range shown by the different algorithms in the case of the Flagship2 sample. Conversely, the classifiers trained with EL-COSMOS show a sharper drop in precision at small recall values. The $H_{\rm E}$ magnitude limit selection appears to be more effective for EL-COSMOS than for Flagship2. In both cases, this simple selection is (not unexpectedly) worse than the machine learning classifiers, but in the case of EL-COSMOS the resulting performance becomes comparable to that of the worse-performing classifiers at 95% recall.

Overall, Fig. 5.5 and Table 5.3 show similar performance when training with either Flagship2 or EL-COSMOS, with the former showing a larger variance at the recall threshold. Such an agreement is an encouraging indication of the robustness of the general conclusions that can be drawn from these results. In both cases, the best-performing

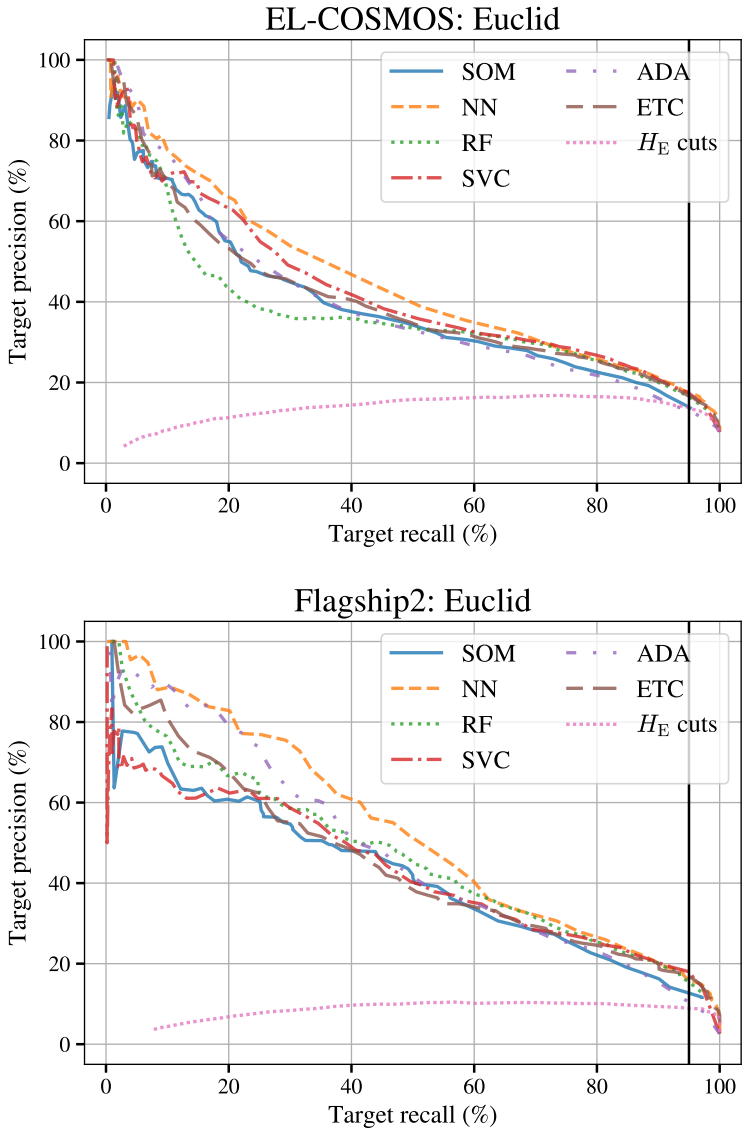

Figure 5.5: Precision vs. recall performance of the different classifiers, using *Euclid* photometry alone for the training. The two panels correspond to the two test catalogues as indicated. The vertical solid line gives our reference recall value of 95%.

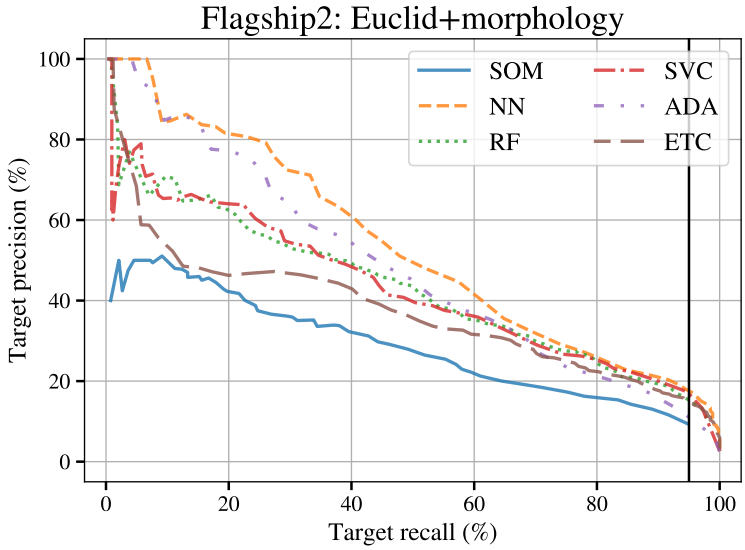

Figure 5.6: Same as Fig. 5.5, but now adding morphological information in terms of half-light radius and axial ratio values.

algorithms are the neural network, the support vector classifier, and the extra-tree classifier. The random forest follows shortly behind, indicating that the bootstrap resampling used in the decision tree training is especially efficient for this task. Last comes the self-organising map, which is not optimised for this kind of task, and the adaptive boosting classifier.

The effect of complementing Euclid infrared photometry with morphological information described by the galaxy half-light radius and axial ratio values, can be seen in Fig. 5.6. The plot shows no large improvement and some classifiers perform worse and there is an even larger variance between the different classifiers, especially at low recall values. The best-performing one is still the neural network, followed by the support vector classifier, the random forest and the extra-tree classifier. Again, the adaptive boosting classifier and the self-organising map fare poorly. A more detailed discussion is left for Sect. 5.4.4.

5.4.3 Adding ground-based photometry

When we combine *Euclid* and ground-based photometry we substitute the $I_{\rm E}$ band with the five ground-based filters, $ugriz$. In this case, the input features of the classifiers are the following seven colour combinations $(u-g)$, $(g-r)$, $(r-i)$, $(i-z)$, $(z-Y_{\rm E})$, $(Y_{\rm E}-J_{\rm E})$, and $(J_{\rm E}-H_{\rm E})$. In addition, we also use $H_{\rm E}$ as the last input feature.

Figure 5.7 shows how the diagnostic plots change when combining *Euclid* and ground-based photometry. We immediately see from Fig. 5.7 how the ground-based data improves the overall performance, yielding curves that are much closer to the ideal shape (see Fig. 5.2). In the bottom panel, we also show (green dot) the precision ($\sim 5.5\%$) and recall ($\sim 92.4\%$) values recovered when using photometric redshifts to simply isolate targets with $0.9 \leq z_{\rm photo} \leq 1.8$, with no extra information to constrain the desired Hα line flux. We note that the photometric redshift selection does not reach the 95% recall

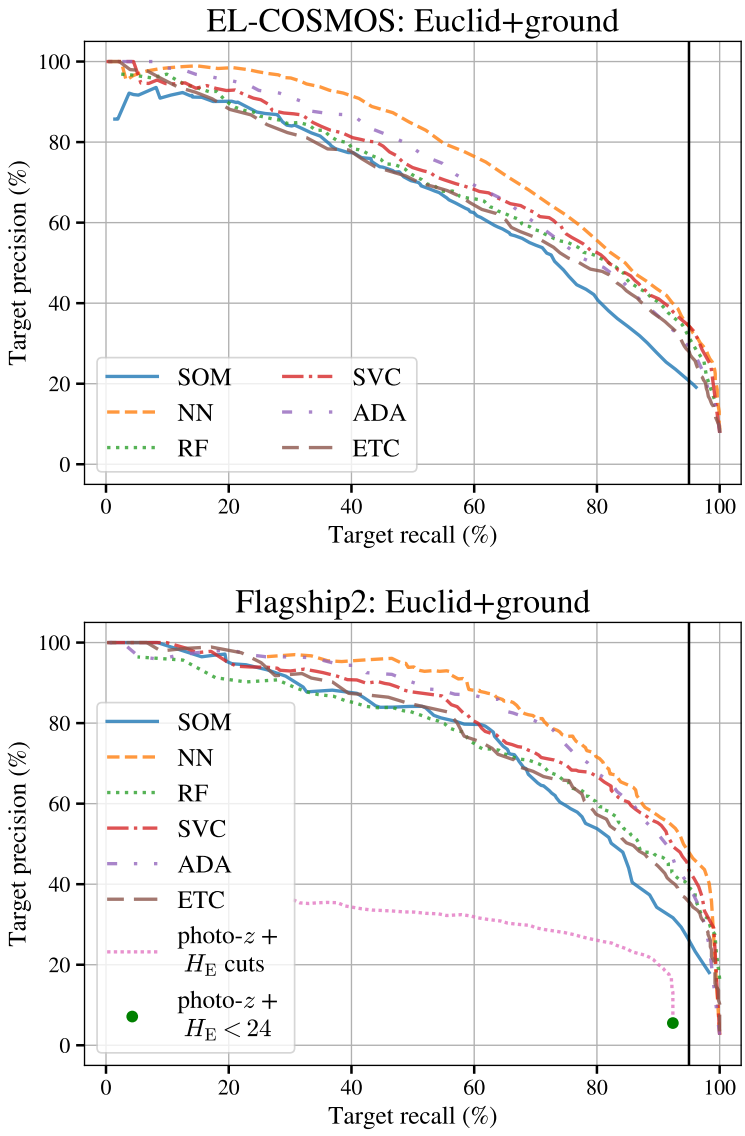

Figure 5.7: Precision versus recall curves for the analyses with *Euclid* photometry and ground-based photometry. *Top*: results for EL-COSMOS. *Bottom*: results for Flagship2. The dotted magenta line represents the combined selection of $H_{\rm E}$ cuts and the photo-z selection. The green point marks the precision and recall value obtained with the photo-z selection alone.

value. We also consider photometric redshift selections with various $H_{\rm E}$ magnitude limits, shown by the magenta dotted line. In Appendix A.2 we present a preliminary test that combines the photometric and the redshift information in the training of a neural network.

The improvement in performance appears to be larger when estimated using Flagship2 than with EL-COSMOS with a difference of $\sim 10\%$ in precision for all algorithms. The reason for this can be related to the colour distribution of the targets. In the EL-COSMOS catalogue the distribution functions of magnitudes and colours for targets show more variance than in Flagship2 where the targets are more localised on colour space. When *Euclid*-only photometry is used, the information is not sufficient for tightly constraining the target region in the parameter space, thus producing similar results from the two catalogues. However, when ground-based photometry is added, in the Flagship2 case it becomes easier to isolate the targets. These differences may be due to the recipes used for assigning spectral energy distributions and synthetic emission lines in the two catalogues.

The relative ranking of the different classifiers derived from the two catalogues is the same. The worst-performing algorithm at the recall threshold is the self-organising map, which shows a steeper drop in precision than the others (see Table 5.3). The remaining algorithms have precision values $> 35\%$ for Flagship2, with the neural network reaching almost 50%. For EL-COSMOS at the recall threshold the values of the precision are always $> 25\%$, peaking at $\sim 34\%$ for both the neural network and the support vector classifiers.

5.4.4 Comparison of the results

In this section, we focus on the results based on the Flagship2 training and discuss the results obtained with the three configurations. We will then focus on the best-performing classifier, the neural network, and discuss in more detail the three cases. We will also show a comparison with a simpler redshift-only selection based on *Euclid* photometric redshifts.

Figures 5.5, 5.6, and 5.7, together with Table 5.3, provide a direct quantitative comparison of the three training configurations: the best performance is obtained by the combined *Euclid* and ground-based photometry. For Flagship2, this more than doubles the precision at the recall threshold with respect to the other two configurations, a clear benefit of the extra information on lower redshift objects provided by the optical bands (see discussion in the following). The addition of morphological information through the half-light radius and ellipticity, conversely, does not introduce any significant improvement: the neural network and the adaptive boosting classifier show only a minimal gain, while all others worsen their performance.

The half-light radius, in particular, does show a trend as a function of redshift, but this relation has a large scatter and weak correlation coefficient. It is possible that other morphological measures that we did not consider, such as the Sérsic index, will be more sensitive to galaxy type and have a greater importance for classification; however, we reserve this investigation for future work. When fed uninformative features, the classification algorithms will tend to ignore them. The majority of the tested classifiers have, in fact, built-in mechanisms to ignore a feature. Specifically, the neural network would reduce, during the training, the weight of the specific feature that appears to be uninformative, while the decision tree-based classifier would not introduce decision rules based on it. Similarly, the support vector classifier would only produce boundaries orthogonal to an uninformative feature. The same cannot be said about a standard self-organising

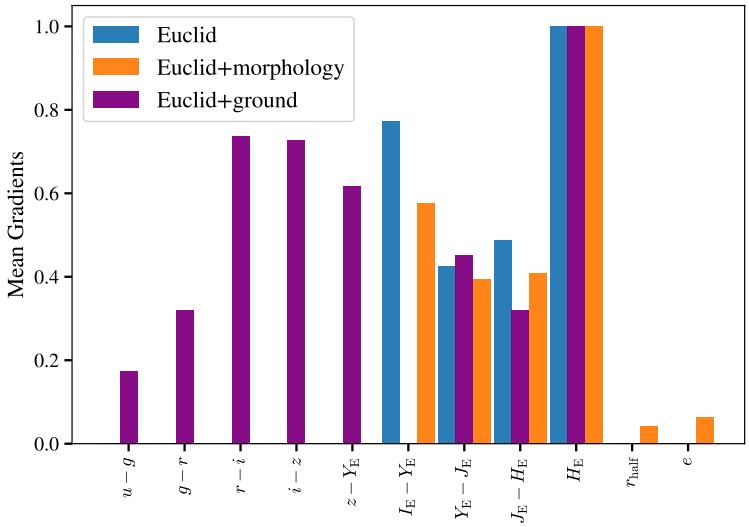

Figure 5.8: Mean gradients of the neural network output as a function of the input for the three training configurations. In blue, orange and purple are respectively plotted the mean gradients of the neural networks trained with *Euclid* photometry, *Euclid* photometry and morphology, and *Euclid* and ground-based photometry. All gradients have been normalised to that corresponding to the *Euclid* $H_{\rm E}$ magnitude.

map: in this case, the effect of an uninformative feature is to spread the classification targets over a larger number of cells, thus reducing the sensitivity.

In order to understand which features are most relevant for classification, which is known as the *saliency* in the machine learning literature, in Fig. 5.8 we show the mean gradients of the network with respect to the input features. We see that the most important feature turns out to be the $H_{\rm E}$ magnitude, followed by the ground-based colours. The dependence on the optical colours and in particular on $(I_{\rm E} - Y_{\rm E})$ in the *Euclid* photometry configuration has two main reasons. First, the optical bands retain low redshift information (see following discussion); second, the correlation between the between $I_{\rm E}$ and $f_{\rm H\alpha+[N\,II]}$ is even stronger than the correlation of the emission line flux and $H_{\rm E}$. The network uses $(I_{\rm E} - Y_{\rm E})$ to extract $I_{\rm E}$ from the pivot magnitude $H_{\rm E}$ and infer this correlation. Lastly, as expected, the morphological parameters are the least important inputs for the neural network.

Having identified the $H_{\rm E}$ magnitude as the most informative feature, we can gain additional intuition about the classifiers by comparing the number counts $N(H_{\rm E})$ of the true targets to those of the samples recovered by the neural network. These are shown in Fig. 5.9. The green histogram gives the number counts for the true targets, i.e., the reference distribution we are trying to reproduce with the classifier. Notably, the counts go to zero for $H_{\rm E} > 22.5$, hence there are no target galaxies fainter than this magnitude. This explains the rapid gain in precision one obtains by simply cutting the full sample (here shown by the orange histogram) at brighter and brighter values of $H_{\rm E}$ (see Table 5.2). Looking at the other histograms, we see that the application of the neural network effectively cuts the distribution down to the correct $H_{\rm E}$. When using only *Euclid* bands (blue histogram), this leaves an excess of sources, which are either outside the

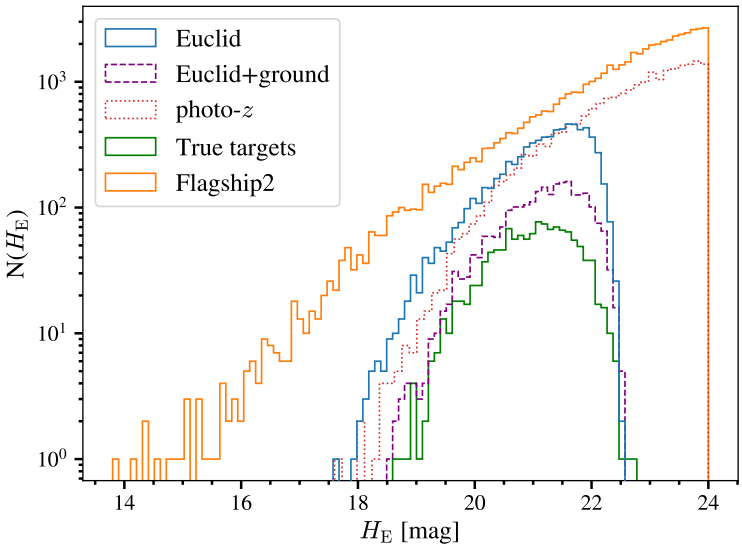

Figure 5.9: H_E-band number counts for samples built from the Flagship2 catalogue. The samples selected with the neural network classifier, using *Euclid* photometry only or combined with ground-based photometry are shown, respectively, by the blue-solid and magenta-dashed histograms. As indicated by the legend, the red-dotted histogram corresponds to a sample selected in redshift only, using *Euclid* photometric redshifts. The counts for the full Flagship2 catalogue and the true target sample are also shown for reference, by the orange and green histograms. Note how the distribution of the true targets (green histogram) dies off at magnitudes fainter than $H_E \simeq 22.5$. The targets in the EL-COSMOS catalogue extend to fainter flux.

redshift range or below the chosen Hα + [N II] flux limit, which are significantly reduced by adding the ground-based information (magenta dashed histogram). Note also how a selection over the target redshift range $[0.9, 1.8]$ using photometric redshifts, clearly does not effectively cut on the $H_{\rm E}$ magnitude, leaving a large population of faint objects. Nevertheless, we remind the reader that this discussion is specific to Flagship2. In the case of EL-COSMOS, also galaxies fainter than $H_{\rm E} \simeq 22.5$ are part of the target sample (see Fig. 5.4 and Table 5.2).

We can use the false positive rate (see Sect. 5.2.1) to interpret the origin of misclassified galaxies as a function of redshift and emission line flux. In the top panel of Fig. 5.10 this quantity is plotted as a function of redshift. The sample produced using *Euclid* photometry alone shows an excess of false positives at $z < 1$. This explicitly shows the inability with only the *Euclid* bands to properly exclude low-redshift galaxies, as well as some with flux below the flux limit. The addition of ground-based photometry effectively cures this, removing all galaxies at $z < 0.9$, leaving only a fraction of misidentified objects fainter than 2×10^{-16} erg s^{-1} cm^{-2} inside the target redshift range. It is interesting to note that the *Euclid*-only and the *Euclid* plus ground curves become indistinguishable at $z \gtrsim 1.4$. This is consistent with the redshift at which the 4000 Å break enters the $Y_{\rm E}$ band (at 9600 Å) and indicates that in this range the combination of $I_{\rm E}$ and $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$ provides, in general, sufficient spectral leverage to break degeneracies to both capture the correct redshift and identify emission line targets. As also shown, a photometric redshift selection is effective at removing low-redshift galaxies, but keeps in the sample all the low-flux galaxies (as is expected, since we are selecting on redshift alone).

In Fig 5.10 bottom panel, instead, we plot the false positive rate as a function of $f_{\rm H\alpha+[N\,II]}$. In this case, the peak and discontinuity evident at the flux limit, 2×10^{-16} erg s^{-1} cm^{-2}, is due to sources just below the flux limit, which enter the sample as false positives. Above the flux limit, instead, false positives arise from galaxies that are outside the redshift range. For this reason, the photo-z selection gives the lowest false positive rate, followed by the *Euclid* and ground-based classification. This does not tell the full story, however. The photo-z selection includes a number of false positives entering the sample at low fluxes, which are the cause of the very low precision shown by this selection. The classifier trained with ground-based photometry provides the best solution by balancing the two conditions of removing objects below the line flux limit and outside the redshift range.

Complementary, it is also interesting to look at the true negative rate (Eq. 5.6) of the whole selected sample, which gives an insight into the fraction of non-targets removed from the sample. When we select galaxies using *Euclid* photometry only, the true negative rate is 87%; the combination with ground-based data increases this metric up to 97%. Conversely, the true negative rate of the photo-z selection is 59%. Again, the better performance of the classifiers in comparison to the photo-z selection reflects the fact that the latter does not make a selection in the emission line limiting flux.

Finally, the machine learning algorithms identify regions in the full colour-magnitude space with a higher density of targets. In the case of the classifier trained on *Euclid* photometry, this is a four-dimensional space. In Appendix A.3 we present slices through the four-dimensional probability maps constructed from each classifier, showing how the selection depends on colour. It is interesting to visualise the boundaries constructed by each classifier. There is no visible separation between target and non-target galaxies in the colour planes and the classification algorithms define complex boundaries in the four-dimensional space. The support vector classifier and the neural network produce particularly smooth boundaries, while the self-organising map and tree-based classifiers

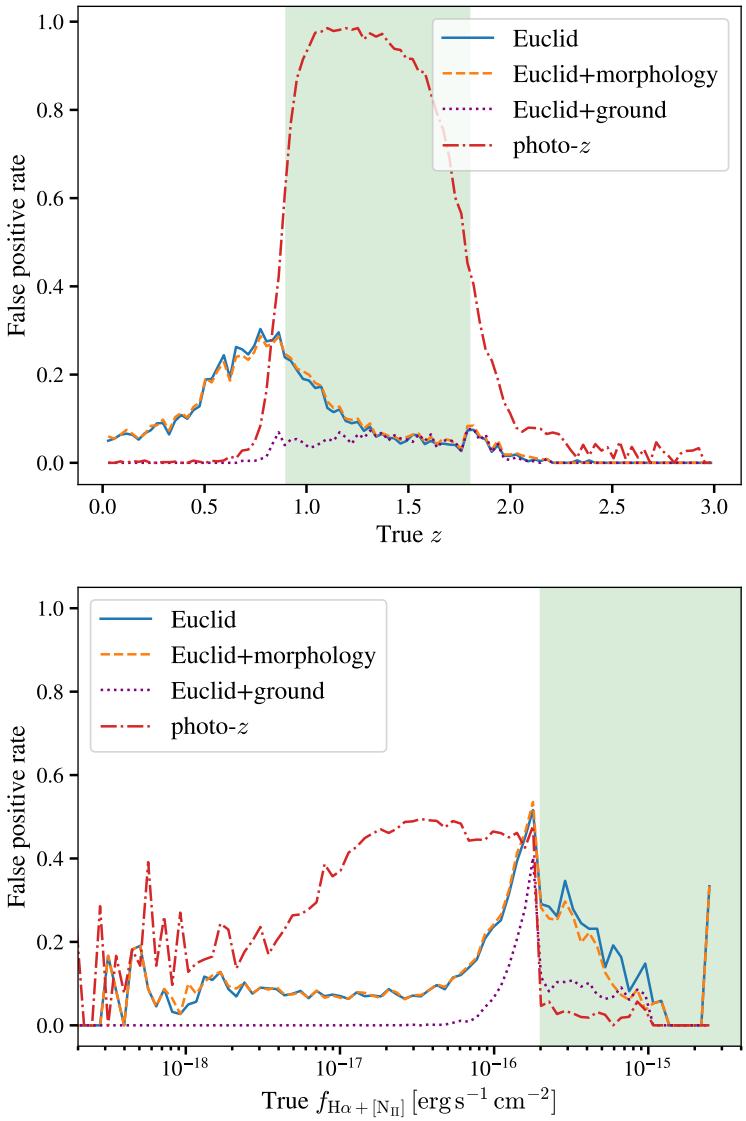

Figure 5.10: False positive rate as function of redshift and $f_{\mathrm{H}\alpha+[\mathrm{N\,II}]}$. The plot allows us to identify the origin of non-targets that enter the selected samples. The solid blue, dashed orange, and dotted purple curves, respectively, correspond to the neural networks trained with *Euclid* photometry, *Euclid* plus morphological data, and *Euclid* plus ground-based photometry. The dash-dotted red line is the false positive rate of the photo-z selection. *Top*: False positive rate as a function of z. The green shaded area marks the target redshift range. *Bottom*: False positive rate as a function of $f_{\mathrm{H}\alpha+[\mathrm{N\,II}]}$. The green shaded area corresponds to the Hα limiting flux. There is a peak in the false positive rate just below the flux limit used to define the target sample, although we note that these galaxies can still give correct redshift measurements.

do not. The irregular boundary is an indication that the classifier is overfitting the training set and will not generalise well. In addition, we verified that the 5% of the targets that we lose by imposing the 95% recall value are uniformly distributed in colour and are not part of any particular object class. We note that the lost targets are mainly faint objects.

5.5 Purity and completeness

The final purity of the spectroscopic sample will depend on the combination of the photometric information with the selection criteria applied to the spectroscopic measurements, as described by the flow diagram of Fig. 5.1. To provide a concrete, yet preliminary, example, we would like to quantify here the improvement in the final redshift purity and sample completeness produced by our photometric selection process. This work is based on a set of simulated spectra that were processed by the *Euclid* spectroscopic measurement pipeline (the SPE processing function). Although the simulated data were not yet fully realistic, they are nevertheless very useful for understanding how a machine learning-based photometric classification can aid in the sample selection. Also, the simulated spectra were built from the EL-COSMOS sample described in Sect. 5.3.1, which helps in making this test self-consistent. Two-dimensional spectral images were generated using the `FastSpec` code based on the spectral energy distribution and morphological parameters of the galaxies. These images were convolved with the NISP instrumental point spread function and realistic noise was added according to the detector model. Multiple exposures were simulated for each source and stacked with one to four exposures. One-dimensional spectra were extracted from the images and input to the *Euclid* spectroscopic measurement processing function to measure the redshift and spectral features.

The spectroscopic measurement pipeline carries out a likelihood analysis using spectral templates to estimate the redshift. It produces a probability distribution function of the redshift that is typically sharply peaked with a few primary redshift solutions. The integral of the peak provides a useful measure of the reliability of the solution. We vary the threshold in this reliability value to select spectroscopic samples and build the relationship between redshift purity and completeness, as shown by the SPE solid blue line in Fig. 5.11.

In the following discussion we focus on the results from the neural network classifier applied to the simulated spectroscopic sample. The purity and completeness values should be taken as indicative of the general trends and not as accurate forecasts of the pipeline performance. The values depend on the specific distribution of simulated sources and instrumental configuration. The target sample is defined as described in Sect. 5.3.3, using the total flux of the Hα and N II complex.

Figure 5.11 shows how redshift purity versus sample completeness plot improves when we complement the pure spectroscopic reliability cut selection (blue solid line) with increasing information provided by the photometric neural network classifier for the two configurations using *Euclid*-only or *Euclid* and ground-based photometry. The curve corresponding to the $H_{\rm E}$ magnitude-limit selection that gives 95% recall (see Table 5.2) is also plotted together with the curve corresponding to the colour selection presented in Sect. 5.4.1. These two curves visually overlap, but the colour cut curve (dash-dot-dotted purple line) is actually higher than the simple magnitude cut curve (dash-dotted orange line). This behaviour was expected as the two selections have very similar precision values at 95% recall and the colour cut has slightly better performance. The figure shows that in the range between 40% and 60% completeness, the photometric

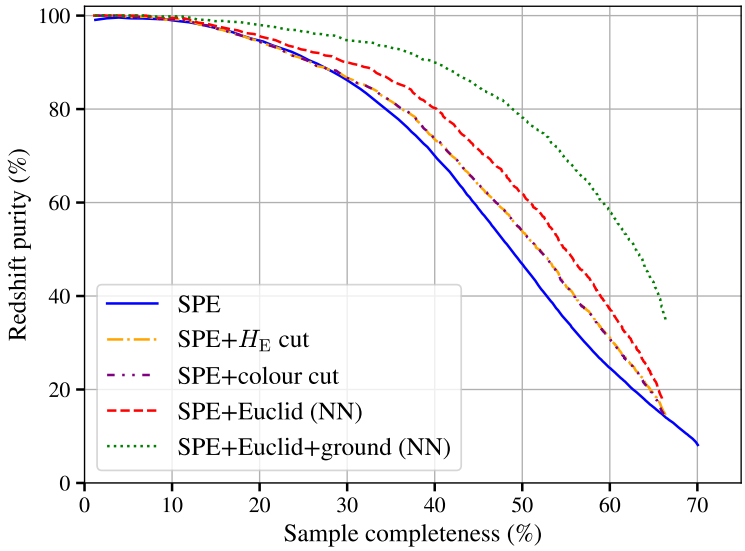

Figure 5.11: Redshift purity and sample completeness as a function of spectroscopic reliability threshold. The solid blue, dashed red, dotted green, and dash-dotted orange lines respectively correspond to a selection using only SPE reliability, SPE reliability combined with a photometric classification based on *Euclid* data, with the classification that uses *Euclid* and ground-based photometry, with the H_E magnitude limit selection, and with the colour selection in the $(I_E - H_E)$-H_E plane. In all cases, the recall of the photometric classification is set to 95%.

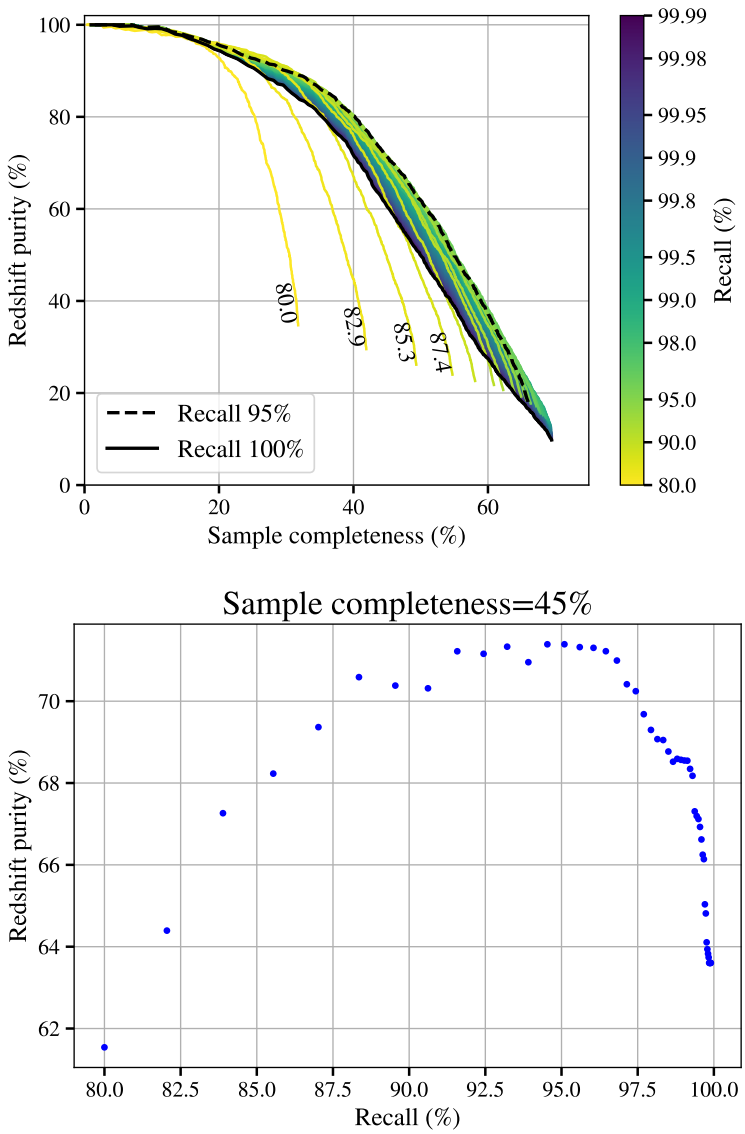

Figure 5.12: Spectroscopic redshift purity and completeness with the addition of the photometric classification. *Top*: the curves are colour-coded as a function of the recall of the photometric classification. The spectroscopic reliability threshold varies along each curve, while varying the threshold on the photometric classification probability shifts the curve. The purity improves as recall increases, reaching a maximum for recall $\sim$ 95% and declining after. For better visualisation, the first lines are labelled with the corresponding recall value. At recall values above 95% the curves are tightly packed. The solid black line corresponds to 100% recall, while the dashed line to 95% recall, the value we chose to benchmark our results. *Bottom*: redshift purity as a function of the recall of the photometric classification, fixing the value of sample completeness to 45%.

classification improves the redshift purity. For example, at a fixed value of 45% sample completeness, the classification based on *Euclid*-only bands improves the purity by $\sim 20\%$, when we add ground-based photometry the improvement rises to $\sim 45\%$. The simple $H_{\rm E}$ magnitude limit selection, at that same completeness value, gives an improvement of a few per cent only ($\lesssim 10\%$), evidencing the importance of exploiting all available photometric information.

To examine the effect of the photometric classification in more detail, in the top panel of Fig. 5.12 we show the redshift purity and sample completeness as a function of the reliability threshold imposed on the spectroscopic redshift measurement. The photometric classification has its own threshold parameter on the classification probability, which when combined with the spectroscopic selection, produces a family of curves. We label these curves based on their recall values. The bottom panel shows the dependence of redshift purity on the photometric selection recall, when the completeness is fixed to 45%. As we see, a recall value of 95% approximately maximises the purity-completeness curve, which justifies the choice made in Sect. 5.2.1. In addition, we verified that the 5% of the targets that we lose with the selection are uniformly distributed in colour and are not part of any particular object class. We note that the lost targets are mainly very faint objects.

The main conclusion from this exercise is that the impact of properly elaborated photometric information on the final purity and completeness of the *Euclid* spectroscopic sample is very significant, with a major improvement especially when ground-based visible bands are included. The precise gain, however, will depend on the galaxy distribution, the survey configuration and the instrument model.

5.6 Conclusions

We have investigated the benefits of combining photometric information with the spectroscopic measurement criteria for selecting *Euclid* spectroscopic samples. *Euclid* spectroscopy will give estimates of the galaxy redshifts, fluxes of the emission lines, and confidence intervals. However, since emission-line galaxies make up only a small fraction of the photometric sample, measurement noise can reduce the redshift purity and completeness of the sample and degrade the figure of merit for the galaxy clustering probe. The addition of photometric criteria in the selection can allow us to improve the purity of the sample by identifying sources that are likely to be bright emission-line galaxies at the target redshift.

To this end, we compared a set of machine learning classification algorithms with the aim of photometrically selecting emission-line target galaxies that are likely to give good redshift measurements in the Euclid Wide Survey. We used two catalogues to benchmark the classification performance, EL-COSMOS and Flagship2. Both catalogues have *Euclid* and ground-based simulated photometry. We produced noisy realisations of the catalogues assuming background-limited observations. The two catalogues yield similar results when using as input *Euclid*-only photometry, but when this is combined with ground-based data, the results using Flagship2 outperform those with EL-COSMOS. This is related to the differences in the Hα luminosity function and colour distribution of the two catalogues. In addition to these two configurations (*Euclid*-only and *Euclid* plus ground), we also considered adding morphological information (half-light radius and the axial ratio). We find that in general, while the addition of ground-based data strongly improves the precision (doubling it in the case of Flagship2), including morphological information (at least in the form provided here) gives negligible improvement.

The purity of the final spectroscopic sample will depend on the combination of the photometric classification with further selection criteria based on the properties of the spectroscopic data (see diagram in Fig. 5.1). To investigate this requires full end-to-end simulations of the spectroscopic reduction pipeline. We presented a preliminary exercise to assess the relative gain when the spectroscopic data are complemented by the photometric selection discussed here. This will be expanded in future work. We showed that in the range between 40% and 60% completeness, the purity is boosted by $\sim 20\%$ when using *Euclid*-only bands, and between 40% and 100% when including ground-based photometry. We consider this a remarkable indication.

The introduction of ground-based data significantly improves the purity of the sample, but in the practical application can also bring additional nuisance in the form of systematic errors. The ground-based photometry will come from multiple surveys and so will not be fully homogeneous. It will also suffer from additional selection effects correlated with the observing conditions that can propagate as systematic errors to the galaxy clustering measurements and cosmological constraints. Thus, the gains in purity from incorporating ground-based data must be carefully weighed against the potential of adding systematic errors, also considering the specific requirements of the science analysis to be carried out. We foresee that ground-based data may be used in analyses where a higher level of purity is desired, such as for studying the galaxy halo occupation distribution or galaxy evolution as a function of environment.

Photometric redshifts can also play a key role in sample selection. We used the *Euclid* photometric redshift estimates to select galaxies in the target redshift range and compared the performance of such a selection to that of the colour-based machine learning classifiers. Figure 5.10 shows that the photo-z selection is very efficient for redshift classification, especially to remove low redshift interlopers, but is not effective in identifying emission-line galaxies. Indeed, the photo-z selection has the highest fraction of false positives from faint galaxies with $f_{\mathrm{H}\alpha+[\mathrm{N\,II}]} < 2 \times 10^{-16}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$, but the lowest for bright ones with $f_{\mathrm{H}\alpha+[\mathrm{N\,II}]} > 2 \times 10^{-16}\,\mathrm{erg\,s^{-1}\,cm^{-2}}$, which means that it makes a better redshift selection than the algorithms presented in this work. Photometric redshifts could be used with additional constraints from spectral energy distribution fits to identify bright emission-line galaxy targets. In particular, the *Euclid* photometric redshift pipeline will output estimates of galaxy physical properties including the star-formation rate and dust attenuation, which will allow us to select emission-line galaxy samples. We expect that a classifier developed based on photometric redshifts and estimates of physical properties from spectral energy distribution fitting would perform similarly to the pure colour and magnitude-based classifiers that we tested, since the underlying photometric information is the same. Analogously, we expect a classifier trained to make a selection in redshift alone to perform similarly to the photo-z selection. Alternative classifiers that use the estimates of galaxy physical properties from the *Euclid* photometric redshift pipeline for sample selection will be investigated in a future work.

It is important to note that in this study some of the complications that will be present in real *Euclid* data were not considered. First, we assume an ideal training set, which is fully representative of the Wide Survey data. In the actual Euclid Wide Survey, the training set will come from the Deep Fields, which will total $\sim 50\,\mathrm{deg}^2$. Shallow and full-depth photometric measurements will be available for the *Euclid* photometry in the Deep Fields; however, we will only have the full-depth measurements for the ground-based photometry. As they are currently trained, the machine learning algorithms learn to classify the targets at a given noise level and it is not necessarily true that they will be able to generalise their results when trained and tested on samples with different noise levels. Therefore, if ground-based photometry is used, it will be necessary to degrade

the measurements to match the noise level in the Wide Survey. Since the ground-based photometry will come from multiple surveys, this operation will not be simple, and residual variations in homogeneity in the noise can lead to systematic variations in the classifier performance.

Moreover, the effective emission-line flux limit will vary across the Wide Survey due to foreground emission including zodiacal light and scattered stellar light (Euclid Collaboration: Scaramella et al., 2022). In this study, we used a fixed flux limit to build the training set of emission-line galaxies. In practice, this does not impose a sharp flux cut in the measured sample. However, when developing a classifier on real data, we will be able to use the Deep Survey to define the training set as the set of galaxies that are correctly measured by the *Euclid* pipeline, without imposing any specific constraints on their physical properties. It will also be possible to construct a classifier that accounts for variations in the noise level across the Wide Survey to optimise the sample.

Finally, a further complication that must be considered is contamination from stars in the galaxy catalogue that can impact the purity. The photometric classifier can be trained to maximise the precision in the presence of stars. This work will require us to incorporate a star-galaxy classifier, which is based both on size and photometric colours. Here, the morphological measurements will be important.

In the next stage of this work, we will consider the full set of spectroscopic and photometric selection criteria in order to compute the redshift purity, sample completeness and ultimately cosmologically relevant figures of merit. This requires running the spectroscopic reduction pipeline on mock data in order to produce end-to-end simulations. Such simulations will allow us to optimise the sample selection criteria, possibly with the use of machine learning classifiers. With *Euclid* observations began in fall 2023, we will be able to further tune the selection based on the actual telescope performance and ultimately construct the spectroscopic galaxy sample that will be used to test the cosmological model.

Part II

Cosmological parameter measurements

CHAPTER 6

Optimal constraints on Primordial non-Gaussianity with the eBOSS DR16 quasars in Fourier space

*The present chapter is based on the paper 'Optimal constraints on Primordial non-Gaussianity with the eBOSS DR16 quasars in Fourier space' by **Marina S. Cagliari**, Emanuele Castorina, Marco Bonici, and Davide Bianchi, accepted by Journal of Cosmology and Astroparticle Physics (Cagliari et al., 2023).*

6.1 Introduction and main results

The late-time distribution of the large-scale structure of the Universe is the result of the evolution, under gravitational interaction, of the set of primordial curvature perturbations. By measuring the n-point functions of a galaxy sample we have therefore the unique opportunity to test the statistical properties of the initial conditions of the Universe. Of particular relevance for LSS probes is the presence of possible primordial non-Gaussianities (PNG). The leading hypothesis for the dynamical generation of the primordial density fluctuations, Inflation (see Baumann, 2011, for a review), offers theoretical guidance to the most generic ways PNG could arise in cosmological correlators, also indicating that PNG are generically smaller than the dominant Gaussian term.

In this work we focus on the so-called local PNG, for which the primordial gravitational potential $\Phi_P(\mathbf{x})$ is a non-linear function of a Gaussian field φ, $\Phi_P = \varphi + f_{\rm NL}(\varphi^2 - \langle \varphi^2 \rangle)$. The amplitude of local PNG is parameterised by a single number $f_{\rm NL}$, and we immediately see that, if the primordial fluctuations are of $\mathcal{O}(10^{-5})$, local PNG are $\mathcal{O}(10^5)$ smaller than the Gaussian term for $f_{\rm NL} = 1$. Local PNG are among the most studied in the literature because they are exactly zero if the inflationary dynamics is driven by a single degree of freedom, the so-called single-field models (Maldacena, 2003; Creminelli & Zaldarriaga, 2004; Cabass et al., 2017). A robust detection of $f_{\rm NL}$ will therefore exclude all such models and point to a more complicated inflationary sector. Conversely, multi-field models of inflation generically predict $f_{\rm NL} \sim \mathcal{O}(1)$ (Senatore & Zaldarriaga, 2012; Alvarez et al., 2014), and could be severely constrained by a strong experimental bound. Measurements of the anisotropies of the cosmic microwave background from the *Planck* satellite put the stringent limit $f_{\rm NL} = 0.8 \pm 5$ (Planck Collaboration et al., 2020b), and upcoming instruments are expected to reduce this error bar by another 50% (Abazajian et al., 2016). Differently, than the CMB, which is sensitive to $f_{\rm NL}$ starting with the three-point function, LSS can probe PNG at the two-point, or power spectrum in Fourier space, level. As first pointed out in Dalal et al. (2008), the quadratic term in the definition of the primordial potential Φ_P induces a correlation between the long-wavelength gravitational field and the small-scale fluctuations. The latter could very well be in the range corresponding to the formation of halos and galaxies, whose number density is therefore

modulated by the large-scale value of Φ_P. Mathematically, we say that the large-scale bias of galaxies is modified in the presence of local PNG, and it reads

$$\delta_g(\mathbf{x}, z) = b(z)\,\delta_m(\mathbf{x}, z) + f_{\rm NL}\,b_\phi(z)\,\Phi_P(\mathbf{x}) \tag{6.1}$$

where δ_g is the galaxy density perturbation and b is the Gaussian linear bias. The new bias coefficient b_ϕ parameterises the actual response of small-scale fluctuations to the presence of local PNG, and it is subject to large theoretical uncertainties due to our incomplete knowledge of galaxy formation physics (Barreira, 2022a,b).[1] Via Einstein's Equations, the presence of Φ_P in the above expression implies that, on large scales, the power spectrum acquires a distinct k^{-2} feature, which is then interpreted as the smoking gun of local PNG. In this respect, knowing the value of b_ϕ is not a fundamental limitation, since what ultimately matters to exclude single field models is a detection rather than the actual value of $f_{\rm NL}$. The possibility to measure local PNG with the galaxy power spectrum has spurred a tremendous amount of research activities, and all major spectroscopic and photometric instruments like DESI (DESI Collaboration et al., 2016), Euclid (Laureijs et al., 2011), SPHEREx (Crill et al., 2020) and the Vera C. Rubin Observatory (LSST Science Collaboration et al., 2009) have the search for PNG as one of their primary science goals. It also serves as an important science case for future facilities (Achúcarro et al., 2022). Current LSS bounds are still far from the CMB one, $|f_{\rm NL}| \sim \mathcal{O}(20-30)$ (Castorina et al., 2019; Mueller et al., 2022; D'Amico et al., 2022; Cabass et al., 2022), but are expected to improve down to $\sigma_{f_{\rm NL}} \sim 1$ with current and future observations (Sailer et al., 2021; Cabass et al., 2023; Ansari et al., 2018; Bragança et al., 2023; Karagiannis et al., 2018).[2]

The main goal of this work is to provide the most stringent and robust constraints on local PNG with current data. We will use the extended Baryon Oscillation Spectroscopic Survey data release 16 quasar (QSO) sample (Ross et al., 2020; Lyke et al., 2020). Our analysis takes advantage of optimal signal weighting that maximises the response of a given galaxy sample to the presence of local PNG. These weights were first derived in Castorina et al. (2019), and are based on optimal quadratic estimators (Tegmark, 1997; Bond et al., 2000; Tegmark et al., 1998). The main reason to use optimal weights lies in Eq. (6.1): the non-Gaussian contribution is proportional to the primordial potential, and therefore it does not evolve over time, while the linear bias term is proportional to the matter density, which grows over time. This suggests that, in a given sample, high-redshift objects should be given more weight than low-redshift ones, since the Gaussian piece is smaller at earlier times. As we will see in Sect. 6.2 in more detail, the optimal analysis downweights the Gaussian signal by $w_0 \sim b(z)D(z)$, where $D(z)$ is the linear growth factor that decreases with increasing redshift, and upweights the PNG term by $\tilde{w} \sim b_\phi$.

The optimal redshift weighting therefore requires some prior knowledge of the response b_ϕ of a given sample to the presence of local PNG. For mass selected halos, analytical models (Slosar et al., 2008; Biagetti, 2019) and simulations (Biagetti et al., 2017; Barreira et al., 2020) suggest that $b_\phi \propto (b-p)$, where p is a number of $\mathcal{O}(1)$, is a very good approximation to the true response. As stated above, the picture is however much more complicated for observed galaxies. In this work, we will present constraints on $f_{\rm NL}$ for

[1] This point could, however, be turned the other way around and suggests that by carefully selecting the galaxy sample one could maximise the response, i.e. the value of b_ϕ, to improve the constraint (Castorina et al., 2018; Sullivan et al., 2023; Barreira & Krause, 2023).

[2] Recent work, see Rezaie et al. (2023), presents evidence of non zero $f_{\rm NL}$ at more than 99% confidence level (c.l.) with DESI imaging data. However, CMB and LSS measure local PNG on the same range of scales, which suggests a non-cosmological origin for the signal reported in Rezaie et al. (2023).

the two values of p mostly used in the literature, $p = 1.0$ and $p = 1.6$. The fact that the bound on local PNG depends on p could raise some concern about the robustness of our results. However, as we already pointed out, what matters is only a possible detection of $f_{\rm NL}$, not what the actual value is. Moreover, galaxy selection based on luminosity, colour, or magnitude, correlates reasonably well with host halo mass (Scoccimarro et al., 2001; Yuan et al., 2022), therefore we do not expect a large deviation from the values used in this work. Nevertheless, we will also show, in a novel application of our framework, how optimal signal weights can be used to put a data-driven prior on the value of b_ϕ or p. As mentioned above, the optimal weights $\tilde{w}$ are proportional to the response of the galaxy number density to the presence of local PNG. This implies that if the input value of b_ϕ we use for the weighting is very different from the true response, the optimal analysis will not improve the bound on $f_{\rm NL}$ over the un-weighted case, or will even worsen the constraints. As a first application of this idea, we will show in Sect. 6.4 that a large value of $p \gtrsim 3$ for the response of the QSOs in DR16 is not favoured by the data, without relying on any numerical simulations. We expect that as the uncertainty on $f_{\rm NL}$ reduces in the near future, our method will provide invaluable information on the most likely value of p to use in the data analysis. It should, however, be kept in mind that assuming a value of p implies that different data sets cannot be combined together or with the CMB. For this reason, we will also show, in Appendix B, constraints for $b_\phi f_{\rm NL}$ for all the data sets used in this work.

Our strongest bounds read

$$\begin{cases} -4 < f_{\rm NL} < 27\,, & 68\%\,\text{c.l.}\,, \\ -18 < f_{\rm NL} < 42\,, & 95\%\,\text{c.l.}\,, \end{cases} \quad \text{for } p = 1.0\,, \tag{6.2}$$

and

$$\begin{cases} -23 < f_{\rm NL} < 21\,, & 68\%\,\text{c.l.}\,, \\ -43 < f_{\rm NL} < 44\,, & 95\%\,\text{c.l.} \end{cases} \quad \text{for } p = 1.6\,, \tag{6.3}$$

which should be compared with a standard Feldman-Kaiser-Peacock (FKP; Feldman et al., 1994) analysis, see Fig. 6.8 and Table 6.2 for the full results. The optimal analysis improves by 10% and 30% over the FKP one for the $p = 1.0$ and $p = 1.6$ cases respectively. It is worth stressing that the power spectrum of DR16Q catalogue is, on all scales, dominated by the shot-noise, and we therefore did not expect much larger gains (Castorina et al., 2019). Our optimal constraints are robust to the treatment of systematic effects. The bounds using a linear method to remove known foregrounds are statistically indistinguishable from the ones obtained with a non-linear algorithm based on Neural-Network (NN; Rezaie et al., 2021). However, compared to the previous eBOSS data release (Castorina et al., 2019), the improvement in the constraint is smaller than what was expected from the increase in volume, and it is most likely due to the presence of residual foregrounds in the maps. This could also be the reason for the more limited improvement of the optimal analysis with respect to the standard one in comparison to the improvement found in DR14 (Castorina et al., 2019).

Our results are roughly comparable with the ones in Mueller et al. (2022), which also used the DR16Q data set. However, there are a number of important differences with our analysis. First, the weights employed in Mueller et al. (2022) are defined for pairs of galaxies, and cannot be automatically applied to individual galaxies, as relevant for a power spectrum analysis. To avoid imaginary weights for a single object, the Authors of Mueller et al. (2022) imposed, by hand, the positivity of the weights, which is not by itself an optimal procedure. In our case, the weights can very well be negative, precisely

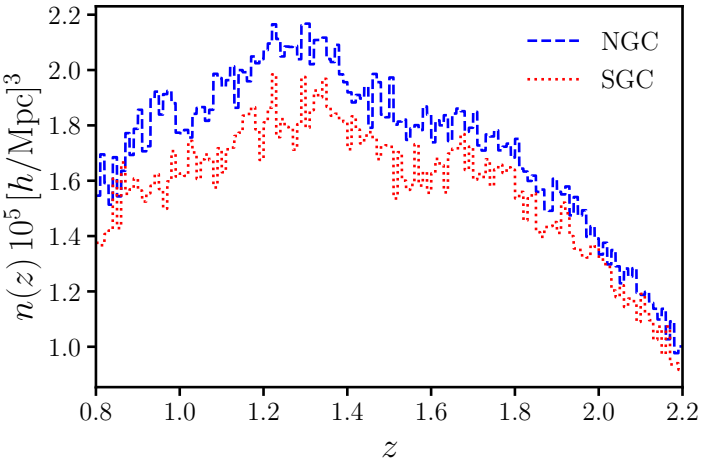

Figure 6.1: The quasar number density as a function of redshift for the DR16Q sample. The dashed blue line corresponds to the NGC and the dotted red line to the SGC. The SGC has a lower density in comparison to the NGC due to the difference in mean depth in the two regions (Ross et al., 2020).

in the region where the signal is: in this way the product of the weights times the signal contributes positively to the total signal-to-noise. More generally, cosmological information is contained in the galaxy fields and not in its non-linear transformations, like for example pairs of galaxies. The other main difference with the work of Mueller et al. (2022) is in the modelling of the signal. As discussed in Sect. 6.2.4, we think the bound of Mueller et al. (2022) is artificially tighter due to an incorrect choice for the effective redshift, $z_{\rm eff}$, at which the theoretical model is evaluated. For most applications the precise definition of $z_{\rm eff}$ does not matter, but it becomes important in searches for local PNG, where the signal is proportional to $b_\phi(z) \sim b(z) - p$. For the DR16Q analysis in Mueller et al. (2022), a too-high value of $z_{\rm eff}$ results in a higher linear bias $b(z)$, which artificially reduces the uncertainty on $f_{\rm NL}$ to keep the product $b_\phi f_{\rm NL} \sim$ constant. In Sect. 6.2.4 we will also clarify on this issue and on what the more accurate definition of $z_{\rm eff}$ is.

The rest of this chapter is organised as follows: in Sect. 6.2 we present the DR16Q data set and the measurements of the power spectrum; in Sect. 6.3 we discuss the modelling of the QSO power spectrum, its convolution with the window function and the definition of the effective redshift; in Sect. 6.4 we present and discuss the constraints on $f_{\rm NL}$; Sect. 6.5 concludes and summarises our results.

All the codes, scripts, measurements and Monte Carlo Markov chains used in this work are freely accessible at `github.com/mcagliari/eBOSS-DR16-QSO-OQE`.

6.2 Data

6.2.1 The eBOSS QSO sample

In this work, we use the eBOSS DR16Q sample (Ross et al., 2020; Lyke et al., 2020). As part of the SDSS-IV experiment (Blanton et al., 2017), the eBOSS data were acquired at the Apache Point Observatory in New Mexico.

The DR16Q sample contains $343\,708$ quasars in the redshift range $0.8 < z < 2.2$. The sample is divided into two fields of view, the North Galactic cap (NGC), which covers an area of $2924\,\mathrm{deg}^2$, and the South Galactic cap (SGC), with an area of $1884\,\mathrm{deg}^2$; in comparison to Data Release 14 (DR14) the area is approximately doubled. The whole sample has a volume of $\sim 20\,(\mathrm{Gpc}/h)^3$. The NGC has a mean density of $n \approx 1.8 \times 10^{-5}\,(\mathrm{Mpc}/h)^{-3}$, while the SGC has a slightly lower density of $n \approx 1.6 \times 10^{-5}\,(\mathrm{Mpc}/h)^{-3}$. The number densities as a function of the redshift of the NGC and the SGC quasars are shown in Fig. 6.1. The number density of SGC is about 10% lower than the NGC number density because of the lower mean depth of the survey in the SGC region. The data of the North and South Galactic cap were released in two separate catalogues, each one with the corresponding random catalogue. The random catalogues are 50 times more dense than the data catalogues, and their redshift distributions are produced by sampling from the observed data redshifts (Ross et al., 2020), a procedure known as shuffling. The use of the shuffling scheme to produce the random catalogues introduces a systematic effect called radial integral constraint (RIC; de Mattia & Ruhlmann-Kleider, 2019). We discuss how to estimate and correct for the RIC effect in Sect. 6.3.2.

Both the data and random catalogues contain three weights for each data point. First, there are the close pairs weights, w_{cp}, which take into account fibre collisions. The second weights are related to the spectroscopic completeness, w_{noz}, and correct for the expected redshift failure rate. Third, there are the imaging systematic weights, w_{sys}. These weights correct for systematic effects at large angular scales, and they are therefore especially important for f_{NL} measurements. In the official data release of eBOSS DR16 (Ross et al., 2020), these weights are computed with linear regression of the imaging properties and Galactic foregrounds. Additional catalogues were released for the quasars (Rezaie et al., 2021). In these catalogues, the imaging systematic weights were computed using neural networks. Neural networks are able to approximate non-linear functions, hence they could in principle produce a better correction than the weights computed with the linear regression. NN methods will however remove part of the signal as well, and could bias negative the constraint on local PNG. Hereafter we will refer to the official DR16Q catalogues as the linear weight catalogue, and to the catalogues with NN systematic weights as NN weight catalogue. The completeness weight contribution to any data point is (Ross et al., 2020)

$$w_{\mathrm{c}} = w_{\mathrm{cp}}\, w_{\mathrm{noz}}\, w_{\mathrm{sys}}\,, \tag{6.4}$$

where w_{sys} can either be from the linear weight catalogues or the NN weight catalogue. The same weighting procedure of Eq. (6.4) applies to the objects in the random catalogues.

6.2.2 Mocks

A set of 1000 synthetic clustering catalogues for each Galactic cap (Zhao et al., 2021) was simulated using the effective Zel'dovich approximation mock method (EZmock; Chuang et al., 2015). The EZmock catalogues were produced assuming a flat ΛCDM cosmology with $\Omega_m = 0.307115$, $\Omega_\Lambda = 0.62885$, $\Omega_b = 0.048206$, $h = 0.6777$, $\sigma_8 = 0.8225$, $n_s = 0.9611$, and $f_{\mathrm{NL}} = 0$. The mocks reproduce the two and three-point clustering statistics of DR16Q.

In the official release of the eBOSS DR16 EZmock catalogues, three sets of mock catalogues are provided. Each set of EZmocks consists of 1000 pairs of data and random catalogues. We refer to the first set of EZmocks as EZmock *realistic*, since the data and random catalogues of this set contain all the known observational systematic effects. Each data catalogue has a corresponding random catalogue, whose redshift distribution

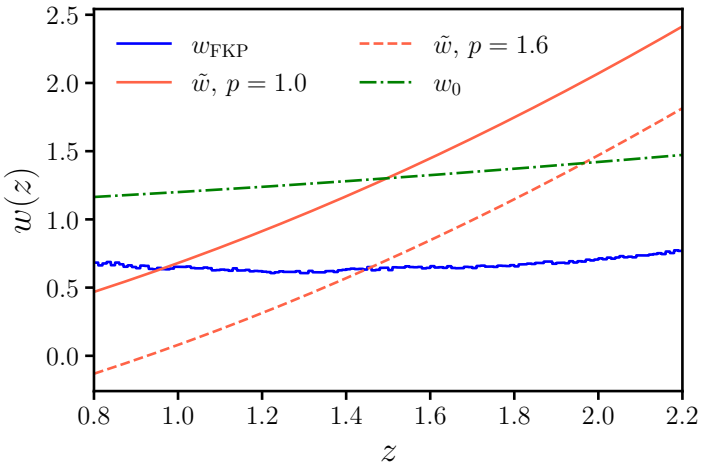

Figure 6.2: Weights as a function of redshift for the eBOSS DR16Q. The solid blue line corresponds to the FKP weights, Eq. (6.6); its almost flat behaviour as a function of redshift is due to $n(z)\,P_{\rm fid} \ll 1$. The optimal weights to estimate $f_{\rm NL}$ are shown in red and dot-dashed green. The solid and dashed red lines respectively correspond to $\tilde{w}(z)$ for $p = 1.0$ and $p = 1.6$. These weights have a clear dependence on the quasar response to the $f_{\rm NL}$ signal.

is produced by shuffling the redshift of the data catalogues. The 1000 data-random pairs of the EZmock realistic set are used to estimate the covariance matrix, $\mathbf{\Sigma}$. The imaging systematic weights of the realistic EZmock are computed only with the linear regression method and not with the neural network. The covariance matrix will thus only contain information about the linear imaging weights. The other two sets are the EZmocks *complete* and the EZmocks *shuffled*. These two sets share the same data catalogues, which do not have any observational systematic effect, while the random catalogues redshift distributions were produced in different ways. In the EZmock shuffled set each data catalogue has a corresponding random catalogue the redshift distribution of which is produced with the shuffling scheme. The EZmock complete set only has two random catalogues, one for each Galactic cap. The redshift distributions of these random catalogues were sampled from the same $n(z)$ interpolation used for the EZmock data catalogues. We use the EZmock complete and shuffled sets to estimate the RIC effect.

6.2.3 Power spectrum estimation

As shown in Castorina et al. (2019), the optimal power spectrum is the cross-correlation of two different fields produced by weighting the underlying catalogues. To arrive at the power spectrum estimator we start with the two quasar density fields (Feldman et al., 1994),

$$\begin{aligned} \tilde{F}(\mathbf{r}) &= \tilde{w}_{\rm tot}\,\left[w_{\rm c}^{\rm qso}\,n_{\rm qso}(\mathbf{r}) - \alpha_{\rm s} w_{\rm c}^{\rm s}\,n_{\rm s}(\mathbf{r})\right]\,, \\ F_0(\mathbf{r}) &= w_{\rm tot,\,0}\,\left[w_{\rm c}^{\rm qso}\,n_{\rm qso}(\mathbf{r}) - \alpha_{\rm s} w_{\rm c}^{\rm s}\,n_{\rm s}(\mathbf{r})\right]\,, \end{aligned} \tag{6.5}$$

where $n_{\rm qso}$ and $n_{\rm s}$ respectively are the number density of the quasar sample and the corresponding random catalogues, $w_{\rm c}^{\rm qso}$ and $w_{\rm c}^{\rm s}$ are the completeness weights from Eq. (6.4)

for the quasars and the randoms. The total weights, $\tilde{w}_{\rm tot}$ and $w_{\rm tot,\,0}$, are the product of the FKP weights (Feldman et al., 1994),

$$w_{\rm FKP}(z) = \frac{1}{1+\bar{n}(z)\,P_{\rm fid}}\,, \tag{6.6}$$

and the optimal weights for a $f_{\rm NL}$ measurements with power spectrum data (Castorina et al., 2019),

$$\tilde{w}(z) = b(z) - p\,, \quad w_0(z) = D(z)\,\left(b(z) + \frac{f(z)}{3}\right)\,. \tag{6.7}$$

Therefore the total weights read (Castorina et al., 2019)

$$\tilde{w}_{\rm tot}(z) = w_{\rm FKP}(z)\,\tilde{w}(z)\,, \quad w_{\rm tot,\,0}(z) = w_{\rm FKP}(z)\,w_0(z)\,. \tag{6.8}$$

In Eq. (6.6), $\bar{n}(z)$ is the mean density as a function of redshift, and $P_{\rm fid} = 3{\times}10^4\,({\rm Mpc}/h)^3$, which corresponds to the expected power on the scales affected by PNG in the sample. In Eq. (6.7), $b(z)$ is the fiducial value of the QSO bias model (Laurent et al., 2017),

$$b(z) = 0.278\,\left((1+z)^2 - 6.565\right) + 2.393\,, \tag{6.9}$$

and $D(z)$ and $f(z)$ are respectively the growth factor and growth rate as functions of redshift. As mentioned in Sect. 6.1, in this work we use $p = 1.0$ and $p = 1.6$. The dependence on the redshift of the weights defined in Eq. (6.6) and Eq. (6.7) is plotted in Fig. 6.2. It shows the difference between the FKP weighting scheme, which is almost constant in redshift, as $n(z)P_{\rm fid} \ll 1$, and the optimal weights, which have a strong dependence on redshift. Finally, the factor $\alpha_{\rm s}$ in Eq. (6.6) is defined as

$$\alpha_{\rm s} = \frac{\sum^{\rm qso} w_{\rm c}}{\sum^{\rm s} w_{\rm c}}\,, \tag{6.10}$$

and it properly normalises the number density of the random catalogues.

Following ref. (Yamamoto et al., 2006), we write the monopole of the cross-correlation between the two weighted fields in Eq. (6.5) as

$$\widetilde{P}_0(k) = A_0^{-1}\,\int\frac{{\rm d}\Omega_k}{4\pi}\left[\int{\rm d}\mathbf{r}_1\,\tilde{F}(\mathbf{r}_1)\,e^{i\mathbf{k}\cdot\mathbf{r}_1}\,\int{\rm d}\mathbf{r}_2\,F_0(\mathbf{r}_2)\,e^{-i\mathbf{k}\cdot\mathbf{r}_2}\mathcal{L}_0(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}_2)\right] - S_0\,, \tag{6.11}$$

where $\mathcal{L}_0$ is the first Legendre polynomial. The normalisation factor A_0 and the shot noise contribution, S_0, are respectively defined as

$$A_0 = \int{\rm d}\mathbf{r}\,w_{\rm tot,\,0}(\mathbf{r})\,\tilde{w}(\mathbf{r})\,\left[w_{\rm c}\,n_{\rm qso}(\mathbf{r})\right]^2\,, \tag{6.12}$$

$$S_0 = A_0^{-1}\,\int{\rm d}\mathbf{r}\,w_{\rm c}^2(\mathbf{r})\,n_{\rm qso}(\mathbf{r})\,\left(1+\alpha_{\rm s}\right)\,w_{\rm tot,\,0}(\mathbf{r})\,\tilde{w}(\mathbf{r})\,\mathcal{L}_0(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})\,. \tag{6.13}$$

We calculate the monopole of the power spectrum using `nbodykit` (Hand et al., 2018), which implements Eq. (6.11) as follows (Bianchi et al., 2015; Hand et al., 2017),

$$\widetilde{P}_0(k) = A_0^{-1}\,\int\frac{{\rm d}\Omega_k}{4\pi}\tilde{F}(\mathbf{k})\,F_0(-\mathbf{k})\,, \tag{6.14}$$

with

$$F_0(\mathbf{k}) = \int \mathrm{d}\mathbf{r}\, F_0(\mathbf{r})\, e^{i\mathbf{k}\cdot\mathbf{r}} \mathcal{L}_0(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) \\ = 4\pi\, Y_{00}(\hat{\mathbf{k}}) \int \mathrm{d}\mathbf{r}\, F_0(\mathbf{r})\, Y^*_{00}(\hat{\mathbf{r}})\, e^{i\mathbf{k}\cdot\mathbf{r}}\,, \tag{6.15}$$

where Y_{00} is the first spherical harmonic. The normalisation and the shot noise are computed as discrete sums over the quasars and the randoms. The normalisation is (Feldman et al., 1994)

$$A_0 = \alpha_\mathrm{s} \sum_i^{N_\mathrm{s}} n_\mathrm{s}(\mathbf{r}_i)\, w_\mathrm{c}(\mathbf{r}_i)\, w_{\mathrm{tot},0}(\mathbf{r}_i)\, \tilde{w}_\mathrm{tot}(\mathbf{r}_i)\,, \tag{6.16}$$

while the shot noise contribution becomes (Feldman et al., 1994)

$$S_0 = A_0^{-1} \left[\sum_i^{N_\mathrm{qso}} w_\mathrm{c}^2(\mathbf{r}_i)\, w_{\mathrm{tot},0}(\mathbf{r}_i)\, \tilde{w}_\mathrm{tot}(\mathbf{r}_i) + \alpha_\mathrm{s}^2 \sum_i^{N_\mathrm{s}} w_\mathrm{c}^2(\mathbf{r}_i)\, w_{\mathrm{tot},0}(\mathbf{r}_i)\, \tilde{w}_\mathrm{tot}(\mathbf{r}_i) \right]. \tag{6.17}$$

To compute the power spectrum estimator in Eq. (6.14) we use a mesh of 512^3 cells. The quasars and the random objects are projected onto the mesh using a triangular-shaped cloud interpolation (Hockney & Eastwood, 1981). In this interpolation each quasar and random is weighted by both its completeness and total weight and we assume Planck (Planck Collaboration et al., 2020a) as fiducial cosmology. The power spectrum is estimated on a logarithmic grid from $k_\mathrm{min} = 3.75 \times 10^{-3}\,(\mathrm{Mpc}/h)^{-1}$ to $k_\mathrm{max} = 2.23 \times 10^{-1}\,(\mathrm{Mpc}/h)^{-1}$ for NGC, and $k_\mathrm{max} = 2.78 \times 10^{-1}\,(\mathrm{Mpc}/h)^{-1}$ for SGC.

A final remark about the normalisation of the power spectrum estimator in Eq. (6.14) and the shot noise contribution in Eq. (6.17). Since Eq. (6.16) is just an approximation of the exact definition in Eq. (6.12), it has been pointed out that this could lead to biased constraints on cosmological parameters (de Mattia et al., 2021). We, therefore, decided to re-normalise the measured power spectra by the limit, at small separation, of the monopole of the window function, $Q_0(0)$, see Sect. 6.2.4. The final estimator of the power spectrum is

$$\hat{P}_0(k) = \frac{A_0}{Q_0(0)} \left(\widetilde{P}_0(k) - S_0 \right). \tag{6.18}$$

We note that we used, both for the linear and NN catalogues, the value of $Q_0(0)$ computed from the randoms of the linear catalogues, since the two catalogues are expected to have the same response at small scales.

In Fig. 6.3 we show the NGC and SGC (left and right columns) observed power spectra of the linear and NN catalogues (dots and squares). The three rows correspond to the different weights used to estimate the $\hat{P}_0(k)$, from top to bottom the rows correspond to the FKP weights, the optimal weight with $p = 1.0$, and with $p = 1.6$. We also plotted the best-fit model for the linear (dashed line) and NN (dash-dotted line) catalogues (see Sect. 6.4). The best fit of the FKP weight case is for the model with $p = 1.6$. For the NN catalogues power spectra, the band powers have been shifted by 5% along the k-axis for better visualisation of the corresponding error bars. The observed power spectra estimated from the two catalogues differ only in the first two bins. The NN catalogues' power spectra have less power in the first bin than their linear counterparts. The excess of power in the linear catalogues' power spectra is expected to be related to large-scale

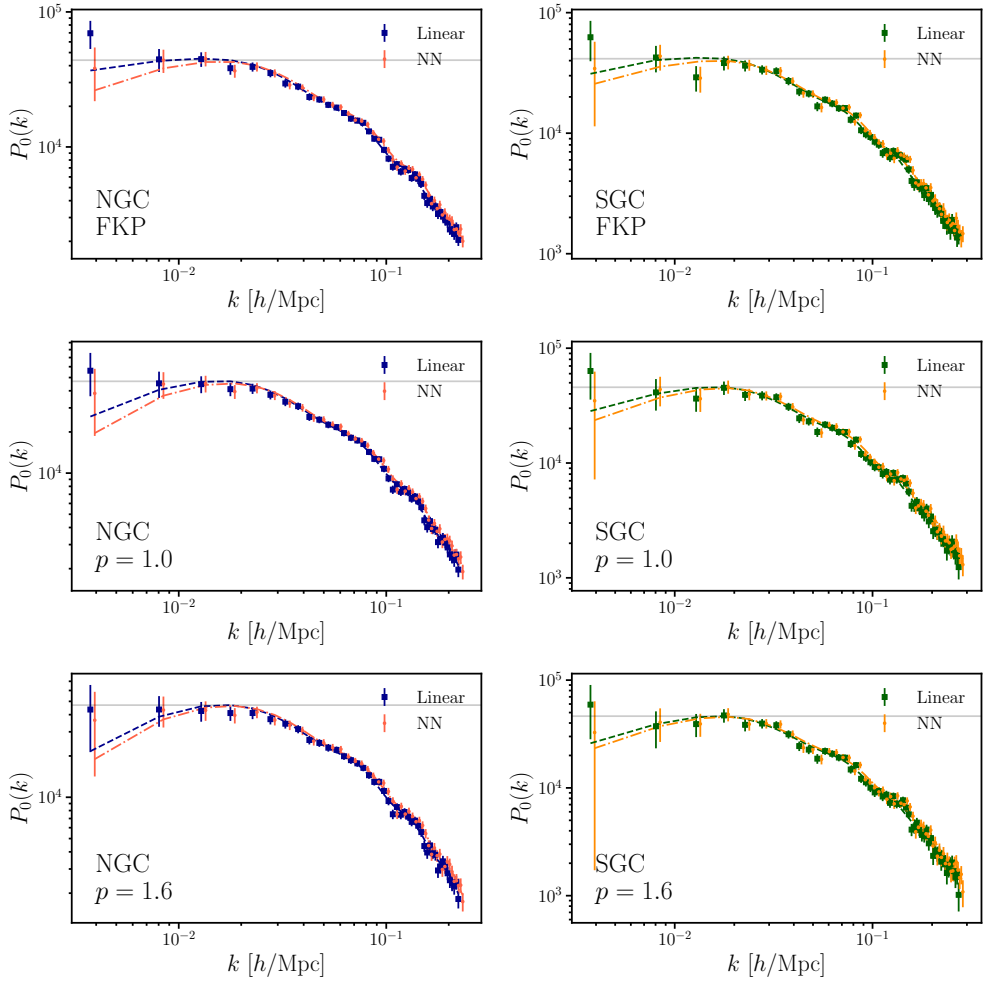

Figure 6.3: Observed power spectra and best-fit model for the NGC (right column), and SGC (left column), and for the three weighting scheme: FKP (top row), optimal weights with $p = 1.0$ (middle row), and with $p = 1.6$ (bottom row). For the NGC (SGC) blue (green) dots and the dashed line correspond to the linear catalogues results, red (yellow) squares and dash-dotted line to the NN catalogues. The horizontal grey line marks the turn-around point in the power spectrum. To better visualise the power spectra error bars, the NN $P(k)$ was shifted by 5%.

systematic effects that the linear regression weights are not able to correct. The solid horizontal line in each panel shows the amplitude of the power spectrum at its peak, and it serves to guide the eye to the fact that the optimal weighted measurements are larger than the corresponding FKP ones. This was expected because the optimal weights upweight high redshift galaxies (see Fig. 6.2), which have a higher bias and therefore a larger clustering amplitude.

6.2.4 Window functions

The window function, $W(\mathbf{s})$, represents the footprint on the sky and the redshift selection function of the survey. It is an essential ingredient to compare a power spectrum model with the observed power spectrum.

In order to evaluate the model of the observed power spectrum we need the multipoles of the window function (see Sect. 6.3.2), defined as follows,

$$Q_\ell(s) \equiv (2\ell+1) \int \mathrm{d}\Omega_s \int \mathrm{d}^3\mathbf{s}_1\, W(\mathbf{s}_1)\, W(\mathbf{s}+\mathbf{s}_1)\, \mathcal{L}_\ell(\hat{\mathbf{s}}_1 \cdot \hat{\mathbf{s}}) \equiv \int \mathrm{d}s_1\, s_1^2\, Q_\ell(s; s_1)\,. \quad (6.19)$$

To compute the window function multipoles we use the pair-counting approach introduced in Wilson et al. (2017). First, with `nbodykit` we calculate the weighted pair counts of the random catalogues as a function of the three-dimensional separation and the cosine of the line-of-sight angle, $RR^{\mathrm{w}}(s,\mu)$. That is done by cross-correlating the random catalogues weighted by $w_{\mathrm{c}}\,\tilde{w}_{\mathrm{tot}}$, and the random catalogues weighted by $w_{\mathrm{c}}\,w_{\mathrm{tot},\,0}$. Second, we compute its multipoles,

$$RR^{\mathrm{w}}_\ell(s) = (2\ell+1) \int \mathrm{d}\mu\, RR^{\mathrm{w}}(s,\mu)\, \mathcal{L}_\ell(\mu)\,. \quad (6.20)$$

Finally, in order to obtain the window function multipoles the quantity above needs to be normalised to take into account the width of the shell over which the pair counting is performed and the density of the random catalogues in comparison to the data catalogues. The window function multipole ℓ is finally defined as

$$Q_\ell(s) = \frac{RR^{\mathrm{w}}_\ell(s)}{4\pi\, s^3\, \mathrm{d}\ln s} \frac{\left(\sum^{\mathrm{qso}} w_{\mathrm{c}}\right)^2 - \sum^{\mathrm{qso}} w_{\mathrm{c}}^2}{\left(\sum^{\mathrm{s}} w_{\mathrm{c}}\right)^2 - \sum^{\mathrm{s}} w_{\mathrm{c}}^2}\,, \quad (6.21)$$

with $\mathrm{d}\ln s = \frac{s_{n+1}-s_n}{s}$, where s is the centre of the n-th separation bin. We stress that each set of optimal weights requires its own multipoles of the window function $Q_\ell(s)$.

To reduce the computational time of the pair counting algorithm we divided the random catalogues into five subsets and computed $RR^{\mathrm{w}}(s,\mu)$ for each of them. These subsets are 10 times denser than the data catalogues. We calculated the window function multipoles of each random subset using Eq. (6.21), with the caveat that the sum over the random $\sum^{\mathrm{s}}$ is now over the subset. The final $Q_\ell(s)$ is the mean of the five subsets.

6.3 Analysis methods

6.3.1 The power spectrum Model

To model the quasar power spectrum in redshift-space we use linear theory. Linear theory is enough to make the prediction for two reasons: first, the local f_{NL} signal is at low k, where structures are still growing with a linear regime. Second, the smaller scales

of this sample are dominated by redshift errors, which dominate over the non-linearities. We write the power spectrum model as follows (**?**),

$$P_{\rm qso}(k,\mu;z) = G(k,\mu;\sigma_{\rm FoG})^2 \, \left[b_{\rm tot}(k;z) + f(z)\,\mu^2\right]^2 \, P_m(k;z) + N \,, \tag{6.22}$$

where P_m is the matter power spectrum in real-space, N is the residual shot noise free parameter, and $f(z)$ is the growth rate. The total quasar bias includes PNG contribution (Dalal et al., 2008; Slosar et al., 2008),

$$b_{\rm tot}(k;z) = b_1 + \Delta b = b_1 + f_{\rm NL}\,(b_1 - p)\,\tilde{\alpha}(k;z) \,, \tag{6.23}$$

where b_1 is the quasar linear bias, and $\tilde{\alpha}(k;z)$ is

$$\tilde{\alpha}(k;z) = \frac{3\,\Omega_m\,H_0^2\,\delta_{\rm c}}{c^2\,k^2\,T(k)\,D(z)} \,. \tag{6.24}$$

In Eq. (6.24), $\delta_{\rm c} = 1.686$ is the critical density in the spherical collapse in an Einstein-De Sitter Universe, Ω_m is the matter density parameter, and H_0 the Hubble parameter, both at $z = 0$, and c is the speed of light. Then, $T(k)$ is the matter transfer function normalised to 1 at low-k, and $D(z)$ is the growth factor normalised to $(1+z)^{-1}$ in the matter-dominated era. Finally, the damping of the power spectrum due to nonlinear redshift-space distortions is included with a Lorentzian function,

$$G(k,\mu;\sigma_{\rm FoG}) = \left[1 + \frac{(k\,\mu\,\sigma_{\rm FoG})^2}{2}\right]^{-1} \,, \tag{6.25}$$

where $\sigma_{\rm FoG}$ accounts for both the typical velocity dispersion of QSOs, as well as their redshift error, which is estimated to be $\sigma_z = 300\,{\rm km\,s^{-1}}$ for DR16Q, with no significant dependence on the redshift (Lyke et al., 2020).

The quasar power spectrum multipoles are then easily computed

$$P_{\ell,\,\rm qso}(k;z) = \frac{2\ell+1}{2} \int_{-1}^{1} {\rm d}\mu \, P_{\rm qso}(k,\mu;z)\,\mathcal{L}_\ell(\mu) \,. \tag{6.26}$$

To evaluate the cosmological quantities in Eq. (6.22) and Eq. (6.24) we assume a Planck fiducial cosmology (Planck Collaboration et al., 2020a), and fix the redshift to an effective value. In the following section, we discuss how the effective redshift is defined and computed. We calculate the cosmological functions with `classy`, the Python wrapper of the `CLASS` CMB Boltzmann solver (Blas et al., 2011).

6.3.2 Convolution with the window function and the effective redshift

The ensemble average of the power spectrum estimator in Eq. (6.11) is (Wilson et al., 2017; Beutler et al., 2019)

$$\langle \hat{P}_0(k) \rangle = \sum_{\ell,\,L} \begin{pmatrix} \ell & L & 0 \\ 0 & 0 & 0 \end{pmatrix}^2 \int {\rm d}s\, s^2 j_0(ks) \int {\rm d}s_1\, s_1^2\, \xi_\ell(s;s_1(z))\, Q_L(s;s_1(z)) \,, \tag{6.27}$$

where $Q_L(s;s_1)$ is defined in Eq. (6.19), $\xi_\ell(s;s_1)$ is the multipole ℓ of the QSO correlation function, $j_A(ks)$ is the spherical Bessel function of order A, and $\left(\begin{smallmatrix} \ell & L & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$ is a Wigner 3-j symbol.

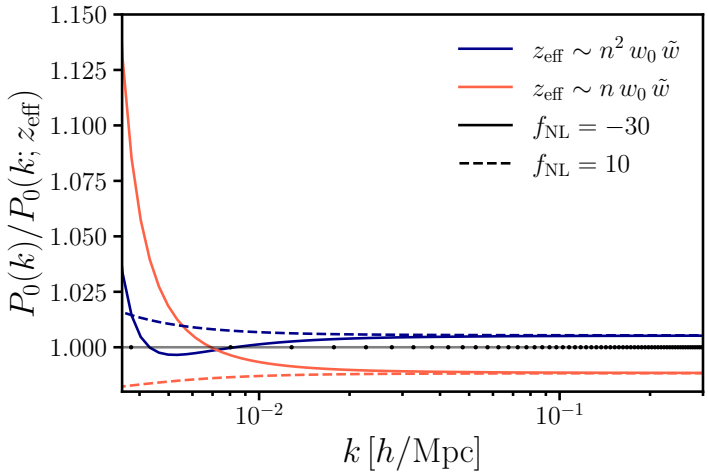

Figure 6.4: The accuracy of the effective redshift approximation. In red, and for two different values of $f_{\rm NL}$, the plot shows the ratio between the theoretical model integrated over the full radial selection function of the DR16Q sample and the same model evaluated at a $z_{\rm eff}$ defined in Eq. (6.30), see Table 6.1. The accuracy of the other most common definition of $z_{\rm eff}$ in the literature is shown with blue lines. The black points correspond to the effective wavenumbers of the measurements of the QSO power spectrum.

	FKP	$p = 1.0$	$p = 1.6$
NGC	1.49	1.65	1.76
SGC	1.50	1.66	1.76

Table 6.1: The effective redshift for the different weights and the two sky regions. The optimal weights increase the $z_{\rm eff}$ of the sample for $p = 1.0$ and $p = 1.6$.

In Eq. (6.27) the redshift evolution of the signal is taken into account by $s_1(z)$ in $\xi_\ell(s; s_1)$, which should be integrated against the $Q_L(s; s_1)$ for a proper model comparison. This could be a time-consuming step if repeated for every point in the parameter space exploration, and it is therefore often approximated. Noticing that in most applications the correlation function is factorisable in time and space, $\xi(s, z) \sim g(s)\, h(z)$, a possible way to speed up the computation of the theoretical model is to separate the integrals over s and s_1. An even more useful approximation is to assume that the model can be evaluated only at some effective redshift, $z_{\rm eff}$, defined by the radial selection function. The expression for the power spectrum then simplifies to

$$P_0(k; z_{\rm eff}) = \sum_{\ell,\, L} \begin{pmatrix} \ell & L & 0 \\ 0 & 0 & 0 \end{pmatrix}^2 \int {\rm d}s\, s^2\, j_0(ks)\, \xi_\ell(s; z_{\rm eff})\, Q_L(s)\,, \tag{6.28}$$

with the multipole of the correlation function of the random catalogues $Q_L(s)$ defined in Eq. (6.19). The final integral is a simple Hankel transform that can be computed quite efficiently Wilson et al. (2017).

The question becomes, then, what is the most accurate definition of $z_{\rm eff}$. The estimator of the power spectrum in Eq. (6.11), and therefore the multiples of the window

functions as well, contains two powers of the radial selection function, which suggests the following definition of $z_{\rm eff}$ for a sample with given $n(z)$ and weights $w(z)$,

$$z_{\rm eff} = \frac{\int {\rm d}z\, n(z)^2\, [\chi(z)^2/H(z)]\, w(z)^2\, z}{\int {\rm d}z\, n(z)^2\, [\chi(z)^2/H(z)]\, w(z)^2}\,, \tag{6.29}$$

where $\chi(z)$ is the comoving distance and $H(z)$ is the Hubble parameter. In practice, the integral above can be estimated via Monte Carlo methods as

$$z_{\rm eff} = \frac{\sum^{\rm qso} z\, n(z)\, w_{\rm c}^2\, w_{\rm FKP}(z)^2\, \tilde{w}(z)\, w_0(z)}{\sum^{\rm qso} n(z)\, w_{\rm c}^2\, w_{\rm FKP}(z)^2\, \tilde{w}(z)\, w_0(z)}\,, \tag{6.30}$$

or with the analogous expression written in terms of the random catalogues. The values of $z_{\rm eff}$ for the samples used in this work are shown in Table 6.1. We see that the optimal weighting scheme increases $z_{\rm eff}$, since it up weights high redshift objects, which have a higher response to the presence of PNG.

The accuracy of our definition of $z_{\rm eff}$ is presented in Fig. 6.4, with the blue lines showing the ratio between the power spectrum model fully integrated over redshift and the monopole evaluated at $z_{\rm eff}$. The dashed line corresponds to $f_{\rm NL} = 10$, while the continuous one to $f_{\rm NL} = -30$, with $p = 1.6$ in both cases. The black points on the horizontal axis show the effective values of the wavenumbers of the measurements. We find that our approximation is sub-percent accurate at high-k, and better than 2.5% accurate on very large scales, thus much smaller than the sample variance of the measurements.

On the other hand, the DR16Q analysis of Mueller et al. (2022) adopts a definition of $z_{\rm eff}$ with one less power of $n(z)$ than Eq. (6.30).[3] This choice produces the red set of curves in Fig. 6.4, which we find are more than a per cent off at high-k, a number that could become significant over many data points, and more than 10% inaccurate at large scales. In particular, Mueller et al. (2022) reports $z_{\rm eff} = 1.83$ for the weights optimised with $p = 1.6$, a number significantly higher than our $z_{\rm eff} = 1.76$. At fixed value of $f_{\rm NL}$, the product $b_\phi\, f_{\rm NL}$ is 10% larger at $z = 1.83$ than at $z = 1.76$. This suggests that the authors of Mueller et al. (2022) would have at least gotten a 10% weaker constraint on $f_{\rm NL}$, had they used the more accurate definition of $z_{\rm eff}$ in Eq. (6.29).

Finally, we write the convolution of the window function in the more convenient form

$$P_0(k; z_{\rm eff}) = \sum_{\ell,\, L} i^\ell \begin{pmatrix} \ell & L & 0 \\ 0 & 0 & 0 \end{pmatrix}^2 \int \frac{{\rm d}q}{2\pi^2}\, q^2\, P_{\ell,\,{\rm qso}}(q; z_{\rm eff}) \int {\rm d}s\, s^2 j_0(ks)\, j_\ell(qs)\, Q_L(s) \tag{6.31}$$

$$= \sum_{\ell,\, L} i^\ell \begin{pmatrix} \ell & L & 0 \\ 0 & 0 & 0 \end{pmatrix}^2 \int \frac{{\rm d}q}{2\pi^2}\, q^2\, P_{\ell,\,{\rm qso}}(q; z_{\rm eff})\, Q_{\ell,\,L}(k, q)\,, \tag{6.32}$$

where we defined

$$Q_{\ell,\,L}(k, q) = \int {\rm d}s\, s^2 j_0(ks)\, j_\ell(qs)\, Q_L(s)\,. \tag{6.33}$$

The integral in Eq. (6.32) is then evaluated as a simple matrix multiplication. This choice allows us to never compute the correlation function multipoles, which are formally divergent in the presence of local PNG.

[3] The published version of Mueller et al. (2022) contains the following definition of the effective redshift, right below their Eq. (11), $z_{\rm eff} = \sum_i z_i\, w_{\rm tot} / \sum_i w_{\rm tot}$, where $w_{\rm tot}^2 = w_{\rm FKP}^2\, w_{\rm c}^2\, |\tilde{w}\, w_0|$. This is a typo, as the analysis of Mueller et al. (2022) actually used $z_{\rm eff} = \sum_i z_i\, w_{\rm tot}^2 / \sum_i w_{\rm tot}^2$, corresponding to the red set of curves in fig. 6.4. We thank Eva-Maria Mueller for the correspondence about this point.

The multipoles of the window function corresponding to the optimal weights with $p = 1.0$ are shown in Fig. 6.5, top panel. In our model, we use only the even multipoles up to $\ell = 4$, and neglect possible odd ones (Beutler et al., 2019).[4]

6.3.3 The integral constraint

The final step to model the observed power spectrum is to correct the convolved power spectrum with the integral constraint effects. Integral constraint effects arise when the survey selection function is estimated from the data themselves (de Mattia & Ruhlmann-Kleider, 2019). Fluctuations over the whole survey average to zero when the observed galaxy mean density is used as the true cosmological mean. This causes a suppression of power at large scales, which we call global integral constraint (GIC; Peacock & Nicholson, 1991; Wilson et al., 2017). On the other hand, an additional radial integral constraint is produced when the radial $n(z)$ is inferred from the data (de Mattia & Ruhlmann-Kleider, 2019). That is the case for eBOSS data, where the random catalogues redshift distribution is obtained by shuffling the data redshift distribution. The RIC causes a suppression of the large-scale fluctuations along the line-of-sight.

The global integral constraint depends on the Hankel transform, $|\tilde{W}_\ell(k)|^2$, of the window function multipole $Q_\ell(s)$. The Hankel transforms are normalised so that $|\tilde{W}_0(0)|^2 = 1$ (Wilson et al., 2017). To correct for the RIC we need to estimate the effect that the shuffling of the random produces on the measured power spectrum. To do so we used both the complete and shuffled EZmocks. First, we compute the mean of the power spectra of the complete EZmocks, $\bar{P}_{\rm c}(k)$, and the mean of the power spectra of the shuffled EZmock, $\bar{P}_{\rm r}(k)$. Then, the radial integral constraint correction is defined via

$$W_{\rm RIC}(k) = \frac{\bar{P}_{\rm c}(k) - \bar{P}_{\rm r}(k)}{\bar{P}_{\rm c}(k)} \,. \tag{6.34}$$

Given $|W(k)|^2$ and $W_{\rm RIC}(k)$ the final expression for the power spectrum is (Wilson et al., 2017; Mueller et al., 2022)

$$P_0^{\rm IC}(k) = P_0(k) - P_0(0)\,|W(k)|^2 - P_0(k)\,W_{\rm RIC}(k)\,, \tag{6.35}$$

where we dropped the dependence on the effective redshift for brevity. In Fig. 6.5 bottom panel we present the effect of the different components of the observed power spectrum model and compare it with the mean power spectrum of the EZmock realistic catalogues. First, the plot shows the importance of the window convolution (dotted red line), which removes power at $k \lesssim 5 \times 10^{-2}\,({\rm Mpc}/h)^{-1}$. Second, the integral constraint correction (dash-dotted green line) only affects the first bin, but it is well within the 68% error bars.

6.3.4 Parameter estimation

To estimate the posterior distribution of the parameters $\boldsymbol{\theta}$ of our model, $\mathbf{T}(\boldsymbol{\theta})$, given our data vector $\mathbf{D}$, we assume a multi-variate Gaussian likelihood,

$$\mathcal{L}(\mathbf{D}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}\sum_{ij}(D_i - T_i(\boldsymbol{\theta}))\,\Sigma_{ij}^{-1}(D_j - T_j(\boldsymbol{\theta}))\right) . \tag{6.36}$$

[4]Wide angle effects and other projection effects are negligible for the DR16Q volume (Castorina & White, 2018a,b; Castorina & Di Dio, 2022; Beutler et al., 2019).

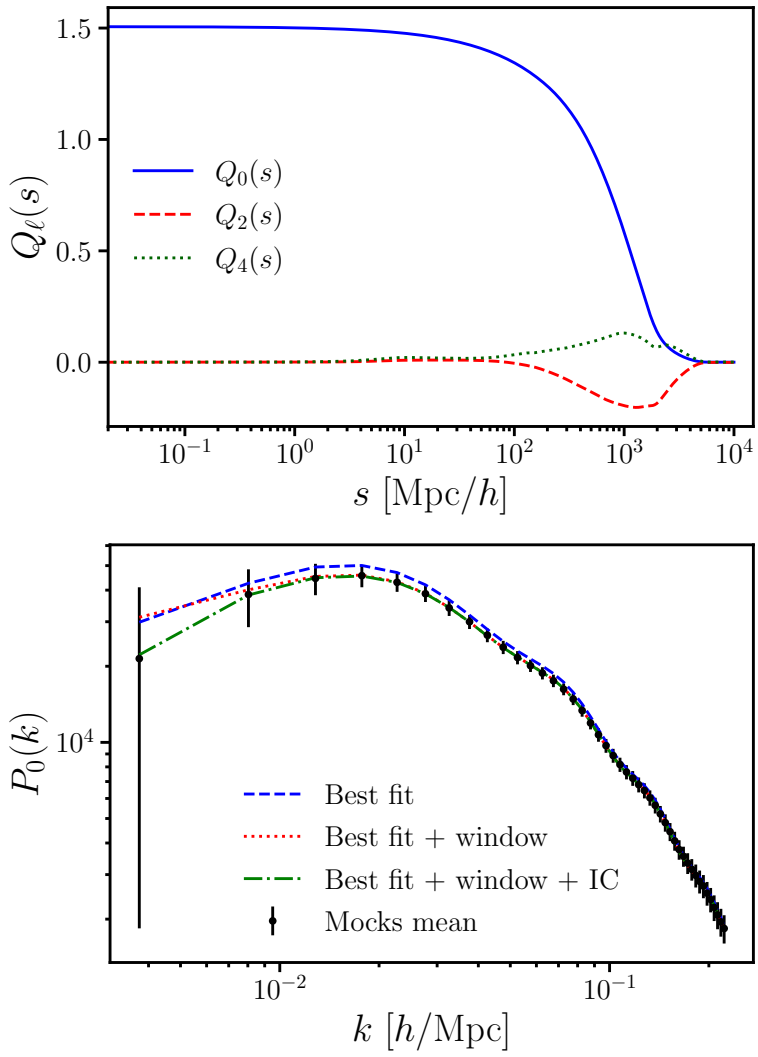

Figure 6.5: *Top:* multipoles of the window function in configuration space. The blue solid line, red dashed, and dotted green respectively are the monopole, quadrupole, and hexadecapole of the window function. *Bottom*: the effect of the different components of the observed power spectrum model compared to the mean power spectrum of the EZmock realistic catalogues (black circles with error bars). The blue dashed line is the monopole of the best-fit model, the dotted red line is the best-fit model convolved with the window function, and the dash-dotted green line is the observed power spectrum corrected with the integral constraint as in Eq. (6.35).

The model parameters $\boldsymbol{\theta}$ are three for each field of view: $\sigma_{\rm FoG}$, b_1 and N (see Sect. 6.3.1), and $f_{\rm NL}$. The rest of the cosmology is fixed to the Planck best-fit values (Planck Collaboration et al., 2020a). Therefore, when fitting the data of one sky patch (single field analysis) the total number of free parameters is four, and when fitting the two fields of view data (joint analysis) the free parameters are seven. In the case of the joint analysis $f_{\rm NL}$ is common for the two fields, while the other three parameters of the model are unique for each field of view, for a total of six parameters. We assume a uniform prior distribution for all the parameters, with the following bounds

$$\begin{aligned} f_{\rm NL} &\in [-500, 500]\,, \\ b_1 &\in [0.1, 6]\,, \\ \sigma_{\rm FoG} &\in [0, 20]\,, \\ N &\in [-5000, 5000]\,. \end{aligned} \tag{6.37}$$

In Eq. (6.36), $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix estimated with the EZmock realistic catalogues (see Sect. 6.2.2). As a finite number of mocks, $N_{\rm m} = 1000$, is used to estimate the covariance matrix, its inverse, the precision matrix, is biased (Hartlap et al., 2007). This bias is corrected by re-scaling the covariance matrix with the inverse of the Hartlap factor,

$$\boldsymbol{\Sigma}' = \frac{N_{\rm m} - 1}{N_{\rm m} - N_{\rm b} - 2}\,\boldsymbol{\Sigma}\,, \tag{6.38}$$

where $N_{\rm b}$ is the number of k-bins in the observed power spectrum. This correction is $\sim 5\%$ for the NGC and $\sim 6\%$ for the SGC.

The Monte Carlo Markov chain (MCMC) algorithm employed in this analysis is the Hamiltonian Monte Carlo (HMC; Duane et al., 1987; Neal, 2011; Betancourt, 2017). HMC is a sampling technique that combines principles from Hamiltonian mechanics and MCMC methods to efficiently explore high-dimensional parameter spaces. Unlike traditional methods, HMC employs a dynamic integration of the target probability distribution. By introducing auxiliary momentum variables, HMC maps the trajectory of the walkers, which explore the posterior distribution, into a Hamiltonian system, whose potential energy is given by minus the logarithm of the joint-likelihood. Then, HMC exploits the gradients of the distribution to follow the Hamiltonian trajectories and efficiently sample the posterior even in a high-dimensional space. In this work, we use the no-U-turn sampler (NUTS; Hoffman & Gelman, 2011; Ge et al., 2018) implementation of HMC, which is able to automatically adapt critical parameters like the step size and the trajectory length. By setting the acceptance rate to 0.9, the chains quickly converge, within a few thousand steps, to $R - 1 \lesssim 10^{-3}$, where R is the Gelman-Rubin statistics (Gelman & Rubin, 1992).

6.4 Constraints and discussion

In this section we present and discuss the constraints on $f_{\rm NL}$ we obtained with the analyses of DR16Q. Figure 6.3 shows the measured data points and error bars of the monopole of the power spectrum with the best-fit model of the joint analyses. The plots are presented for the two sky regions, NGC (left column) and SGC (right column), and the three weighting schemes, the standard FKP with a model assuming $p = 1.6$ (top row), the optimal weights for $p = 1.0$ (middle row) and $p = 1.6$ (bottom row).

Figures 6.6 and 6.7 show the two-dimensional posterior of the joint analyses for the linear and NN catalogue, respectively. Both figures are organised as follows: the left

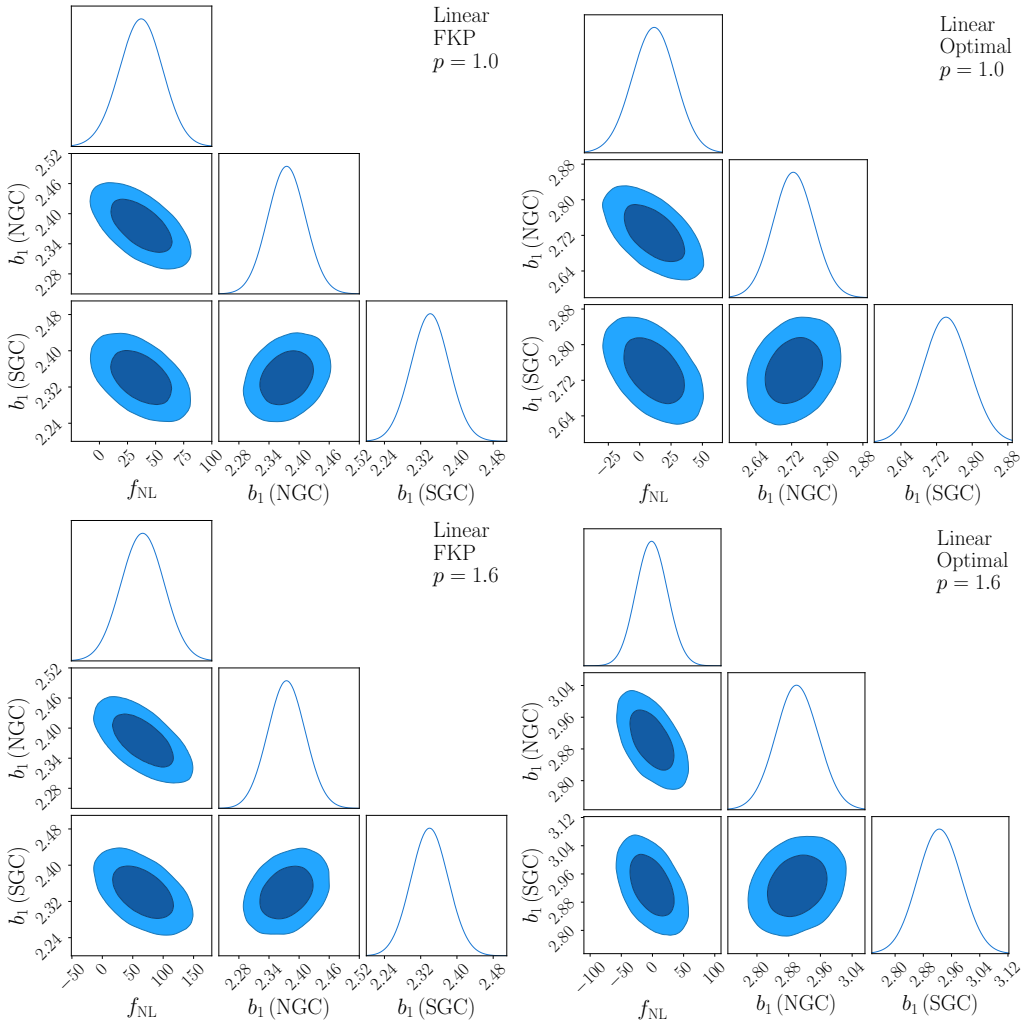

Figure 6.6: Two-dimensional posterior distributions for $f_{\rm NL}$ and the quasar bias, b_1, of NGC and SGC from the joint analysis of the linear catalogues. The plots on the top correspond to $p = 1.0$, and the bottom plots to $p = 1.6$. On the left are the results for the FKP weighting scheme and on the right for the optimal weights.

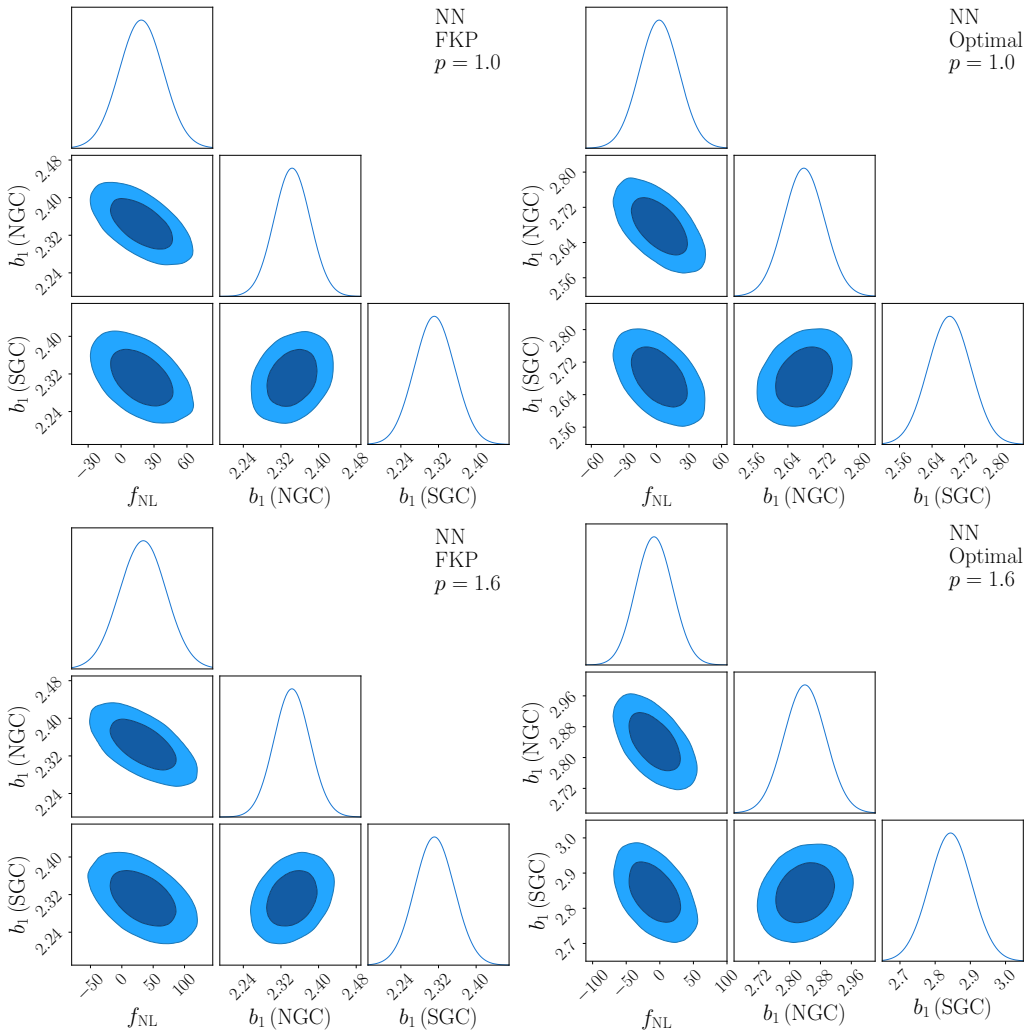

Figure 6.7: Same as figure 6.6, but for the NN catalogues.

	p		C.L.	Linear	NN	eBOSS DR14Q
joint	1.0	FKP	68%	$19 < f_{\rm NL} < 53$	$0 < f_{\rm NL} < 36$	
			95%	$4 < f_{\rm NL} < 71$	$-16 < f_{\rm NL} < 54$	$-39 < f_{\rm NL} < 41$
		Optimal	68%	$-4 < f_{\rm NL} < 27$	$-14 < f_{\rm NL} < 19$	
			95%	$-18 < f_{\rm NL} < 42$	$-30 < f_{\rm NL} < 34$	$-51 < f_{\rm NL} < 21$
	1.6	FKP	68%	$34 < f_{\rm NL} < 97$	$0 < f_{\rm NL} < 66$	
			95%	$6 < f_{\rm NL} < 129$	$-32 < f_{\rm NL} < 98$	$-74 < f_{\rm NL} < 81$
		Optimal	68%	$-23 < f_{\rm NL} < 21$	$-33 < f_{\rm NL} < 15$	
			95%	$-43 < f_{\rm NL} < 44$	$-54 < f_{\rm NL} < 40$	$-81 < f_{\rm NL} < 26$
NGC	1.0	FKP	68%	$26 < f_{\rm NL} < 67$	$1 < f_{\rm NL} < 44$	
			95%	$6 < f_{\rm NL} < 87$	$-23 < f_{\rm NL} < 63$	$-34 < f_{\rm NL} < 61$
		Optimal	68%	$-6 < f_{\rm NL} < 34$	$-20 < f_{\rm NL} < 21$	
			95%	$-27 < f_{\rm NL} < 52$	$-42 < f_{\rm NL} < 41$	$-56 < f_{\rm NL} < 38$
	1.6	FKP	68%	$49 < f_{\rm NL} < 125$	$-2 < f_{\rm NL} < 78$	
			95%	$10 < f_{\rm NL} < 159$	$-38 < f_{\rm NL} < 121$	$-67 < f_{\rm NL} < 112$
		Optimal	68%	$-39 < f_{\rm NL} < 28$	$-53 < f_{\rm NL} < 17$	
			95%	$-80 < f_{\rm NL} < 54$	$-95 < f_{\rm NL} < 47$	$-87 < f_{\rm NL} < 42$
SGC	1.0	FKP	68%	$-15 < f_{\rm NL} < 41$	$-17 < f_{\rm NL} < 45$	
			95%	$-36 < f_{\rm NL} < 72$	$-43 < f_{\rm NL} < 78$	$-64 < f_{\rm NL} < 31$
		Optimal	68%	$-17 < f_{\rm NL} < 29$	$-21 < f_{\rm NL} < 32$	
			95%	$-35 < f_{\rm NL} < 55$	$-40 < f_{\rm NL} < 63$	$-61 < f_{\rm NL} < 26$
	1.6	FKP	68%	$-22 < f_{\rm NL} < 78$	$-34 < f_{\rm NL} < 81$	
			95%	$-61 < f_{\rm NL} < 135$	$-82 < f_{\rm NL} < 146$	$-122 < f_{\rm NL} < 63$
		Optimal	68%	$-28 < f_{\rm NL} < 36$	$-37 < f_{\rm NL} < 40$	
			95%	$-51 < f_{\rm NL} < 74$	$-66 < f_{\rm NL} < 84$	$-92 < f_{\rm NL} < 42$

Table 6.2: Summary of the $f_{\rm NL}$ 68% and 95% constraints of this work. The results for NGC, SGC and the joint analysis are presented, and compared with the eBOSS DR14Q constraints (Castorina et al., 2019).

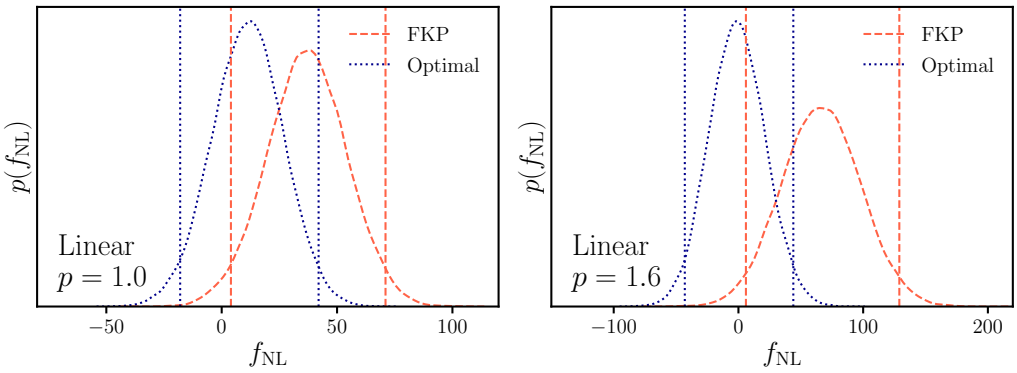

Figure 6.8: One dimensional posterior distribution for $f_{\rm NL}$ for the joint analysis of the linear catalogues. The dashed red curve is the posterior distribution obtained with the FKP weight analysis and the dotted blue curve with the optimal weights analysis. The vertical lines mark the corresponding 95% constraints. On the left, the results for $p = 1.0$, and on the right for $p = 1.6$.

column corresponds to the analysis of the power spectrum monopoles measured with the standard FKP weights and the right column to the analysis of the optimally weighted power spectrum; the top row presents the results for $p = 1.0$, and the bottom row to $p = 1.6$. In the plots, we show three of the seven fit parameters: $f_{\rm NL}$, and the linear bias b_1 of the two sky caps. In all the analyses there is almost no correlation between the biases of the two sky regions, which was expected as they correspond to independent fields of view. Moreover, the linear biases of the two Galactic caps are always consistent with each other, and the bias estimated with the optimal weights is larger than the bias estimated using the standard FKP weights. The reason behind the latter behaviour is the higher effective redshift of the sample when using the optimal weighting scheme, as discussed in Sect. 6.2.3. Another effect visible in Figs. 6.6 and 6.7 is how the correlation between the linear bias and $f_{\rm NL}$ changes between the FKP weights and the optimal weights. Even though this effect is present for both p values it is more evident in the case with $p = 1.6$. The optimal weights reduce the correlation between $f_{\rm NL}$ and the linear bias. We observe this same behaviour for the other fit parameters.

Figure 6.8 presents the comparison of the one-dimensional $f_{\rm NL}$ posterior distributions obtained with the joint analyses of the linear catalogues. The left panel corresponds to the case with $p = 1.0$ and the right panel to $p = 1.6$. In both panels, the red dashed line is the $f_{\rm NL}$ posterior of the FKP analysis and the dotted blue line represents the posterior of the optimal weight analysis. The corresponding vertical lines mark the 95% constraints. For both values of p the 95% constraints estimated with the standard FKP weights do not contain $f_{\rm NL} = 0$. Given that CMB, which measures $f_{\rm NL} = 0.8 \pm 5$, and LSS probe the primordial power spectrum over the same range of scales, the results suggest the presence of residual contamination in the FKP catalogues produced with the linear systematic weights. The optimal weights shift the posterior to values more consistent with $f_{\rm NL} = 0$. Serendipitously, this suggests that higher redshifts QSOs in the samples might be less affected by systematic effects. Nevertheless, the most important difference between the standard FKP and optimal weights is that the optimal weights give tighter constraints. For $p = 1.0$ the optimal weights improve the 95% constraint of about 10%, for $p = 1.6$ this improvement is a little less than 30%. The comparison between

the optimally weighted and the unweighted constraints is difficult, due to the large systematic effects still present in the FKP catalogues, especially the NGC ones. At 95% c.l. the joint analysis for $p = 1.0$ does not contain $f_{\rm NL} = 0$ for the linear catalogues, and it barely contains it at the 68% c.l. for the NN ones, see Table 6.2. Nevertheless, the larger improvements with $p = 1.6$ than with $p = 1.0$ point in the direction of $b_\phi \sim b_1 - 1.6$.

The 68% and 95% constraints on $f_{\rm NL}$ are written down in Table 6.2, where we also make a comparison with the results of Castorina et al. (2019) on the eBOSS DR14Q data. We always get tighter constraints with the linear catalogues rather than with the NN catalogues. The improvement in the constraints using the optimal weights in comparison to the standard FKP ones is the same for the two Galactic caps. On the other hand, the improvements with respect to the eBOSS DR14Q analysis are between 10% and 20% for the two catalogues and are smaller than the one expected by the doubling of the survey volume. This can be attributed to large-scale systematic effects that are still present in the sample. In particular, the constraints from the single field analyses show smaller improvements with respect to the eBOSS DR14Q sample than the joint analysis. In the case of SGC, the constraints are even worse than in the older ones. Nevertheless, the single-field analyses give posterior distributions consistent with each other and their combination in the joint analysis produces the tightest bounds. The best constraints of this work are from the joint analysis of the linear catalogues with the optimal weights. The 95% constraints for $p = 1.0$ are $-18 < f_{\rm NL} < 42$ corresponding to $\sigma_{f_{\rm NL}} \sim 15$ and for $p = 1.6$ they are $-43 < f_{\rm NL} < 44$ with $\sigma_{f_{\rm NL}} \sim 22$.

We also repeated the parameters estimation assuming a value of $p = 3.0$ in the optimal weights. In this case, we expect a worse constraint on $f_{\rm NL}$ from both the FKP and the optimal analysis compared to the cases discussed above. However, as discussed in Sect. 6.1, we can use these constraints to show how the optimal analysis can provide a data-driven estimate of the value of p. If the value of p used in the optimal weights is not close to the true one, then the optimal analysis will not improve over the standard case or will improve less than an analysis with a value of $\tilde{w} \sim b_\phi$ closer to the actual response. Note that this approach assumes that the local PNG signal we are looking for is non-zero. In the case of $p = 3.0$, the NGC analysis always shows evidence for non-zero $f_{\rm NL}$ at the 95% c.l., thus we focus on SGC. For the linear catalogue, the constraints are $-140 < f_{\rm NL} < 81$ for the FKP weights and $-129 < f_{\rm NL} < 76$ for the optimal analysis. This improvement by 8% should be compared to the 20% and 56% reduction of the error bar for $p = 1.0$ and $p = 1.6$ respectively in the SGC analysis. This implies that larger values of $p \gtrsim 3$ are disfavoured for this sample.

6.5 Conclusions

In this work, we presented the most stringent constraint on the amplitude of local Primordial Non-Gaussianities with Large-Scale Structure data, in particular with the eBOSS DR16Q data set. Assuming the QSOs response to $f_{\rm NL}$ is proportional to $b_1 - p$, where b_1 is the linear bias, our strongest bounds read

$$\begin{aligned} -4 < f_{\rm NL} < 27\,, \quad \text{for } p = 1.0\,, \\ -23 < f_{\rm NL} < 21\,, \quad \text{for } p = 1.6\,, \end{aligned} \tag{6.39}$$

at 68% c.l..

Our goal was to show that the optimal signal weighting reduces the error bars on $f_{\rm NL}$ compared to standard analysis, and we robustly find improvement between 10-30% depending on the analysis setup. While our optimal constraints are always consistent

with no local PNG, the comparison with previous eBOSS data releases does not allow us to exclude the presence of residual systematic effects in the data. Nevertheless, the DR16Q catalogues and the analysis presented here represent an important step forward in the direction of robust and optimal analysis of PNG with LSS data. We have also shown how optimal weights could provide a data-driven prior on the largely unknown value of p, and we were able to exclude $p \gtrsim 3$.

This work can be extended in several directions. First, we have not attempted an optimal noise weighting of the power spectrum data. This could be done using optimal quadratic estimators (Tegmark et al., 1998), for which new algorithms have been recently presented (Philcox, 2021b). A fully optimal analysis will allow us to get closer to the full Fisher information contained in the power spectrum. Secondly, it is well known that the Bispectrum is the most sensitive probe to local PNG. A careful study of the optimal weights for higher-point statistics is still missing and could revolutionise the way we constrain $f_{\rm NL}$. We intend to return to these interesting problems in future work.

CHAPTER 7

Preliminary applications of machine learning to LSS analysis

7.1 Introduction

Next-generation redshift surveys aim to measure late-time cosmological parameters with unprecedented precision (LSST Science Collaboration et al., 2009; Laureijs et al., 2011; DESI Collaboration et al., 2016; Akeson et al., 2019). To reach this goal successfully these experiments were designed to rapidly acquire data for a large number of objects. However, the increment of galaxies with a known redshift and the larger volumes of these surveys will also pose new challenges to cosmological analyses. In particular, the combination of the field two-point statistics with higher-order statistics is of key importance to extract all the information from LSS data (e.g., Gil-Marín et al., 2015; Agarwal et al., 2021).

The measurement of these statistics has high computational requirements and their modelling is a non-trivial task as well (e.g., Scoccimarro, 2015; Philcox, 2021a; Pardede et al., 2022). In the last years, the cosmological community has been trying to understand how to solve these problems and many machine learning-based solutions have been proposed. A first solution that has the potential to solve both problems is *field level analysis*. In this kind of analysis, we do not need to extract any summary statistics from the galaxy field as the algorithms directly analyse the three-dimensional distribution of the galaxies. Jasche et al. (2010) and Jasche & Wandelt (2012) proposed the first field-level algorithms that reconstruct the initial conditions of the Universe to recover the cosmological parameters. Another very popular solution is the use of convolutional or graph neural networks (e.g., Ravanbakhsh et al., 2017; Villaescusa-Navarro et al., 2021; Villanueva-Domingo & Villaescusa-Navarro, 2022) to directly analyse the galaxy field. A priori, this second class of algorithms can directly output the cosmological parameters and does not require the modelling of any summary statistics. In this case, we talk about *likelihood-free analysis*, which is a solution to the modelling problem.

In a likelihood-free analysis, the approach involves training a neural network to establish the relationship between a compression of the data and the corresponding cosmological parameters. The neural network effectively serves as an approximation of the posterior distribution for the parameters based on the given data compression. This methodology is versatile and can be applied to various data compression of the field. For instance, it can be utilised for analysing conventional summary statistics of the field, such as the power spectrum or the bispectrum (Hahn et al., 2023a,b) or it can be extended to handle more unconventional compression, such as the scattering transform (Cheng et al., 2020; Valogiannis et al., 2023; Régaldo-Saint Blancard et al., 2023), or even the output of a field-level algorithm (Lemos et al., 2023).

This chapter introduces preliminary results from a field-level analysis algorithm intended for application to VIPERS (Guzzo et al., 2014; Scodeggio et al., 2018). The unique pencil beam geometry of VIPERS allows for flattening it in one direction, generating images that serve as input for a two-dimensional convolutional neural network. The choice of two-dimensional CNNs is strategic, as they exhibit lower memory requirements and faster training times compared to their three-dimensional counterparts or graph neural networks, which are typically employed for larger surveys. This optimisation is particularly advantageous for efficiently handling the characteristics of the VIPERS survey. We build the CNN to analyse simultaneously the two fields of view of VIPERS and to measure the matter density parameter, $\Omega_{\rm m}$, and the amplitude of the linear matter power spectrum at $8\,\mathrm{Mpc}\,h^{-1}$, σ_8.

This work is still limited to the analysis of dark matter halo distributions generated with dark matter simulations. Its novelty lies in the use of Lagrangian perturbation theory simulation for the training of the algorithm, in the application of the survey mask to the simulated light cones in order to reproduce the survey geometry, and in the use of only observational information to build the model inputs, such as the halos positions in the sky (right ascension and declination) together with their redshifts. Additionally, we test the algorithm with halos produced with an N-body simulation in order to understand if it generalises over different simulation types. We train and test the algorithm both in real and redshift space.

The chapter is organised as follows: Section 7.2 describes the CNN architecture and Sect. 7.3 the data used for the training and testing. In Sect. 7.4 we present the results of the different analyses. We conclude in Sect. 7.5.

7.2 Model

We say that a neural network is convolutional if at least one of its layers applies a convolution to the neuron features (see Eq. 2.6) instead of the standard linear transformation (see Eq. 2.2). Convolution can be easily defined for any N-dimensional tensor and it is a very efficient operation whenever the data have a grid-like structure. For this work, we make use of two-dimensional CNNs.

We build a CNN model that takes as input the two VIPERS fields of view, W1 and W4, simultaneously. We sketch the architecture of the CNN in Fig. 7.1. The first section of the CNN consists of two parallel convolutional models that respectively take as input W1 and W4. The first convolutional operation has a kernel specifically designed to output a square image starting from each one of the pixelised fields. Given the kernel dimension k_i along axis i, its stride s_i, which corresponds to the shift of the kernel between two convolutions, and the padding p_i, which is the number of null pixels we add around the input tensor before the convolution, we can determine the final dimension of the convolutional layer output along axis i as follows

$$d_i^{\rm out} = \left\lfloor \frac{d_i^{\rm in} + 2\,p_i - k_i}{s_i} + 1 \right\rfloor . \tag{7.1}$$

Conversely, we can exploit Eq. (7.1) to determine the convolution parameters given the input image shape and an output required dimension. The first convolution reduces the two pixelised fields to squares with the same dimension divided in 8 filters, which undergo a batch normalisation (Ioffe & Szegedy, 2015). The following three layers are a repetition of the convolution operation and the batch normalisation. After each convolution, the dimensions of the images are halved in both directions, while the number of

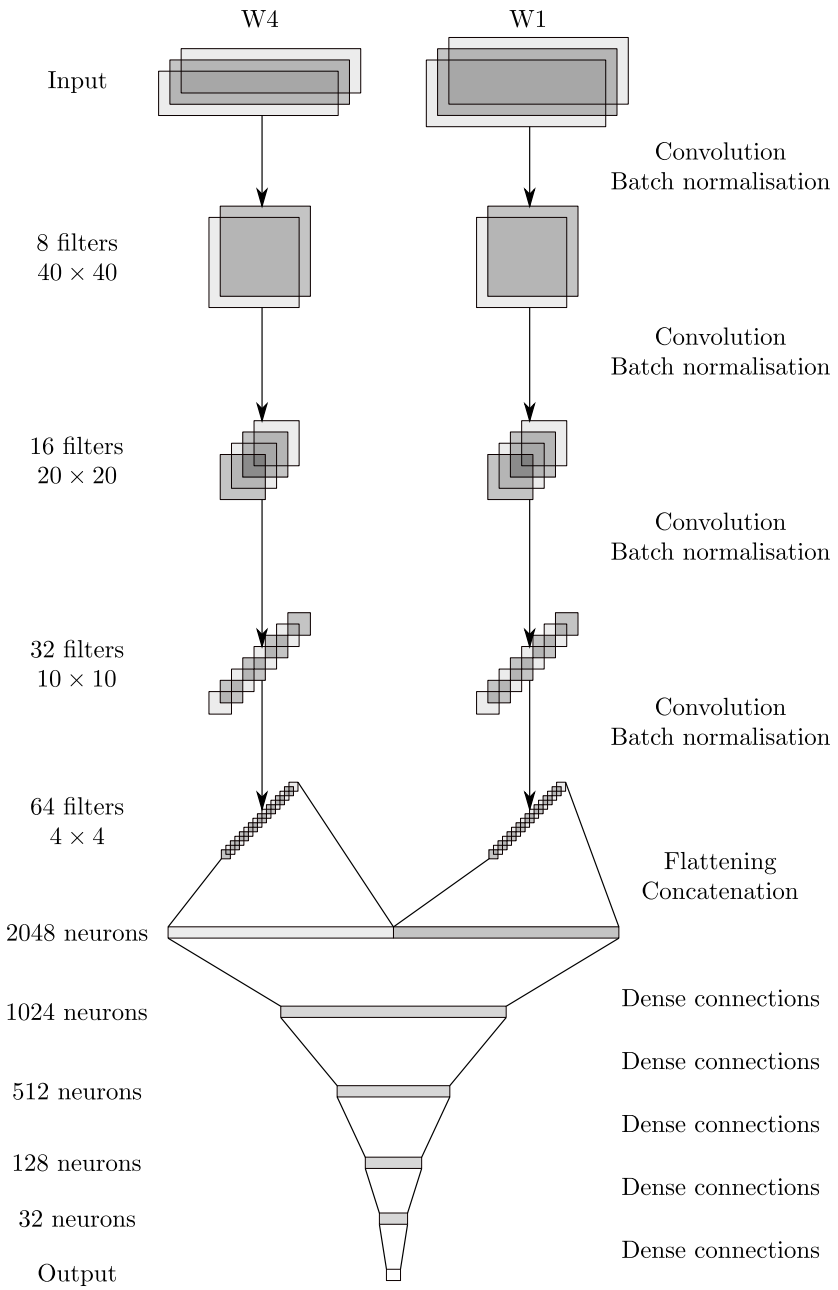

Figure 7.1: Schematic representation of the CNN architecture. The network takes as input the two fields of view, which are pixelised in right ascension and redshift and cut into three declination slices. The images undergo a sequence of convolutional layers separately. For visualisation purposes, we represent only one filter out of four for the hidden convolutional layers. The outputs of the last convolution are flattened and concatenated. Finally, the hidden features are compressed into the output through a series of dense layers. The network outputs either the mean value of the cosmological parameters of interest or their first and second moments.

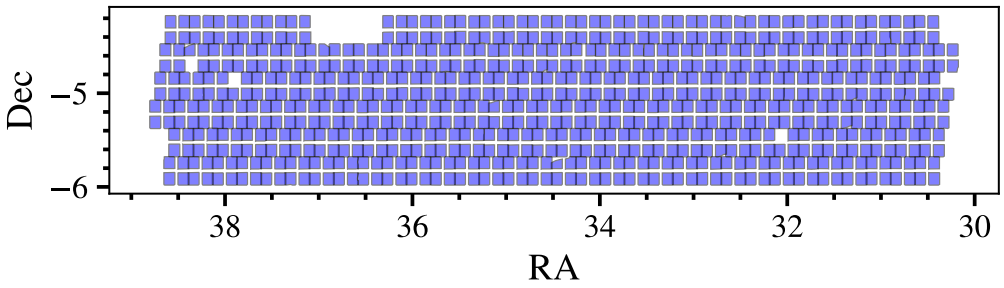

Figure 7.2: Sky footprint of VIPERS W1.

filters is doubled. The final convolutional output for each VIPERS field is a tensor with dimension $4 \times 4 \times 64$. These outputs are flattened and concatenated before the final dense layers of the network, which compress the convolutional output into the cosmological parameters. All the neurons of the CNN, except for the outputs, use a rectified linear unit as activation function (ReLU; Agarap, 2018).

For this work, we developed two different CNNs, which differ in their outputs and loss functions. The first CNN is a *regression* network that simply outputs the value of the cosmological parameters and uses as loss function the mean square error loss (Eq. 2.4),

$$L^{\mathrm{MSE}}(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_a \frac{1}{n_\mathrm{b}} \sum_{i \in \mathrm{batch}} \left(y_{i,\,a} - \widehat{y}_{i,\,a}\right)^2 , \tag{7.2}$$

where n_b is the batch size and a indexes the regression parameters. On the other hand, the second network performs *inference* in the sense that it outputs both the first and second moment, $\boldsymbol{y}_a = [\mu_a, \sigma_a]$, of the a-th cosmological parameter distribution. In cosmology, this method, which is also known as moment network, was originally proposed by Jeffrey & Wandelt (2020) and developed further by Villaescusa-Navarro et al. (2022). The loss function of this model reads as follows

$$\begin{aligned} L^{\mathrm{m}}(\widehat{\boldsymbol{y}}, \boldsymbol{y}) = &\ln\left(\sum_a \frac{1}{n_\mathrm{b}} \sum_{i \in \mathrm{batch}} \left(\widehat{y}_{i,\,a} - \mu_{i,\,a}\right)^2\right) \\ &+ \ln\left(\sum_a \frac{1}{n_\mathrm{b}} \sum_{i \in \mathrm{batch}} \left(\left(\widehat{y}_{i,\,a} - \mu_{i,\,a}\right)^2 - \sigma_{i,\,a}^2\right)^2\right) . \end{aligned} \tag{7.3}$$

If we assume that the posterior distribution of the parameters is Gaussian the output of the moment network completely defines it. Conversely, if the posterior is not Gaussian we are approximating it to a normal distribution with this parameterisation. In this case, a better solution would be a neural density estimator (Hahn et al., 2023a), but we postpone to future work the study of this model. Finally, we use Adam (Kingma & Ba, 2014) as the CNN optimiser.

7.3 Data

Even though this work is entirely based on simulation data, broadly speaking it also makes use of VIPERS data (Guzzo et al., 2014). As mentioned above the final objective

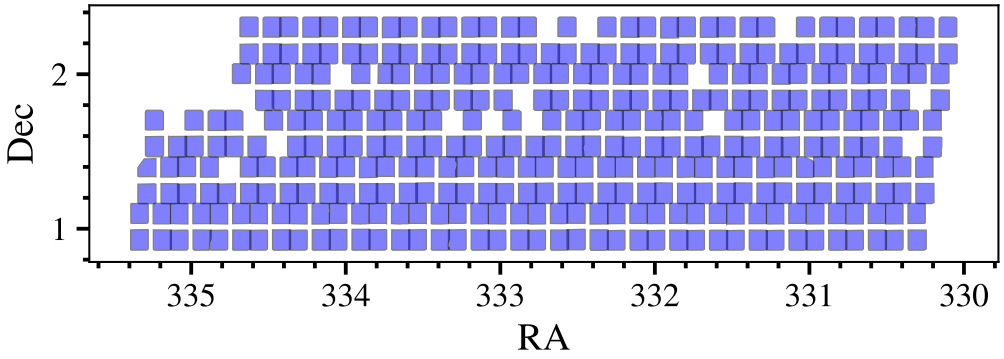

Figure 7.3: Sky footprint of VIPERS W4.

of this project is to apply the developed ML algorithm to VIPERS and make a field-level analysis. VIPERS is a spectroscopic survey with a magnitude limit of $i_{\rm AB} < 22.5$ that covers the redshift range $0.5 < z < 1.5$. For this work, we limit the redshift range to $0.6 < z < 1.0$, as it corresponds to the region with the highest density and completeness. The survey is divided into two fields of view, W1 and W4, which are based on the corresponding wide fields of the CFHTLS photometric catalogue. The two fields cover an area of $\sim 24\,\mathrm{deg}^2$ and we show their footprints on the sky in Figs. 7.2 and 7.3.

As a first step in the direction of real data analysis, we use these same footprints to cut VIPERS-like light cones from the simulated data. By applying the two field masks we impose to the simulations the same angular selection of VIPERS and the corresponding angular component of the window function.

7.3.1 Lagrangian perturbation theory simulations

The CNN training data are produced using the PINpointing Orbit-Crossing Collapsed HIerarchical Objects code (PINOCCHIO; Monaco et al., 2002). PINOCCHIO traces the progression of a group of particles arranged on a uniform grid, employing an ellipsoidal model for calculating collapse times and recognising dark matter halos. Additionally, it uses third-order Lagrangian perturbation theory (3LPT; Munari et al., 2017) to displace the halos from their initial positions. The 3LPT version of the code recovers the halo power spectrum of an N-body simulation within 10% up to a scale of $k \sim 0.5\,h\,\mathrm{Mpc}^{-1}$ in real space. The loss in accuracy with respect to an N-body simulation corresponds to a faster generation of the halo catalogues, possibly making PINOCCHIO a cheaper alternative to generate catalogues for ML applications.

Using PINOCCHIO we generate a set of 2500 simulated cosmologies. All the simulations have a flat ΛCDM cosmology with $h = 0.6777$ and $n_s = 0.96$. Between the simulations we vary the matter density parameter, $\Omega_{\rm m}$, the baryon density parameter, $\Omega_{\rm b}$, and the matter power spectrum amplitude at $8\,h^{-1}\,\mathrm{Mpc}$, σ_8. The 2500 different cosmologies are extracted from a Latin hypercube with $\Omega_{\rm m} \in [0.1, 0.5]$, $\Omega_{\rm b} \in [0.03, 0.07]$, and $\sigma_8 \in [0.6, 1.0]$. We compute the initial power spectrum of the simulation with the cosmic linear anisotropy solving system (CLASS; Lesgourgues, 2011) as the linear matter power spectrum at $z = 0$. We run the simulations with a box size of $400\,\mathrm{Mpc}\,h^{-1}$ and a grid size of 600. The output of a simulation consists of a light cone in the redshift range $0.6 < z < 1.0$ and an aperture of $5\,\mathrm{deg}$. The minimum halo mass is $M_{\rm min} \simeq 10^{11}\,M_\odot\,h^{-1}$.

From each light cone, we cut 32 realisations of the two VIPERS fields. To reduce the overlap of the realisation we cut them displaced from the centre of the cone, which we rotate and mirror to augment the number of data. At the end of the process, we have 80 000 simulations in 2500 cosmologies for training, testing, and validation of the CNN.

7.3.2 N-body simulations

We test the ML model trained with the fast Lagrangian simulation with the Dark Energy and Massive Neutrino Universe simulations (DEMNUni; Carbone et al., 2016). DEMNUni is a set of high-resolution N-body simulations in four different cosmologies: a standard flat ΛCDM simulation and three νΛCDM generalisation to universes with different total masses for neutrinos. For this work, we are only interested in the standard ΛCDM realisation, which is characterised by a Planck 2013 cosmology (Planck Collaboration et al., 2014).

The DEMNUni simulations were produced using the tree particle mesh-smoothed particle hydrodynamics algorithm (TreePM-SPH) implemented in GADGET-3 (Springel, 2005). Viel et al. (2010) modified the original GADGET-3 code to include massive neutrinos. The simulations have Zel'dovich initial conditions at $z_{\rm in} = 99$. They cover a comoving volume of $\left(2\,{\rm Gpc}\,h^{-1}\right)^3$, which contains 2048^3 dark matter particles and the same number of neutrinos when present. Each simulation contains 62 snapshots in the range $0 < z < 99$ logarithmically spaced in the scale factor a. From each snapshot, a halo catalogue was extracted using the GADGET-3 friends-of-friends algorithm (FoF; Springel et al., 2001; Dolag et al., 2009). The minimum halo mass was set to $M_{\rm FoF} \simeq 2.5 \times 10^{12}\,M_\odot\,h^{-1}$ (Castorina et al., 2015). The DEMNUni halo catalogues are full-sky light cones from which we cut 100 realisations of each VIPERS field within the redshift range of interest.

7.3.3 Pixelisation and pre-processing

For the analysis, we adopted a two-dimensional CNN. Therefore, we need to produce two-dimensional images as input for the network starting from the three-dimensional VIPERS-like simulations. Figures 7.2 and 7.3 show that both VIPERS fields are very thin in the declination direction. This suggests flattening the field in this direction to produce an image in right ascension and redshift. However, such a choice would wash out the three-dimensional information of the survey. The solution we adopted is to cut declination slices from the field and flatten those to produce the right ascension and redshift images. Analogously to RGB images, which are the combination of the three primary-colour images, the VIPERS field corresponds to a series of two-dimensional images at different declinations.

We cut both the fields in three declination slices, which we then pixelise in the two remaining directions. The pixelisation is uniform in each one of the two directions. We chose the bin width to correspond to $\sim$10 Mpc h^{-1} at the mean redshift of the field. This corresponds to $\Delta z = 0.005$, which results in 80 redshift bins, and $\Delta{\rm RA} = 0.3$ deg, which produces 30 and 20 right ascension bins respectively for W1 and W4. The halo field is pixelised with a simple nearest-grid-point method and the halo masses are ignored in this process. Finally, the three images corresponding to each realisation are normalised by the maximum pixel value between them. This normalisation choice has two advantages, first, it removes the absolute value information making the analysis more realistic as for real data we do not know the true number density of objects. Second, it normalises the pixel values between 0 and 1 making CNN processing easier and faster. If the pixel

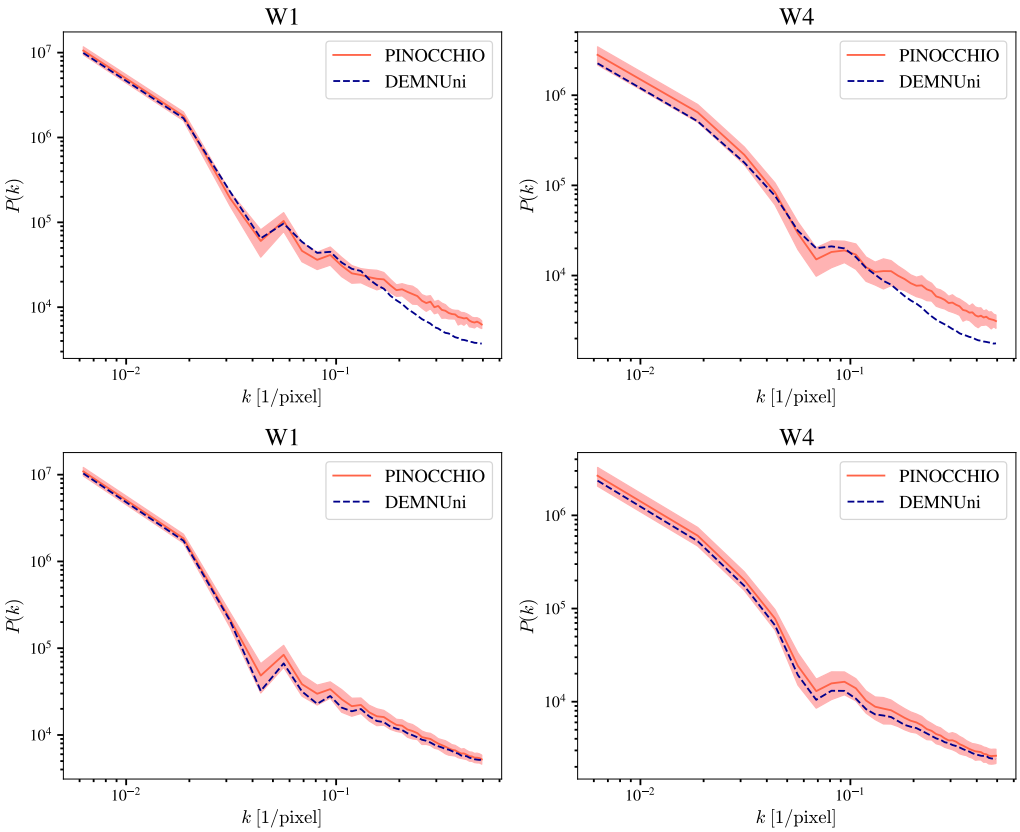

Figure 7.4: Un-normalised two-dimensional pixel power spectrum in redshift (top) and real (bottom) space. The left panels present the power spectrum for W1, while the right panels for W4. The red solid line represents the mean power spectrum of the 96 PINOCCHIO realisations and the shaded area corresponds to the 1σ error of their distribution. The dashed blue line is the mean power spectrum of the 100 DEMNUni VIPERS realisations.

values cover many orders of magnitude, a better solution would be to take the logarithm of the normalised values. However, that is not the case for this work.

We use this procedure to pixelise both the PINOCCHIO simulations and the DEMNUni simulation. However, as the minimum halo mass of the two simulations is different before pixelising the fields, we apply the same halo mass cut to both the simulation types, $M_{\rm halo} \geq 3 \times 10^{12}\, M_\odot\, h^{-1}$.

We expect the field-level analysis to extract information from statistics higher than the two-point. However, if the simulations already differ at the power spectrum level we cannot expect the network to generalise over the simulation types. As a first test, we compare the mean power spectra of the DEMNUni simulation and a PINOCCHIO simulation in the same cosmology. We compute the power spectrum from the two-dimensional images in right ascension and redshift that we feed to the CNN instead of the traditional three-dimensional power spectrum. Even though the meaning of this power spectrum is less intuitive it is helpful to understand the CNN input data. In the

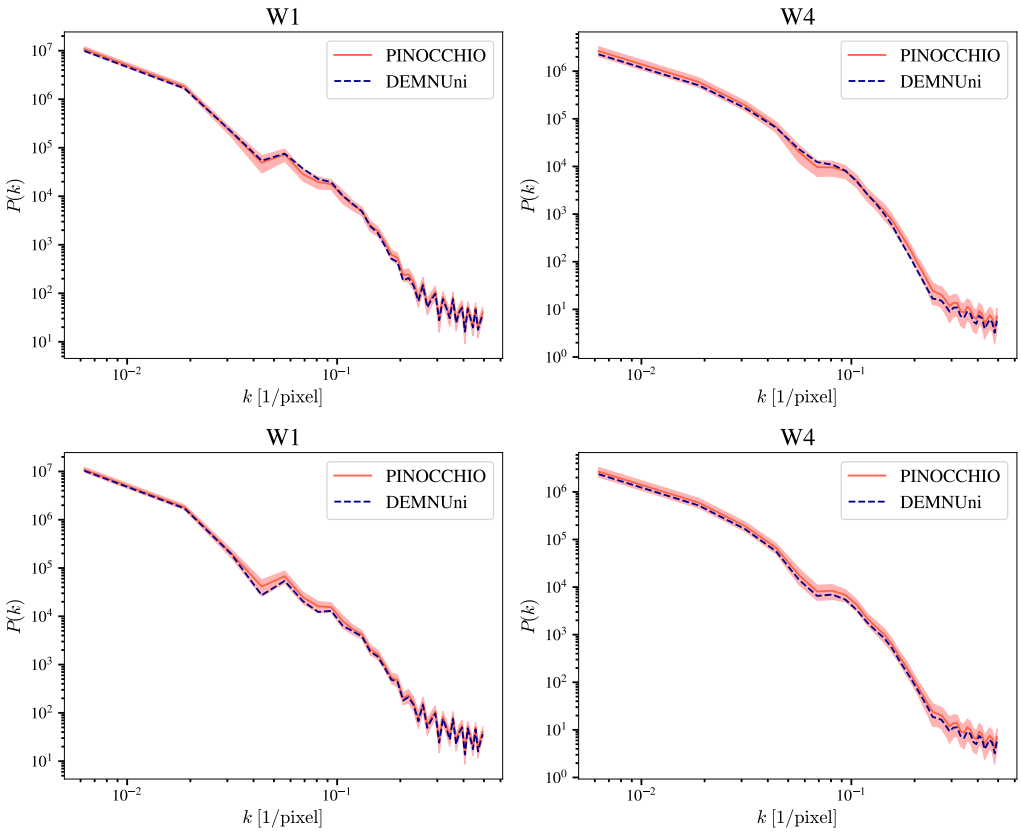

Figure 7.5: Un-normalised two-dimensional pixel power spectrum of the smoothed images in redshift (top) and real (bottom) space. Same as Fig. 7.4.

case of PINOCCHIO, we generated three simulations in the DEMNUni cosmology and obtained 96 VIPERS realisations that we can compare with the 100 DEMNUni VIPERS realisations. Figure 7.4 top panels show the mean of the un-normalised power spectra of the 100 DEMNUni realisations (dashed blue line) for the two VIPERS field (left and right respectively W1 and W4) and the mean of the un-normalised power spectra of PINOCCHIO (solid red line). The shaded area corresponds to the 1σ error in the PINOCCHIO power spectrum. The plots show that at high ks, where the signal is dominated by the redshift space distortion, the two power spectra diverge one from the other. In particular, the absence of power loss in the PINOCCHIO power spectrum proves that the peculiar velocities are not correctly simulated by the Lagrangian code.

A first solution is to remove the redshift space distortion by analysing the fields in real space. Figure 7.4 bottom panels present the mean power spectra in real space. We do not observe anymore the discrepancy at high ks. However, it seems that the DEMNUni power spectrum systematically has less power than the PINOCCHIO power spectrum. This is true for both fields, but the effect is more relevant for W4.

An alternative solution we adopted to solve the difference in the power spectra at high ks is a smoothing of the field. We apply to the pixelised images a symmetric Gaussian filter with $\sigma = 1.7$, which roughly corresponds to a minimum scale of $\sim$17 Mpc h^{-1}.

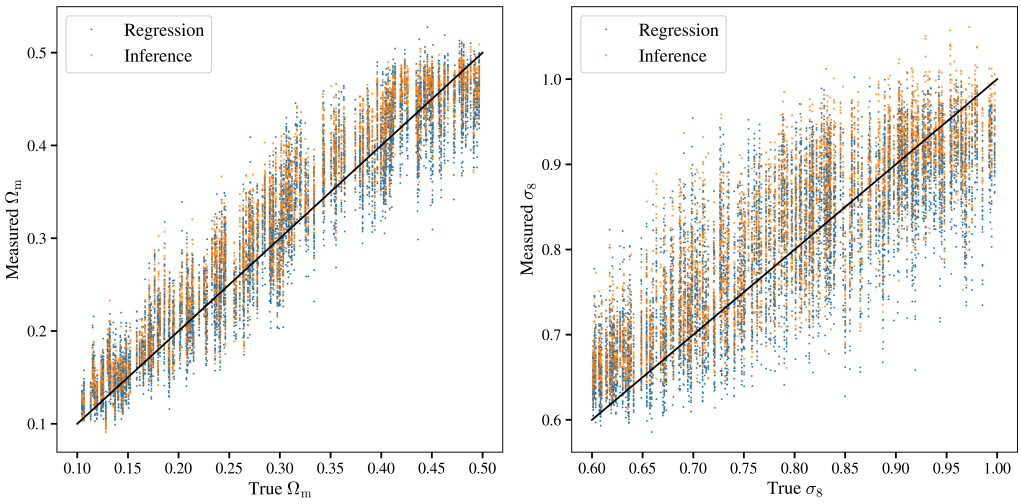

Figure 7.6: Measured and true values of the cosmological parameters for the CNN trained in redshift space without smoothing. The test is performed with PINOCCHIO simulations. The blue points correspond to the output of the regression CNN, while the orange squares to the inference or moment network.

We present the mean power spectra of the smoothed images in Fig. 7.5 in redshift (top panels) and real (bottom panels) space. In redshift space, thanks to the smoothing, the differences related to the peculiar velocities are washed out and the power spectra of the two simulations are more consistent. We note that both in the smoothed and non-smoothed case there is a small discrepancy between the W4 power spectra at low ks. This difference, which we observe also in real space, is within the 1σ error for both fields, but may lead to biased CNN outputs.

7.4 Results

In the following section, we present the test results of the CNNs trained with various input configurations. To summarise, all the CNNs are trained with 75% of the PINOCCHIO simulations, with an additional 15% reserved for validation to monitor performance during training and identify the best model. Early stopping is implemented based on the validation loss. The remaining 10% of the PINOCCHIO simulations is used for testing. We remark that simulations for training, validation, and testing are selected to ensure non-overlapping cosmologies.

Figure 7.6 illustrates the results of a PINOCCHIO self-test for the CNN trained in redshift space without smoothing of the pixelised field. The results show that the CNN is able to extract the two cosmological parameters of interest when tested on unseen cosmologies produced with the same simulation code used for the training data. We note that at the boundaries of the intervals slight biases in the outputs are observed, which is a common behaviour for machine learning algorithms, attributed to reaching the training prior limits (Villaescusa-Navarro et al., 2020).

We quantify the performance of the CNN PINOCCHIO self-test by computing the mean squared relative error and the bias over the whole parameter range. The mean squared relative error is consistent between the regression and inference CNNs, mea-

suring ~ 0.02 for $\Omega_{\rm m}$ and ~ 0.007 for σ_8. However, the bias tends to be higher for the inference network compared to the regression network. Specifically, for the latter, it is ~ 0.006 and ~ 0.009 for $\Omega_{\rm m}$ and σ_8 respectively, while for the inference network, it is ~ 0.02 for both the parameters.

In the next two sections, we present the results of testing the CNNs trained in four different configurations: with or without smoothing and in redshift or real space. We compare the test results obtained with the DEMNUni realisations and with the PINOCCHIO realisations in the same cosmology.

7.4.1 Analyses without smoothing

Figure 7.7 presents the results of the regression CNN tested with both PINOCCHIO and DEMNUni simulations (red and blue markers) in redshift and real space (top and bottom panels). We observe similar behaviour in redshift and real space. In both analyses, the distribution of σ_8 is consistent with the true value, exhibiting no apparent bias. However, for $\Omega_{\rm m}$ we note a bias towards higher matter density parameters. This is evident from Table 7.1, where we summarise the results of all the analyses. The mean value of the measured $\Omega_{\rm m}$ is consistently higher than the true value ($\Omega_{\rm m}^{\rm true} = 0.27$). Notably, the distributions are more compact in redshift space compared to real space.

While the errors in the PINOCCHIO analysis are smaller than those in the DEMNUni simulations, the similarities of the output distributions of the two simulation types, particularly for σ_8, are still remarkable. The PINOCCHIO distributions have sharper features than DEMNUni, but they tend to have the main peak in the same position. This simple result is of utmost importance. It shows that the CNN performs a consistent compression that maps both simulation types in a latent space where they have a similar representation. This demonstrates the CNN ability to generalise across different simulation scenarios to some extent. If the two distributions coincide it is possible to build a model, e.g., a neural density estimator, that de-biases the results.

Figures 7.8 and 7.9 depict the output distributions of the inference CNNs in redshift and real space, respectively. Notably, in the moment network analysis, the bias in measuring $\Omega_{\rm m}$ is even more pronounced than in the regression CNN analyses, see Table 7.1. In the redshift space analysis, the true combination of the two cosmological parameters (green star) lies at the edge of the two-dimensional distribution, while in the real space analysis, it is distinctively separated from the measured distribution (red dots and blue squares). Nevertheless, the similarity in the output distributions of PINOCCHIO and DEMNUni analyses, particularly in real space, suggests the generalisation power of the CNN and the potential for de-biasing the results.

As for the width of the distributions, we observe that in redshift space the regression CNN produces larger errors compared to the moment network and that the inference CNN second moments are smaller than the actual width of the first moment distributions. This is particularly evident for DEMNUni, where the second-moment means are about half the error in the first-moment distribution for both cosmological parameters. In real space, the $\Omega_{\rm m}$ second moment provides a smaller estimate of the error, while for σ_8, it produces results comparable to the regression and first moment distributions. Given the limited number of data points, the output distributions are highly non-Gaussian, making the use of only the first and second moments insufficient to fully describe them. Therefore, we cannot conclusively state that the inference network is unable to recover the second moment of the distribution; however, the current results seem to point in that direction.

Finally, returning to the power spectrum comparison (see Sect. 7.3.3) we note that the

Simulation	Field	Space	CNN	$\Omega_{\rm m}$	σ_8	$\sigma_{\Omega_{\rm m}}$	σ_{σ_8}
PINOCCHIO	No smoothing	Redshift	Regression	0.300 ± 0.032	0.820 ± 0.065	–	–
			Inference	0.302 ± 0.029	0.851 ± 0.062	0.024 ± 0.003	0.051 ± 0.003
		Real	Regression	0.307 ± 0.038	0.832 ± 0.081	–	–
			Inference	0.319 ± 0.041	0.894 ± 0.075	0.036 ± 0.013	0.084 ± 0.009
	Smoothing	Redshift	Regression	0.276 ± 0.033	0.805 ± 0.072	–	–
			Inference	0.261 ± 0.025	0.843 ± 0.057	0.033 ± 0.004	0.076 ± 0.009
		Real	Regression	0.302 ± 0.033	0.823 ± 0.079	–	–
			Inference	0.316 ± 0.025	0.844 ± 0.072	0.043 ± 0.006	0.084 ± 0.007
DEMNUni	No smoothing	Redshift	Regression	0.313 ± 0.047	0.803 ± 0.077	–	–
			Inference	0.318 ± 0.041	0.821 ± 0.081	0.026 ± 0.004	0.049 ± 0.004
		Real	Regression	0.313 ± 0.049	0.804 ± 0.093	–	–
			Inference	0.320 ± 0.052	0.855 ± 0.077	0.038 ± 0.010	0.080 ± 0.010
	Smoothing	Redshift	Regression	0.285 ± 0.044	0.754 ± 0.084	–	–
			Inference	0.276 ± 0.040	0.795 ± 0.087	0.033 ± 0.005	0.072 ± 0.012
		Real	Regression	0.328 ± 0.044	0.787 ± 0.091	–	–
			Inference	0.330 ± 0.038	0.825 ± 0.078	0.044 ± 0.007	0.079 ± 0.009

Table 7.1: Results of the CNN tests in all the different configurations using both PINOCCHIO and DEMNUni. We report the mean and standard deviation of the output distributions of the different networks. The regression CNN only outputs the value of the cosmological parameter, while the inference CNN outputs the first and second moments of the parameters. The true value of the parameters are $\Omega_{\rm m}^{\rm true} = 0.27$ and $\sigma_8^{\rm true} = 0.83$.

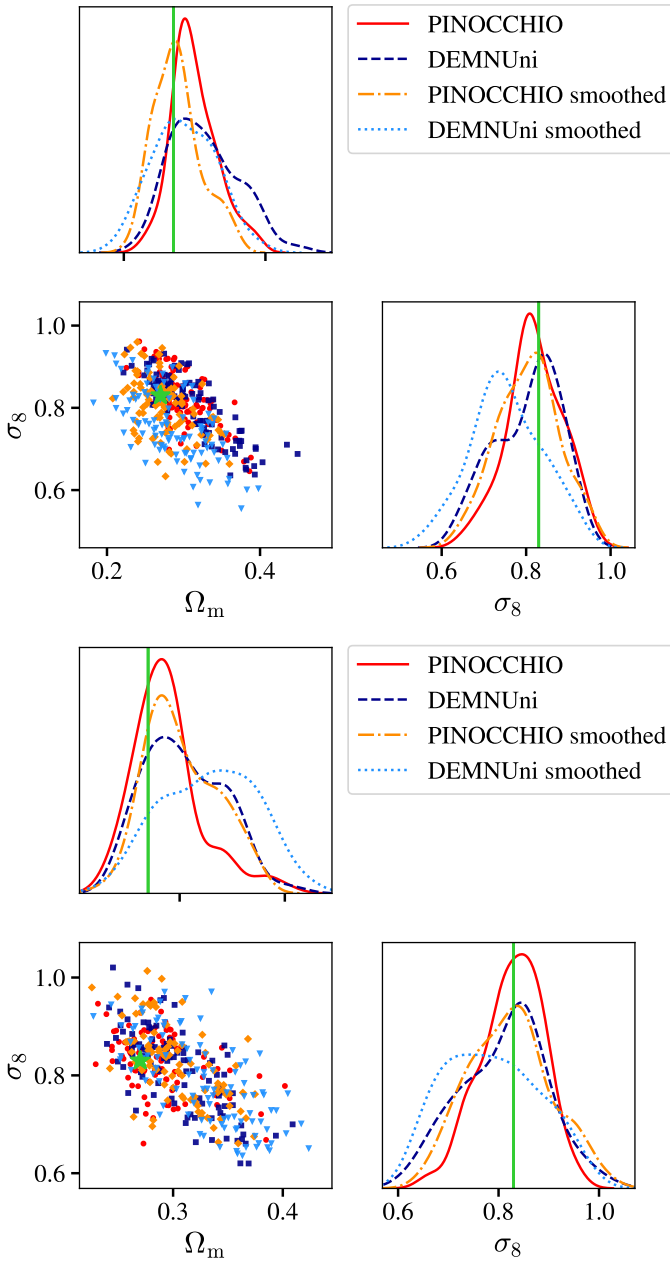

Figure 7.7: Two-dimensional output distributions of the regression CNN. The red dots and solid distributions correspond to the PINOCCHIO test results without smoothing, while the blue squares and dashed distributions to the DEMNUni results. The orange diamonds and dash-dotted distributions are the PINOCCHIO results with smoothing in the field, and the light blue triangles and dotted distributions are the DEMNUni results in the same configuration. The green star with the green vertical lines represents the true value of the cosmological parameter of the simulations. *Top*: redshift space analysis. *Bottom*: real space analysis.

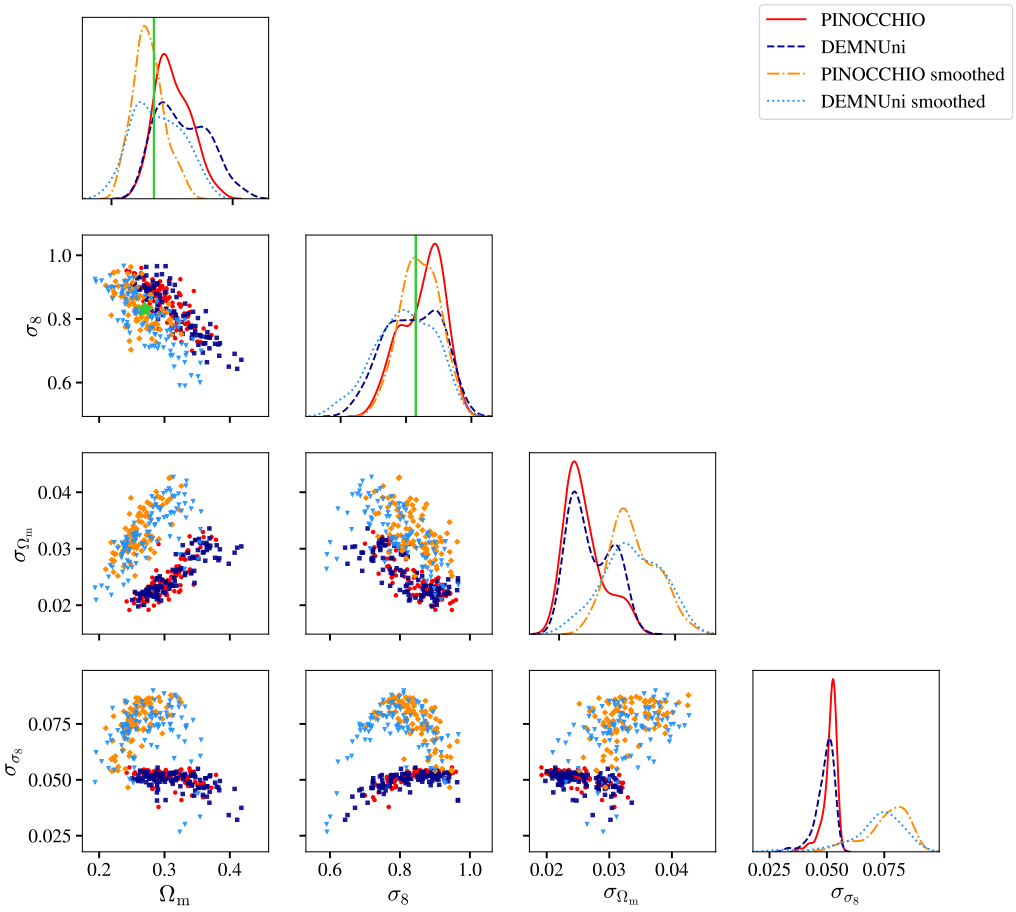

Figure 7.8: Two-dimensional output distributions of the inference CNN in redshift space. As in Fig. 7.7, the red dots (orange diamonds) and solid (dash-dotted) distributions correspond to the outputs of the 96 PINOCCHIO realisations and the blue squares (light blue triangles) and dashed (dotted) distributions to the 100 DEMNUni realisations in the non-smoothed (smoothed) analysis. The green star and vertical lines represent the true value of the parameters.

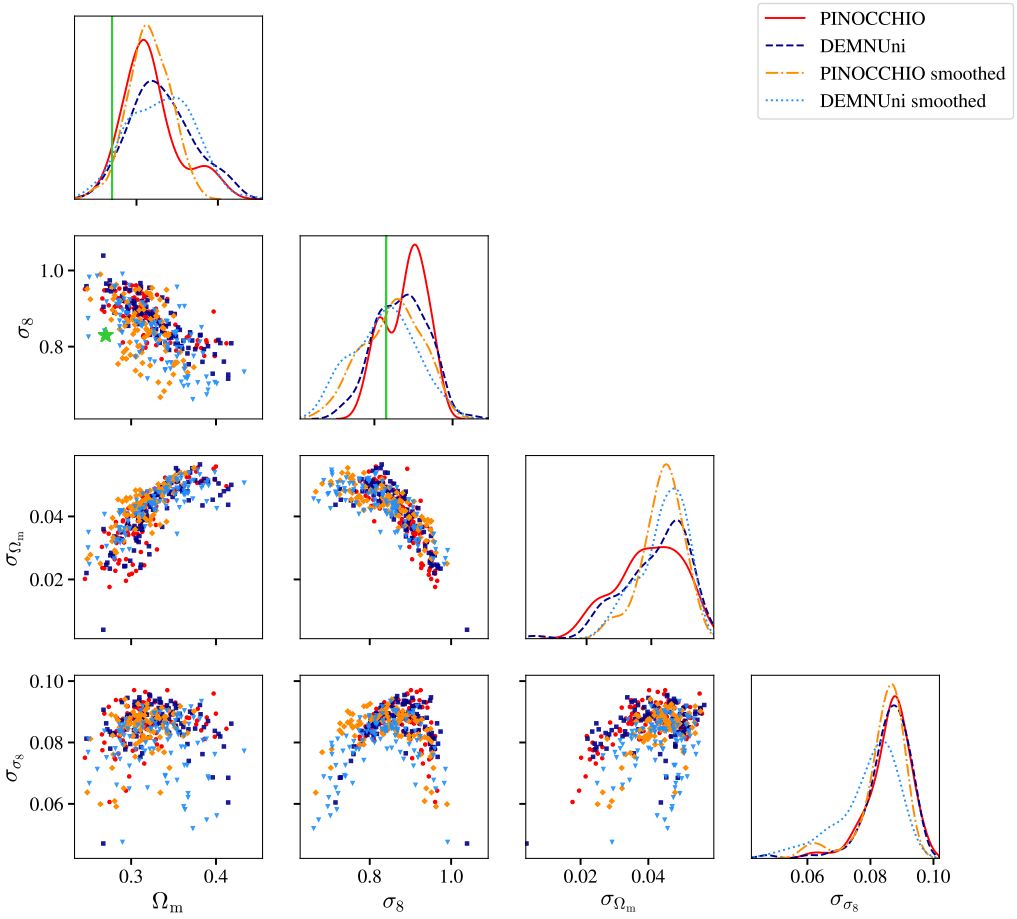

Figure 7.9: Two-dimensional output distributions of the inference CNN in real space. As in Fig. 7.8.

bias in the matter density parameter estimation is almost the same in real and redshift space. Moreover, while the bias is larger for DEMNUni, the results from the two simulations align with each other. This behaviour implies that the bias may not originate from the incorrect redshift space distortion modelling of PINOCCHIO or that the CNN has the ability to marginalise over the smaller scales where the PINOCCHIO redshift space clustering exhibits an excess of power compared to the N-body simulation (see Fig. 7.4 top panels).

7.4.2 Analyses with smoothing

In this section, we present the results from the CNN trained using the smoothed pixelised field. Figure 7.7 shows the output distributions of the regression networks trained in redshift and real space (top and bottom panels) for PINOCCHIO (orange diamonds) and DEMNUni (light blue triangles). The smoothing of the field de-biases the $\Omega_{\rm m}$ outputs in redshift space, but the issue remains in real space. Now, the distributions of the PINOCCHIO and DEMNUni tests are less consistent, especially in real space. The PINOCCHIO distribution appears more compact than the DEMNUni distribution, but their means and standard deviations are still consistent, see Table 7.1. We note that the smoothing removed the bias of the matter density parameter only in redshift space, but it has reduced the estimated σ_8 values both in real and redshift space.

We show the results of the inference CNNs in Figs. 7.8 and 7.9. Analogously to the regression network, the smoothing removes the bias in redshift space but not in real space. However, in redshift space, the distribution of the two simulation outputs differs, with DEMNUni results exhibiting longer tails and more banana-shaped two-dimensional distribution compared to the PINOCCHIO ones. The differences in the distributions are smaller in real space.

As for the estimates of the parameter second moments, the behaviour differs from the non-smoothed analysis. In the case of PINOCCHIO, the inference second moment is consistent, if not larger (especially in real space), with the error estimated from the regression. However, the first moments of the parameters have more compact distributions than the regression results and the error estimated from their second moment. For DEMNUni, the situation is different, indicating that now the two simulation distributions are less consistent. In redshift space, for both $\Omega_{\rm m}$ and σ_8, the estimated second moment is smaller than the observed error. In real space, the mean of the matter density parameter second moment is consistent with the regression error and is larger than the first moment error, while for σ_8, the inference first-moment distribution is more compact than the regression one and its width is consistent with the estimated second moment.

In redshift space, we observe a clear separation between the second-moment distributions of the smoothed and non-smoothed analyses, while in real space they cover the same parameter ranges. This seems to indicate that the CNN exploits the peculiar velocity information, as its estimated error increases when we wash out this information with the field smoothing. The idea that the CNN is actually exploiting the small-scale peculiar velocity information is reinforced by the fact that we do not observe the shift of the second moment estimated values in the smoothed and non-smoothed real space analyses.

Even though it was expected from the behaviour of the power spectra after the smoothing (see Figs. 7.4 and 7.5), we find very peculiar that the smoothing removes the bias in redshift space, but it does not alleviate it in real space. This suggests again that the bias is not solely due to incorrect modelling of halo peculiar velocities. At the same time, the smoothing in redshift space reduces the overlap of the PINOCCHIO and

DEMNUni results, indicating that it does not help the network in generalising over the simulation types.

7.5 Discussion and conclusions

In this study, we have presented preliminary results for a machine learning-based field-level analysis of VIPERS, focusing on the measurement of the two cosmological parameters, $\Omega_{\rm m}$ and σ_8. While the current investigation is confined to dark matter halos, it already incorporates the survey geometry into the simulated data, using only observational information such as the halo right ascension, declination, and redshift, to construct inputs for the convolutional neural network. Exploiting the pencil-like geometry of the two VIPERS fields of view, we represent the inputs as two-dimensional images in right ascension and redshift. To retain the three-dimensional information, we divide the VIPERS fields into three declination slices before generating the images.

We trained the network with 3LPT simulations produced with PINOCCHIO and tested it both with PINOCCHIO and an N-body simulation, DEMNUni. We developed two different CNNs: the first is a regression network that only outputs the parameter values, while the second network performs inference by measuring the first and second moments of the parameter distributions. We make the analysis both in real and redshift space, considering two minimum scales in pixalisation, $\sim 10\,\mathrm{Mpc}\,h^{-1}$ and $\sim 17\,\mathrm{Mpc}\,h^{-1}$, for the non-smoothed and smoothed fields, respectively. The results from the non-smoothed field have a bias in the measurements of $\Omega_{\rm m}$ both in real and redshift space. This effect is mitigated by the smoothing of the pixelised field in redshift space.

However, the most significant aspect of the results lies in the comparison of the PINOCCHIO and DEMNUni output distributions, rather than the accurate estimation of the cosmological parameters. If the distributions from the two simulations are consistent, it indicates that the CNN compresses them in a latent parameter space where their representation is equivalent. This opens the possibility of de-biasing the results using alternative methods, such as neural density estimators. Furthermore, an overlap of these distributions would suggest that the CNNs exhibit a degree of generalisation across different simulation types. We observe the highest overlap between the PINOCCHIO and DEMNUni distributions in the analyses without pixelisation smoothing, especially for the regression network. To quantitatively measure this overlap, we plan to employ distribution comparison statistics, e.g., as the Kolmogorov-Smirnov test (Kolmogorov, A. L., 1933; Smirnov, 1948) or the Kullback-Leibler divergence (Kullback & Leibler, 1951). Nevertheless, these first results point in the direction that the CNN can generalise over the simulation types. If we were to confirm this preliminary finding it would make machine learning-based algorithms very cost-efficient, as the use of approximate simulations would make their training faster and cheaper.

Regarding the inference network, the results seem to indicate that the CNN may struggle to provide robust predictions for the parameter second moments. In the non-smoothed case, the second moments are consistently smaller than the actual standard deviation of the parameter distributions (see Table 7.1). To make a conclusive statement on this issue, we need a larger set of data points in the DEMNUni cosmology for both simulations. Additionally, studying the behaviour of the inference CNN across the entire parameter range would provide valuable insights. Another approach could involve computing the χ^2 statistics for the PINOCCHIO test set, either in parameter bins or over the entire range (de Santi et al., 2023).

The next phase of the work will consist of quantifying the differences between the PINOCCHIO and DEMNUni simulation distributions and understanding whether the

small variations observed in the means of the measured values are related to systematic effects and if we can eventually mitigate them. We also plan to continue the testing with other N-body simulations in different cosmologies, e.g., the MultiDark simulations,[1] to assess if the CNNs have consistent behaviour. If that is the case, it would be possible to start implementing the neural density estimator to de-bias the results and extract posterior distribution for the parameters.

The following step will involve introducing galaxies into the dark matter simulations. We plan to use a standard halo occupation distribution algorithm (HOD; Zheng et al., 2005, 2007) to populate the halos. A HOD model depends on a set of free parameters that regulate the expected number of galaxies that will appear in a halo of a given mass. The choice of these parameters is not trivial and we have at least two possibilities to generate the training set for the galaxy field analysis.

The first solution is to fit the HOD parameters in each cosmology, ensuring that the galaxy clustering reproduces the two-point correlation function of VIPERS. The alternative approach is to marginalise over these parameters by generating a training set with different HOD parameters, enabling the network to learn them together with the cosmological parameters. Even though the second solution may prove less efficient due to degeneracies between the HOD and cosmological parameters, we favour it. This approach allows us to cross-check the network performance by using the measured HOD parameters to generate new mock catalogues and compare their standard clustering statistics with the observed data. The final test will involve comparing the CNN results with the official VIPERS analyses (Rota et al., 2017).

In conclusion, this work provides an initial glimpse into the potential of a machine learning algorithm that uses observational information for field-level analysis. A significant outcome so far is the feasibility of training a network with fast simulations, such as 3LPT simulations, and applying it to a completely different simulation like an N-body simulation. While we are still working to conclusively determine this, the current results suggest that training ML algorithms for field-level analysis can be more computationally efficient than expected, making them a competitive alternative to summary statistics-based analyses.

[1] https://www.cosmosim.org/cms/data/projects/multidark-bolshoi-project/.

Conclusions

The general goal of this thesis has been developing and testing improved methods to analyse large-scale structures and extract cosmological information. In particular, I focused my work on novel machine learning-based algorithms to augment the scientific information inferred from the observational data.

In Part I, I presented two studies related to the analysis of photometric data, with the scope of improving the confidence of photometric redshift estimates and optimising the selection from survey data of galaxy samples for clustering analyses. In Chapt. 4, I discussed a novel method that improves photometric redshift measurements by exploiting the spectroscopic information of angular neighbours. This algorithm, which we dubbed NezNet, is a graph neural network model that classifies angular galaxy pairs as true or false redshift neighbours. The graph neural network helps identify catastrophic errors in the photometric redshift measurements and effectively reduces the dispersion of the final photometric sample by a factor of 2 and the fraction of catastrophic errors by a factor of ~ 4.

Chapter 5 reports the detailed study of photometric selection to improve the purity and completeness of the *Euclid* galaxy clustering spectroscopic sample. In this work, I compared the performance of six machine learning classifiers with standard photometric selection based on colour and magnitude cuts. The results show that, compared to standard methods, machine learning algorithms, particularly neural networks and support vector classifiers, are capable of identifying more complex boundaries in the colour-magnitude multidimensional space. I showed that the combination of the *Euclid* spectroscopic selection with the neural network photometric selection improves the redshift purity of the final sample by $\sim 20\%$ and $\sim 50\%$ when using *Euclid* photometry and *Euclid* with ground-based photometry, respectively.

Both these works open interesting perspectives and I plan to further improve them for real data applications. Regarding NezNet, I intend to extend the architecture of the graph neural network to analyse groups of angular neighbours, rather than just pairs, and to directly estimate the redshift of a photometric galaxy given its spectroscopic neighbours. Then, NezNet could be applied to slitless galaxy samples, such as *Euclid* data, as, in this case, it would be possible to select samples where galaxies with only photometric information and galaxies with spectroscopic information densely overlap.

As for the neural network photometric selection, the final aim is to implement it in the *Euclid* galaxy clustering selection pipeline. We will start working on real *Euclid* data as soon as they become available to verify the algorithm performance when using data from the Euclid Deep Field for the training and accounting for realistic stellar contamination. For this additional selection, I expect that information on galaxy morphology will

become more relevant than found in my current investigation. Another variable that needs stricter control is the varying noise level of the ground-based data. Possible solutions include reducing all the data to the same photometric noise during pre-processing or teaching the network to marginalise over it.

In Part II of the thesis, I have discussed the measurement of cosmological parameters from large-scale structure data, presenting two methods that are either alternative to or improve over more standard approaches. In Chapt. 6, I illustrated my work applying an optimal quadratic estimator to measure local primordial non-Gaussianities from the eBOSS QSO large-scale distribution. With this analysis, I obtained one of the most stringent constraints on $f_{\rm NL}$ from large-scale structure data up to date. The signal optimal quadratic estimator method reduces the bias in the results and gives tighter bounds, $\sigma_{f_{\rm NL}} \sim 16$, on the $f_{\rm NL}$ parameter, with an improvement of $\sim 13\%$ over the standard approach. In the case that quasars have a lower response to local primordial non-Gaussianities, the optimal constraint becomes $\sigma_{f_{\rm NL}} \sim 21$, with a $\sim 30\%$ improvement over standard analyses.

In Chapt. 7, I presented a preliminary application of convolutional neural networks to a two-dimensional field-level analysis of large-scale structures. This study is still limited to the analysis of dark matter halo distributions derived from numerical simulations; however, it uses a realistic survey geometry to generate the training data and it exploits observational information, such as the halo angular position and redshift, to build the network inputs. A major novelty of this work is that the convolutional neural network is trained using samples built with a fast third-order Lagrangian perturbation theory code. This is much faster than using an N-body simulation and allows for quickly building a large training set. I tested the network performance using both the 3LPT and N-body simulations in order to understand the ability of the algorithm to generalise between these two types of simulation with different resolutions and details. I showed that, pixelising the data in cells of $\sim 10\,{\rm Mpc}\,h^{-1}$, the convolutional neural network trained with the 3LPT simulations recovers consistent values for the parameters investigated, when applied to either the 3LPT or the N-body halo catalogues.

Both these studies suggest interesting future developments. To improve even more the constraints of $f_{\rm NL}$, I plan to perform a complete optimal analysis, which combines the optimal quadratic estimator of the PNG signal with a full inverse-noise analysis. Moreover, a joint analysis of the power spectrum and bispectrum would allow us to break some of the degeneracies in the inference parameters. Such improvements, combined with the larger data sets that will soon become available, will help us reach the so-long sought objective of $\sigma_{f_{\rm NL}} \sim 1$ and distinguish between single and multi-field inflation models.

Regarding my machine learning-based field-level analysis, a significant amount of work is still needed. First, I intend to rigorously quantify the differences between the results from the 3LPT and N-body simulations, as well as test the network on different N-body simulations. Proving convincingly that we can use fast simulations to train machine learning algorithms would be a major result by itself, allowing us to reduce the training cost of any field-level neural network. This will make them competitive alternatives to future, more standard analyses that plan to combine two-point and higher-order statistics to capture extra cosmological information at a significant computational cost.

In conclusion, my thesis has addressed the evolving landscape of cosmology, which is currently facing challenges related to the established ΛCDM model and the advent of next-generation cosmological surveys. With the start of a high-precision era for large-scale structure cosmology, thanks to fourth-generation galaxy surveys like *Euclid*, DESI, and LSST, the demand for novel analysis methods becomes imperative. Focusing on

large-scale structure studies, this work explores alternative algorithms to analyse LSS data at different stages of the process, with a primary emphasis on machine learning models. Overall, this thesis contributes to the advancement of cosmological analyses by introducing innovative methodologies to extract cosmological information from the data, paving the way to more efficient and accurate large-scale structure studies in the era of *Euclid* and DESI.

Appendices

APPENDIX A

Photometric selection additional tests

A.1 Colour-magnitude projection planes

We present in Fig. A.1 the colour-colour and colour-magnitude distributions for targets and non-targets in the Flagship2 catalogue for three different combinations. The contours contain 99%, 50%, and 25% of the samples. There is a nearly complete overlap of the targets and non-targets in the colours.

A.2 Photo-z as input variables

As an additional test, we trained a neural network with Flagship2 data in the *Euclid* plus ground-based configuration with the additional information of the measured photo-z. In principle, the neural network can extrapolate the redshift from the photometric information. However, by directly providing the photo-z we may facilitate the selection process as the network will expend less effort in extracting the redshift information.

The precision at 95% recall is 47.9% in the case without photo-z, as reported in Table 5.3. When we add the photo-z of the galaxy as an input feature the precision rises to 50.1%. Nevertheless, the addition of the photo-z to the input information makes the classifier dependent on the redshift precision, on the assumptions in the photo-z estimation algorithm, and finally on the photo-z estimation algorithm itself. We postpone to a future work the detailed study of these dependencies.

A.3 Selection probability maps

Figure A.3 gives a visualisation of the selection probability for each classifier in planes through the parameter space. We show the results from the Flagship2 catalogue for the case when classifiers are trained with *Euclid* photometry alone. Each row shows the colour-colour and colour-magnitude plots for a given algorithm. The parameter space is four-dimensional, and the two-dimensional planes are made by fixing two of the parameters to their median values.

One notices that the different classifiers identify a similar region of maximum probability for a given pair of features. The shape and the gradients of these regions, however, vary for each algorithm. This is due to the differences in the selection algorithms and possible projection effects when the boundaries are represented on the planes. In the case of the single classifiers (top three rows: self-organising map, neural network, and support vector classifier) they are compact and well-defined, unlike for the cases of voting classifiers based on decision trees (bottom three rows). Also, the contours and gradients are less smooth for self-organising map than for the neural network and the

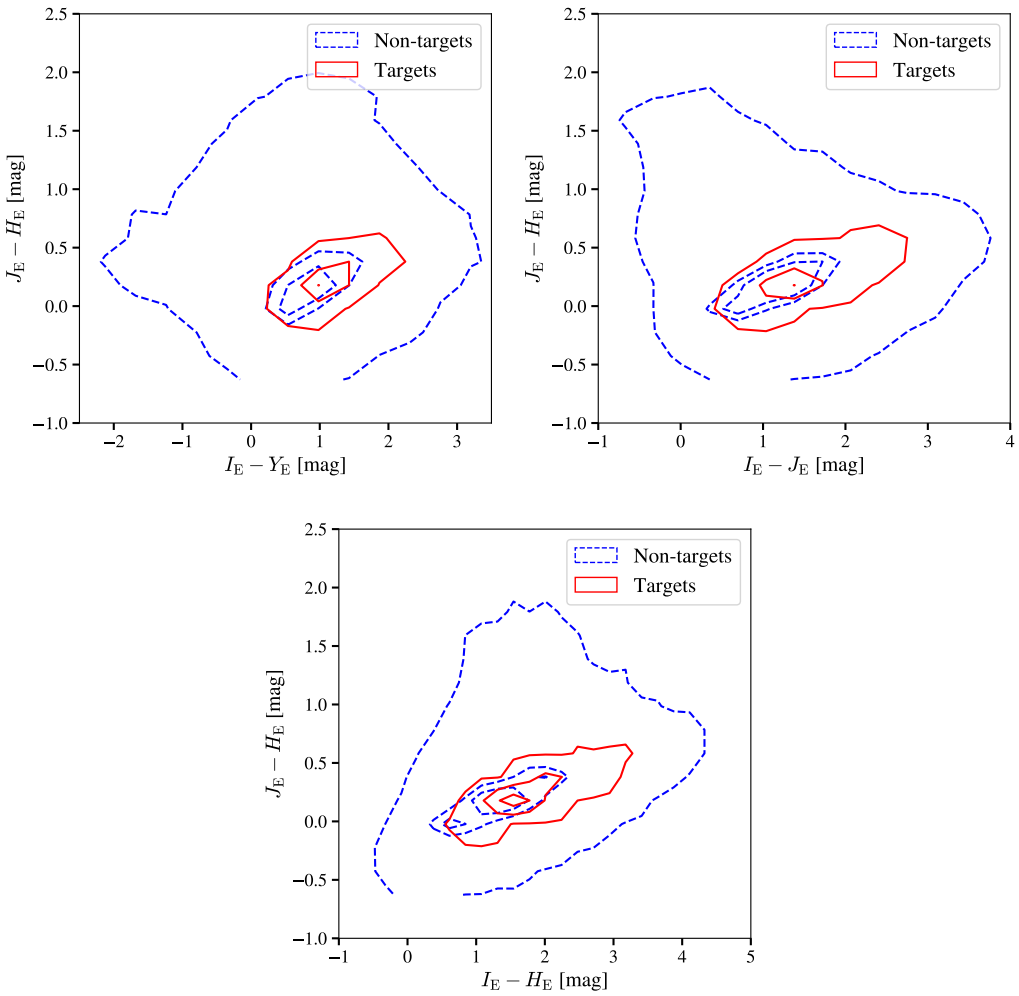

Figure A.1: Target and non-target distributions in colour-colour and colour-magnitudes planes for the Flagship2 catalogue. The contours contain 99%, 50%, and 25% of the samples.

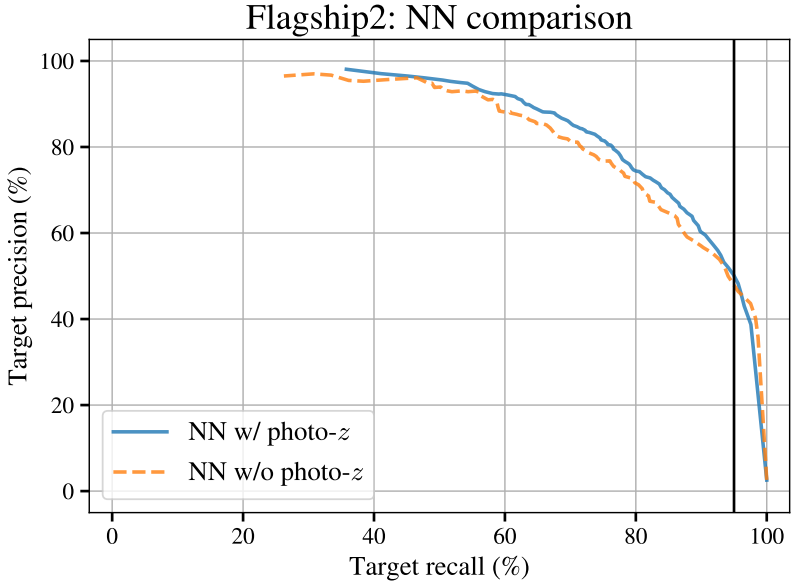

Figure A.2: Comparison of the precision versus recall curves of two neural networks trained with and without photo-zs as an input feature. The two neural networks were trained with *Euclid* and ground-based photometry, but, in the case of the solid blue line, the algorithm takes the photo-z of the galaxy as an additional feature. The solid vertical line corresponds to 95% recall.

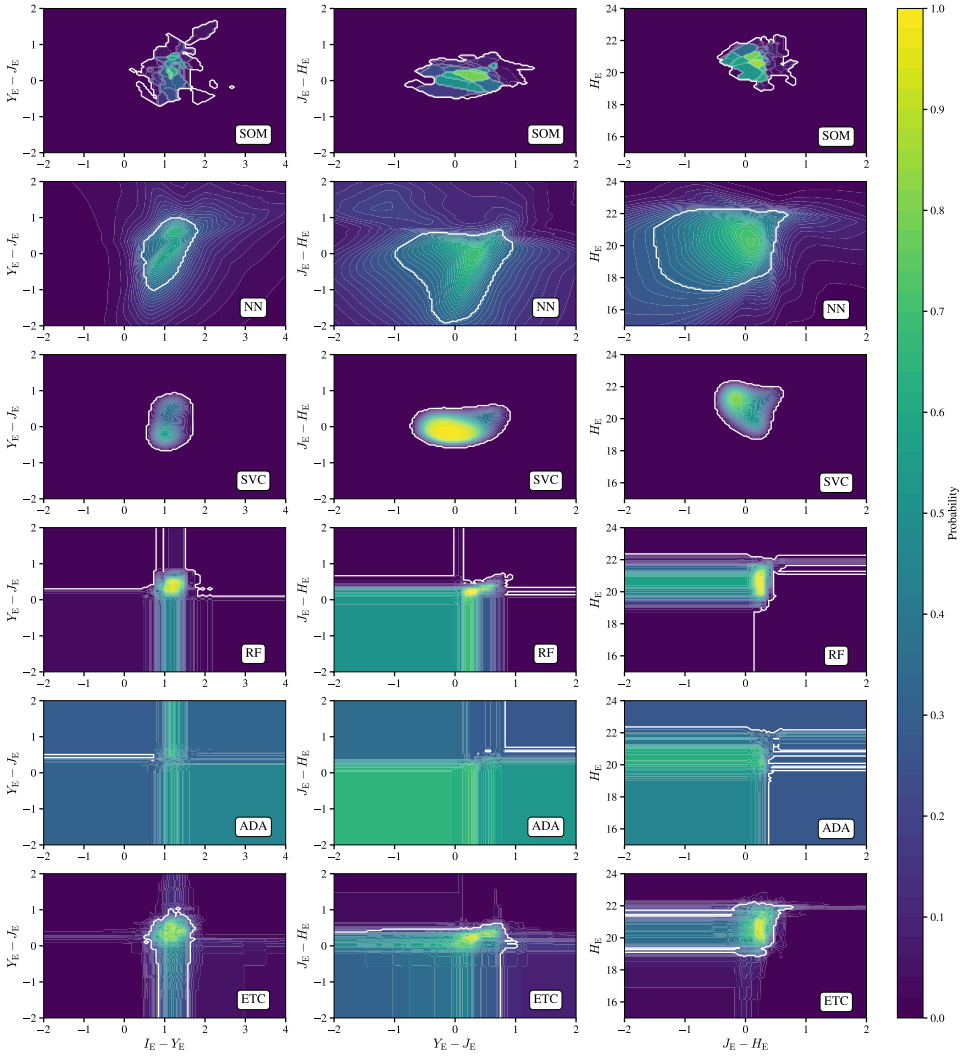

Figure A.3: Probability maps in colour-colour and colour-magnitude planes, for the six classifiers tested in this paper, trained using Flagship2 *Euclid* photometry only. The thick white contour marks the probability threshold that gives 95% recall.

support vector classifier. The probability gradient of the support vector classifier is very steep, especially in comparison to the neural network.

The probability maps for the voting classifiers (bottom three rows) show orthogonal contours. This is due to the common base classifier of these algorithms, the decision tree, which tends to produce decision rules orthogonal to one another. At the same time, the three algorithms have very different probability contours. These differences are related to the batch selection rule used to train the decision trees (see Sect. 5.2). We expect that the algorithms that give a classification model with irregular and steep contours (such as the self-organising map) or stepped contours (such as the decision trees) will be prone to over-fitting and will show poorer performance than algorithms that give smooth probability contours.

Fitting $b_\phi\, f_{\rm NL}$

		C.L.	Linear	NN
NGC	FKP	68%	$131 < b_\phi\, f_{\rm NL} < 310$	$2 < b_\phi\, f_{\rm NL} < 192$
		95%	$35 < b_\phi\, f_{\rm NL} < 394$	$-102 < b_\phi\, f_{\rm NL} < 280$
	Optimal $p = 1.0$	68%	$-27 < b_\phi\, f_{\rm NL} < 200$	$-115 < b_\phi\, f_{\rm NL} < 123$
		95%	$-174 < b_\phi\, f_{\rm NL} < 288$	$-255 < b_\phi\, f_{\rm NL} < 227$
	Optimal $p = 1.6$	68%	$-176 < b_\phi\, f_{\rm NL} < 121$	$-222 < b_\phi\, f_{\rm NL} < 76$
		95%	$-371 < b_\phi\, f_{\rm NL} < 234$	$-409 < b_\phi\, f_{\rm NL} < 192$
SGC	FKP	68%	$-48 < b_\phi\, f_{\rm NL} < 198$	$-83 < b_\phi\, f_{\rm NL} < 188$
		95%	$-162 < b_\phi\, f_{\rm NL} < 322$	$-207 < b_\phi\, f_{\rm NL} < 323$
	Optimal $p = 1.0$	68%	$-93 < b_\phi\, f_{\rm NL} < 175$	$-129 < b_\phi\, f_{\rm NL} < 174$
		95%	$-202 < b_\phi\, f_{\rm NL} < 317$	$-230 < b_\phi\, f_{\rm NL} < 349$
	Optimal $p = 1.6$	68%	$-121 < b_\phi\, f_{\rm NL} < 165$	$-144 < b_\phi\, f_{\rm NL} < 182$
		95%	$-237 < b_\phi\, f_{\rm NL} < 311$	$-265 < b_\phi\, f_{\rm NL} < 354$

Table B.1: Summary on the 68% and 95% constraints on $b_\phi\, f_{\rm NL}$ for the NGC and SGC.

Table B.1 summarises the constraint on $b_\phi f_{\rm NL}$ for the different measurements of the power spectrum. As in Chapt. 6, all bounds are compatible with zero PNG with the exception of the FKP analysis of NGC. Figure B.1 shows the 2D posterior of the parameters in SGC.

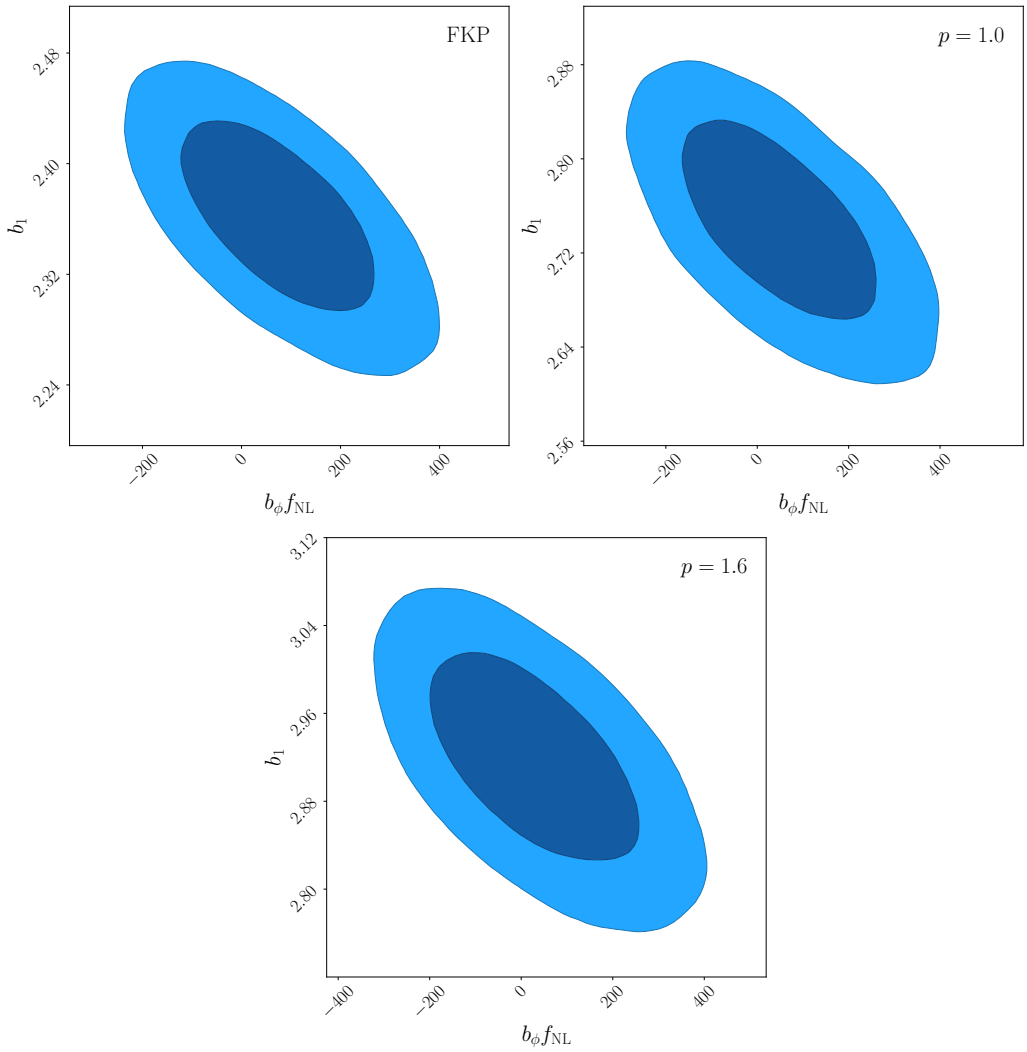

Figure B.1: Two dimensional posterior distributions for $b_\phi\ f_{\rm NL}$ and and the quasar linear bias, b_1, from the SGC linear catalogue analysis. The posterior distribution of the analysis with the FKP weighting scheme, the optimal weights with $p = 1.0$, and $p = 1.6$ are shown respectively on the top left, top right, and bottom panel.

Bibliography

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org

Abazajian, K. N., Adshead, P., Ahmed, Z., et al. 2016, arXiv:1610.02743

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, *Phys. Rev. D*, 98, 043526

Abbott, T. M. C., Aguena, M., Alarcon, A., et al. 2022, *Phys. Rev. D*, 105, 023520

Achúcarro, A., Biagetti, M., Braglia, M., et al. 2022, arXiv:2203.08128

Addison, G. E., Bennett, C. L., Jeong, D., Komatsu, E., & Weiland, J. L. 2019, *Astrophys. J.*, 879, 15

Adler, R. J., Casey, B., & Jacob, O. C. 1995, *Am. J. of Phys.*, 63, 620

Agarap, A. F. 2018, arXiv:1803.08375

Agarwal, N., Desjacques, V., Jeong, D., & Schmidt, F. 2021, *JCAP*, 3, 021

Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv:1902.05569

Alam, S., Ata, M., Bailey, S., et al. 2017, *Mon. Not. R. Astron. Soc.*, 470, 2617

Alarcon, A., Gaztanaga, E., Eriksen, M., et al. 2021, *Mon. Not. R. Astron. Soc.*, 501, 6103

Alvarez, M., Baldauf, T., Bond, J. R., et al. 2014, arXiv:1412.4671

Ansari, R., Arena, E. J., Bandura, K., et al. 2018, arXiv:1810.09572

Aragon-Calvo, M. A., van de Weygaert, R., Jones, B. J. T., & Mobasher, B. 2015, *Mon. Not. R. Astron. Soc.*, 454, 463

Arnouts, S., Moscardini, L., Vanzella, E., et al. 2002, *Mon. Not. R. Astron. Soc.*, 329, 355

Barreira, A. 2022a, *JCAP*, 2022, 013

Barreira, A. 2022b, *JCAP*, 2022, 033

Barreira, A., Cabass, G., Schmidt, F., Pillepich, A., & Nelson, D. 2020, *JCAP*, 2020, 013

Barreira, A. & Krause, E. 2023, *JCAP*, 2023, 044

Bauer, E. & Kohavi, R. 1999, *Machine learning*, 36, 105

Baum, W. A. 1957, *Astron. J.*, 62, 6

Baumann, D. 2011, in Theoretical Advanced Study Institute in Elementary Particle Physics: Physics of the Large and the Small, 523

Bautista, J. E., Paviot, R., Vargas Magaña, M., et al. 2021, *Mon. Not. R. Astron. Soc.*, 500, 736

Beck, R. & Sadowski, P. 2019, Refined Redshift Regression in Cosmology with Graph Convolution Networks, https://ml4physicalsciences.github.io/2019/files/NeurIPS_ML4PS_2019_80.pdf

Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv:1403.5237

Betancourt, M. 2017, arXiv:1701.02434
Beutler, F., Castorina, E., & Zhang, P. 2019, *JCAP*, 2019, 040
Biagetti, M. 2019, *Galaxies*, 7, 71
Biagetti, M., Lazeyras, T., Baldauf, T., Desjacques, V., & Schmidt, F. 2017, *Mon. Not. R. Astron. Soc.*, 468, 3277
Bianchi, D., Gil-Marín, H., Ruggeri, R., & Percival, W. J. 2015, *Mon. Not. R. Astron. Soc.*, 453, L11
Biondo, S. M., Ramos, E. M., Deiros, M. M., et al. 2000, *Journal of the American College of Surgeons*, 191, 191
Blake, C., Brough, S., Colless, M., et al. 2011, *Mon. Not. R. Astron. Soc.*, 415, 2876
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, *Astron. J.*, 154, 28
Blas, D., Lesgourgues, J., & Tram, T. 2011, *JCAP*, 2011, 034
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, *Astron. Astrophys.*, 363, 476
Bond, J. R., Jaffe, A. H., & Knox, L. 2000, *Astrophys. J.*, 533, 19
Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992, in Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92 (New York, NY, USA: Association for Computing Machinery), 144
Bottini, D., Garilli, B., Maccagni, D., et al. 2005, *Publ. Astron. Soc. Pac.*, 117, 996
Boyd, C., Tolson, M. A., & Copes, W. 1987, *The Journal of Trauma: Injury, Infection, and Critical Care*, 27, 27
Bragança, D., Donath, Y., Senatore, L., & Zheng, H. 2023, arXiv:2307.04992
Breiman, L. 2001, Mach. Learn., 45, 5–32
Brescia, M., Cavuoti, S., Razim, O., et al. 2021, *Frontiers in Astronomy and Space Sciences*, 8, 70
Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. 2017, *IEEE Signal Processing Magazine*, 34, 18
Cabass, G., Ivanov, M. M., Philcox, O. H. E., Simonović, M., & Zaldarriaga, M. 2022, *Phys. Rev. D*, 106, 043506
Cabass, G., Ivanov, M. M., Philcox, O. H. E., Simonović, M., & Zaldarriaga, M. 2023, *Physics Letters B*, 841, 137912
Cabass, G., Pajer, E., & Schmidt, F. 2017, *JCAP*, 01, 003
Cagliari, M. S., Castorina, E., Bonici, M., & Bianchi, D. 2023, arXiv:2309.15814
Cagliari, M. S., Granett, B. R., Guzzo, L., et al. 2022, *Astron. Astrophys.*, 660, A9
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, *Astrophys. J.*, 533, 682
Carbone, C., Petkova, M., & Dolag, K. 2016, *JCAP*, 2016, 034
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, *Astrophys. J.*, 712, 511
Castorina, E., Carbone, C., Bel, J., Sefusatti, E., & Dolag, K. 2015, *JCAP*, 2015, 043
Castorina, E. & Di Dio, E. 2022, *JCAP*, 2022, 061
Castorina, E., Feng, Y., Seljak, U., & Villaescusa-Navarro, F. 2018, *Phys. Rev. Lett.*, 121, 101301
Castorina, E., Hand, N., Seljak, U., et al. 2019, *JCAP*, 2019, 010
Castorina, E. & White, M. 2018a, *Mon. Not. R. Astron. Soc.*, 476, 4403
Castorina, E. & White, M. 2018b, *Mon. Not. R. Astron. Soc.*, 479, 741
Cheng, S., Ting, Y.-S., Ménard, B., & Bruna, J. 2020, *Mon. Not. R. Astron. Soc.*, 499, 5902
Chuang, C.-H., Kitaura, F.-S., Prada, F., Zhao, C., & Yepes, G. 2015, *Mon. Not. R. Astron. Soc.*, 446, 2621
Colless, M., Peterson, B. A., Jackson, C., et al. 2003, arXiv:astro-ph/0306581

Collister, A. A. & Lahav, O. 2004, *Publ. Astron. Soc. Pac.*, 116, 345
Comparat, J., Delubac, T., Jouvel, S., et al. 2016, *Astron. Astrophys.*, 592, A121
Creminelli, P. & Zaldarriaga, M. 2004, *JCAP*, 2004, 006
Crill, B. P., Werner, M., Akeson, R., et al. 2020, in Space Telescopes and Instrumentation 2020: Optical, Infrared, and Millimeter Wave, ed. M. Lystrup, M. D. Perrin, N. Batalha, N. Siegler, & E. C. Tong, Vol. 11443, International Society for Optics and Photonics (SPIE), 114430I
Cropper, M., Pottinger, S., Niemi, S., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9904, Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave, ed. H. A. MacEwen, G. G. Fazio, M. Lystrup, N. Batalha, N. Siegler, & E. C. Tong, 99040Q
Cucciati, O., Granett, B. R., Branchini, E., et al. 2014, *Astron. Astrophys.*, 565, A67
Dalal, N., Doré, O., Huterer, D., & Shirokov, A. 2008, *Phys. Rev. D*, 77, 123514
D'Amico, G., Lewandowski, M., Senatore, L., & Zhang, P. 2022, arXiv:2201.11518
de la Torre, S., Jullo, E., Giocoli, C., et al. 2017, *Astron. Astrophys.*, 608, A44
de Mattia, A. & Ruhlmann-Kleider, V. 2019, *JCAP*, 2019, 036
de Mattia, A., Ruhlmann-Kleider, V., Raichoor, A., et al. 2021, *Mon. Not. R. Astron. Soc.*, 501, 5616
de Santi, N. S. M., Shao, H., Villaescusa-Navarro, F., et al. 2023, *Astrophys. J.*, 952, 69
DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv:1611.00036
Dodelson, S. & Schmidt, F. 2020, Modern Cosmology (Elsevier Science)
Dolag, K., Borgani, S., Murante, G., & Springel, V. 2009, *Mon. Not. R. Astron. Soc.*, 399, 497
Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. 1987, *Physics Letters B*, 195, 216
Dunlop, J. S., McLure, R. J., Robertson, B. E., et al. 2012, *Mon. Not. R. Astron. Soc.*, 420, 901
Einstein, A. 1915, *Sitzungsberichte der Koumlniglich Preussischen Akademie der Wissenschaften*, 844
Einstein, A. 1917, *Sitzungsberichte der Koumlniglich Preussischen Akademie der Wissenschaften*, 142
Eisenstein, D. J., Annis, J., Gunn, J. E., et al. 2001, *Astron. J.*, 122, 2267
Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, *Astrophys. J.*, 633, 560
Etherington, J. & Thomas, D. 2015, *Mon. Not. R. Astron. Soc.*, 451, 660
Euclid Collaboration: Blanchard, A., Camera, S., Carbone, C., et al. 2020, *Astron. Astrophys.*, 642, A191
Euclid Collaboration: Desprez, G., Paltani, S., Coupon, J., et al. 2020, *Astron. Astrophys.*, 644, A31
Euclid Collaboration: Gabarra, L., Mancini, C., Rodriguez Munoz, L., et al. 2023, arXiv:2302.09372
Euclid Collaboration: Pocino, A., Tutusaus, I., Castander, F. J., et al. 2021, *Astron. Astrophys.*, 655, A44
Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, *Astron. Astrophys.*, 662, A112
Euclid Collaboration: Schirmer, M., Jahnke, K., Seidel, G., et al. 2022, *Astron. Astrophys.*, 662, A92
Feldman, H. A., Kaiser, N., & Peacock, J. A. 1994, *Astrophys. J.*, 426, 23
Finoguenov, A., Guzzo, L., Hasinger, G., et al. 2007, *Astrophys. J. Suppl.*, 172, 182

Flaugher, B. 2005, *International Journal of Modern Physics A*, 20, 3121
Freund, Y. & Schapire, R. E. 1997, *Journal of Computer and System Sciences*, 55, 119
Friedmann, A. 1922, *Zeitschrift fur Physik*, 10, 377
Ge, H., Xu, K., & Ghahramani, Z. 2018, in International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain, 1682
Gelman, A. & Rubin, D. B. 1992, *Statistical Science*, 7, 457
Geurts, P., Ernst, D., & Wehenkel, L. 2006, *Machine learning*, 63, 3
Gil-Marín, H., Noreña, J., Verde, L., et al. 2015, *Mon. Not. R. Astron. Soc.*, 451, 539
Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. 2017, arXiv:1704.01212
Gonçalves, R. S., Carvalho, G. C., Andrade, U., et al. 2021, *JCAP*, 2021, 029
Gong, L. & Cheng, Q. 2018, arXiv:1809.02709
Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (MIT Press), `http://www.deeplearningbook.org`
Grattarola, D. & Alippi, C. 2020, arXiv:2006.12138
Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, *Astron. J.*, 131, 2332
Guzzo, L., Pierleoni, M., Meneux, B., et al. 2008, *Nature*, 451, 541
Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, *Astron. Astrophys.*, 566, A108
Guzzo, L. & Vipers Team. 2017, *The Messenger*, 168, 40
Hahn, C., Eickenberg, M., Ho, S., et al. 2023a, *JCAP*, 2023, 010
Hahn, C., Lemos, P., Parker, L., et al. 2023b, arXiv:2310.15246
Hamilton, A. J. S. 1998, in Astrophysics and Space Science Library, Vol. 231, The Evolving Universe, ed. D. Hamilton, 185
Hamilton, W. L. 2020, Synthesis Lectures on Artificial Intelligence and Machine Learning, 14, 14
Hand, N., Feng, Y., Beutler, F., et al. 2018, *Astron. J.*, 156, 160
Hand, N., Li, Y., Slepian, Z., & Seljak, U. 2017, *JCAP*, 2017, 002
Hartlap, J., Simon, P., & Schneider, P. 2007, *Astron. Astrophys.*, 464, 399
Hastie, T., Tibshirani, R., & Friedman, J. 2001, The Elements of Statistical Learning, Springer Series in Statistics (New York, NY, USA: Springer New York Inc.)
Henghes, B., Thiyagalingam, J., Pettitt, C., Hey, T., & Lahav, O. 2022, *Mon. Not. R. Astron. Soc.*, 512, 1696
Hockney, R. W. & Eastwood, J. W. 1981, Computer Simulation Using Particles
Hoffman, M. D. & Gelman, A. 2011, arXiv:1111.4246
Hornik, K., Stinchcombe, M., & White, H. 1989, *Neural Networks*, 2, 2
Hubble, E. 1929, *Proceedings of the National Academy of Science*, 15, 168
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, *Astron. Astrophys.*, 457, 841
Ilbert, O., Capak, P., Salvato, M., et al. 2009, *Astrophys. J.*, 690, 1236
Ioffe, S. & Szegedy, C. 2015, arXiv:1502.03167
Jasche, J., Kitaura, F. S., Wandelt, B. D., & Enßlin, T. A. 2010, *Mon. Not. R. Astron. Soc.*, 406, 60
Jasche, J. & Wandelt, B. D. 2012, *Mon. Not. R. Astron. Soc.*, 425, 1042
Jeffrey, N. & Wandelt, B. D. 2020, arXiv:2011.05991
Kaiser, N. 1986, in NATO Advanced Study Institute (ASI) Series C, Vol. 180, Galaxy Distances and Deviations from Universal Expansion, ed. B. F. Madore & R. B. Tully, 271–272
Karagiannis, D., Lazanu, A., Liguori, M., et al. 2018, *Mon. Not. R. Astron. Soc.*, 478, 1341

Kingma, D. P. & Ba, J. 2014, arXiv:1412.6980
Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. 2017, arXiv:1706.02515
Kohonen, T. 1982, *Biological Cybernetics*, 43, 59
Kohonen, T. 1990, *Proceedings of the IEEE*, 78, 1464
Kolmogorov, A. L. 1933, *G. Ist. Ital. Attuari*, 4, 4
Kullback, S. & Leibler, R. A. 1951, *The Annals of Mathematical Statistics*, 22, 22
Lahav, O. 1994, *Vistas in Astronomy*, 38, 251
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, *Astrophys. J. Suppl.*, 224, 24
Landy, S. D. & Szalay, A. S. 1993, *Astrophys. J.*, 412, 64
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193
Laurent, P., Eftekharzadeh, S., Le Goff, J.-M., et al. 2017, *JCAP*, 2017, 017
Le Fèvre, O., Saisse, M., Mancini, D., et al. 2003, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes, ed. M. Iye & A. F. M. Moorwood, 1670–1681
Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, *Astron. Astrophys.*, 439, 845
LeCun, Y. 1989, Generalization and network design strategies, ed. R. Pfeifer, Z. Schreter, F. Fogelman, & L. Steels (Elsevier)
LeCun, Y., Bengio, Y., & Hinton, G. 2015, *Nature*, 521, 436
Lemaître, G. 1927, *Annales de la Société Scientifique de Bruxelles*, 47, 49
Lemos, P., Parker, L. H., Hahn, C., et al. 2023, in Machine Learning for Astrophysics, 18
Lesgourgues, J. 2011, arXiv:1104.2932
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. 2017, arXiv:1708.02002
Linder, E. V. 2003, *Phys. Rev. Lett.*, 90, 091301
LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv:0912.0201
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, *Astrophys. J. Suppl.*, 250, 8
Maciaszek, T., Ealet, A., Gillard, W., et al. 2022, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 12180, Space Telescopes and Instrumentation 2022: Optical, Infrared, and Millimeter Wave, ed. L. E. Coyle, S. Matsuura, & M. D. Perrin, 121801K
MacQueen, J. B. 1967
Madau, P. 1995, *Astrophys. J.*, 441, 18
Malavasi, N., Pozzetti, L., Cucciati, O., Bardelli, S., & Cimatti, A. 2016, *Astron. Astrophys.*, 585, A116
Maldacena, J. 2003, *JHEP*, 2003, 013
Maraston, C. 2005, *Mon. Not. R. Astron. Soc.*, 362, 799
Masters, D., Capak, P., Stern, D., et al. 2015, *Astrophys. J.*, 813, 53
Monaco, P., Theuns, T., & Taffoni, G. 2002, *Mon. Not. R. Astron. Soc.*, 331, 587
Moosavi, V., Packmann, S., & Vallés, I. 2014, SOMPY: A Python Library for Self Organizing Map (SOM), available at github.com/sevamoo/SOMPY
Moutard, T., Arnouts, S., Ilbert, O., et al. 2016, *Astron. Astrophys.*, 590, A102
Mueller, E.-M., Rezaie, M., Percival, W. J., et al. 2022, *Mon. Not. R. Astron. Soc.*, 514, 3396
Munari, E., Monaco, P., Sefusatti, E., et al. 2017, *Mon. Not. R. Astron. Soc.*, 465, 4658
Murtagh, F. 1991, *Neurocomputing*, 2, 2
Neal, R. 2011, in Handbook of Markov Chain Monte Carlo, 113–162
Newman, J. A. 2008, *Astrophys. J.*, 684, 88
Newman, J. A. & Gruen, D. 2022, *Annual Review of Astronomy and Astrophysics*, 60, 363

O'Shea, K. & Nash, R. 2015, arXiv:1511.08458
Pardede, K., Rizzo, F., Biagetti, M., et al. 2022, *JCAP*, 2022, 066
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, *Astron. Astrophys.*, 621, A26
Paszke, A., Gross, S., Chintala, S., et al. 2017, in NIPS-W
Peacock, J. A. & Nicholson, D. 1991, *Mon. Not. R. Astron. Soc.*, 253, 307
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, 12, 2825
Peebles, P. J. E. 1980, The large-scale structure of the universe
Perlmutter, S., Aldering, G., della Valle, M., et al. 1998, *Nature*, 391, 51
Pezzotta, A., de la Torre, S., Bel, J., et al. 2017, *Astron. Astrophys.*, 604, A33
Philcox, O. H. E. 2021a, *Mon. Not. R. Astron. Soc.*, 501, 4004
Philcox, O. H. E. 2021b, *Phys. Rev. D*, 103, 103504
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, *Astron. Astrophys.*, 571, A16
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020a, *Astron. Astrophys.*, 641, A6
Planck Collaboration, Akrami, Y., Arroja, F., et al. 2020b, *Astron. Astrophys.*, 641, A9
Pozzetti, L., Hirata, C. M., Geach, J. E., et al. 2016, *Astron. Astrophys.*, 590, A3
Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, *Astron. J.*, 165, 126
Ravanbakhsh, S., Oliva, J., Fromenteau, S., et al. 2017, arXiv:1711.02033
Régaldo-Saint Blancard, B., Hahn, C., Ho, S., et al. 2023, arXiv:2310.15250
Rezaie, M., Ross, A. J., Seo, H.-J., et al. 2023, arXiv:2307.01753
Rezaie, M., Ross, A. J., Seo, H.-J., et al. 2021, *Mon. Not. R. Astron. Soc.*, 506, 3439
Riess, A. G., Casertano, S., Yuan, W., et al. 2021, *Astrophys. J. Lett.*, 908, L6
Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, *Astron. J.*, 116, 1009
Ross, A. J., Bautista, J., Tojeiro, R., et al. 2020, *Mon. Not. R. Astron. Soc.*, 498, 2354
Ross, A. J., Ho, S., Cuesta, A. J., et al. 2011, *Mon. Not. R. Astron. Soc.*, 417, 1350
Rota, S., Granett, B. R., Bel, J., et al. 2017, *Astron. Astrophys.*, 601, A144
Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, *Astronomy and Computing*, 10, 121
Sadeh, I., Abdalla, F. B., & Lahav, O. 2016, *Publ. Astron. Soc. Pac.*, 128, 104502
Sailer, N., Castorina, E., Ferraro, S., & White, M. 2021, *JCAP*, 2021, 049
Saito, S., de la Torre, S., Ilbert, O., et al. 2020, *Mon. Not. R. Astron. Soc.*, 494, 199
Salvato, M., Ilbert, O., & Hoyle, B. 2019, *Nature Astronomy*, 3, 212
Scoccimarro, R. 2015, *Phys. Rev. D*, 92, 083532
Scoccimarro, R., Sheth, R. K., Hui, L., & Jain, B. 2001, *Astrophys. J.*, 546, 20
Scodeggio, M., Guzzo, L., Garilli, B., et al. 2018, *Astron. Astrophys.*, 609, A84
Scoville, N., Aussel, H., Brusa, M., et al. 2007, *Astrophys. J. Suppl.*, 172, 1
Senatore, L. & Zaldarriaga, M. 2012, *JHEP*, 2012, 24
Slosar, A., Hirata, C., Seljak, U., Ho, S., & Padmanabhan, N. 2008, *JCAP*, 2008, 031
Smirnov, N. 1948, *The Annals of Mathematical Statistics*, 19, 19
Springel, V. 2005, *Mon. Not. R. Astron. Soc.*, 364, 1105
Springel, V., Yoshida, N., & White, S. D. M. 2001, *New Astron.*, 6, 79
Stanford, S. A., Masters, D., Darvish, B., et al. 2021, *Astrophys. J. Suppl.*, 256, 9
Sullivan, J. M., Prijon, T., & Seljak, U. 2023, *JCAP*, 2023, 004
Tanaka, M., Coupon, J., Hsieh, B.-C., et al. 2018, *Publications of the Astronomical Society of Japan*, 70, S9
Tegmark, M. 1997, *Phys. Rev. D*, 55, 5895

Tegmark, M., Eisenstein, D. J., Strauss, M. A., et al. 2006, *Phys. Rev. D*, 74, 123507
Tegmark, M., Hamilton, A. J. S., Strauss, M. A., Vogeley, M. S., & Szalay, A. S. 1998, *Astrophys. J.*, 499, 555
Tonry, J. & Davis, M. 1979, *Astron. J.*, 84, 1511
Tosone, F., Cagliari, M. S., Guzzo, L., Granett, B. R., & Crespi, A. 2023, *Astron. Astrophys.*, 672, A150
Valogiannis, G., Yuan, S., & Dvorkin, C. 2023, arXiv:2310.16116
Verde, L., Treu, T., & Riess, A. G. 2019, *Nature Astronomy*, 3, 891
Viel, M., Haehnelt, M. G., & Springel, V. 2010, *JCAP*, 2010, 015
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021, arXiv:2109.09747
Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., et al. 2022, *Astrophys. J. Suppl.*, 259, 61
Villaescusa-Navarro, F., Wandelt, B. D., Anglés-Alcázar, D., et al. 2020, arXiv:2011.05992
Villanueva-Domingo, P. & Villaescusa-Navarro, F. 2022, *Astrophys. J.*, 937, 115
Wang, Y., Sun, Y., Liu, Z., et al. 2018, arXiv:1801.07829
Weaver, J. R., Kauffmann, O., Shuntov, M., et al. 2021, in American Astronomical Society Meeting Abstracts, Vol. 53, American Astronomical Society Meeting Abstracts, 215.06
Wilson, D., Nayyeri, H., Cooray, A., & Häußler, B. 2020, *Astrophys. J.*, 888, 83
Wilson, M. J., Peacock, J. A., Taylor, A. N., & de la Torre, S. 2017, *Mon. Not. R. Astron. Soc.*, 464, 3121
Wu, Z., Pan, S., Chen, F., et al. 2019, arXiv:1901.00596
Yamamoto, K., Nakamichi, M., Kamino, A., Bassett, B. A., & Nishioka, H. 2006, *Publ. Astron. Soc. Jpn.*, 58, 93
York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, *Astron. J.*, 120, 1579
Yuan, S., Hadzhiyska, B., Bose, S., & Eisenstein, D. J. 2022, *Mon. Not. R. Astron. Soc.*, 512, 5793
Zaheer, M., Kottur, S., Ravanbakhsh, S., et al. 2017, arXiv:1703.06114
Zhao, C., Chuang, C.-H., Bautista, J., et al. 2021, *Mon. Not. R. Astron. Soc.*, 503, 1149
Zheng, Z., Berlind, A. A., Weinberg, D. H., et al. 2005, *Astrophys. J.*, 633, 791
Zheng, Z., Coil, A. L., & Zehavi, I. 2007, *Astrophys. J.*, 667, 760
Zhou, J., Cui, G., Hu, S., et al. 2018, arXiv:1812.08434

List of Publications

First author publications

- "Euclid: Constraining ensemble photometric redshift distributions with stacked spectroscopy", **M. S. Cagliari**, B. R. Granett, L. Guzzo, M. Bolzonella, L. Pozzetti, I. Tutusaus, S. Camera, and Euclid Collaboration. In: Astronomy & Astrophysics 660 (2022), A9.

Co-author publications

- "Augmenting photometric redshift estimates using spectroscopic nearest neighbours", F. Tosone, **M. S. Cagliari**, L. Guzzo, B. R. Granett, and A. Crespi. In: Astronomy & Astrophysics 672 (2023), A150.

Publications under review

- "Optimal constraints on Primordial non-Gaussianity with the eBOSS DR16 quasars in Fourier space", **M. S. Cagliari**, E. Castorina, M. Bonici, D. Bianchi. Accepted with revision by Journal of Cosmology and Astroparticle Physics.

Publications in preparation

- "Euclid: Testing photometric selection of emission-line galaxy targets", **M. S. Cagliari**, B. R. Granett, L. Guzzo, M. Bertermin, M. Bolzonella, S. de la Torre, P. Monaco, M. Moresco, W. J. Percival, C. Scarlata, Y. Wang, M. Ezziati, O. Ilber, V. Le Brun, and Euclid Collaboration. Under review of the Euclid Consortium Publication Board. To be submitted to Astronomy & Astrophysics.

Acknowledgements

Finally, at the end of this path, I will admit that I am happy that I took it. In these last three years, I learned so many things to improve myself as a person and as a scientist. I also met many incredible people who helped me out and without whom I could have not reached the finish line.

My first thanks go to my supervisors Ben and Gigi, and to Ema, who I consider my mentor. I thank Ben for always being available to help me out with all different kinds of problems and for having both guided me and given me the liberty of choosing my own projects. I am convinced that this was a very precious opportunity. I also thank Gigi, who has always supported my work with his knowledge, his experience, and the enthusiasm I sometimes missed showing me that it is possible to continue loving this work. I also thank him for believing I was worth the expense of sending me to New York City. Last, but not least of these first acknowledgements is Emanuele, whom I thank for teaching me not only science but most importantly what it means to be a researcher and showing me how small my world was. I also thank him for reconciling Marina and Cagliari so many times. I thank all three of them for coping with my bad writing and for writing so many letters for me.

My second and most sincere thanks go to Federico and Marco, my two great senpais. It all started like a roller-coaster to Barolo and it continued as one. Without the two of them, I would have never reached the end. With them, I joked and I learned. The only thing I can hope is to become for someone else a post-doc as good as the two of them were for me. I thank Federico for having helped me out with all the small and trivial problems, for all our Friday discussions about the new manga chapters, and for all the magnificent memes. I also thank Marco for always finding time for me (and Julia), for his not-so-politically-correct jokes, and for the calls at whatever time it was in your time zone. In the end, they showed me that we can take different paths and I know that it is also thanks to them that in the future I will be able to take the one I think is best for me. I hope that we will always remain `bruttaggente`.

I also thank Maria and Davide. I thank Maria for coping with my bad jokes, which I am sure sometimes terrified her, and for listening to my worries and doubts. I thank Davide for all the constructive discussion and all the explanations he gave me. I also thank Petra for her kindness and all the help she gave me with the bureaucracy. Finally, I cannot conclude my thanks to the 'Cosmo@MI' group without thanking our beloved mascot and cluster: Doraemon. The motto is 'who goes slow goes far and goes safe'. Without it, I would have done half of the work I accomplished, but sometimes I would have also done double the work. I will look at the silver lining of all the issues I had with it as opportunities to learn.

I want to thank Ana and Claudia as well. Even though I was not in their groups, they supported me as a young researcher and as a person. I thank Ana for making me part of the plans for all her fabulous jokes to our poor victim Giovanni and for telling me that everything is going to be fine, which it was in the end. I can only hope that everything will continue to be fine and also 'perfectly perfect'. I thank Claudia for being the delight of my lunch breaks and all the funny stories she told us.

Now I finally thank 'Da Office': my three disc boys, Cristiano, Simone, and Pietro, my extra-galactic mate, Giovanni, the little ones, first Rossella and now also Matilde, and Annamaria of course. Sharing the office with them has been so incredibly fun that no words can describe how hard we laughed together. I thank Cristiano for his kindness, his enthusiasm, and our very nice discussions about statistics, for which I will always be there. I thank Simone for sharing so keenly my sense of humour, for our lunch-break 'arguments', and for helping me out with any computer problem. I thank Pietro for showing me what real scientific enthusiasm is, for being a delightful presence in the office, and for his pristine love and respect for Annamaria. Last, but not least I thank the best student of the XXXVI PhD cycle, Giovanni. Even though we joked so much about it I truly believe that he is the best of us. I thank Giovanni for being such a good joke when needed, for helping us with the PhD bureaucracy, and for representing us in these three years. I thank Rossella for sharing my passion for animes and for coping with all the eccentricities of us older students. Finally, I thank Matilde, even though she just arrived it has been exciting to work with her and I hope to continue to do so. I think that together with Rossella she will carry on the legacy of this office. To Rossella and Matilde I wish for a productive PhD. My wise words for them are to travel, to have fun as much as they can, and to not get discouraged by difficult times as everything is going to be fine, as someone once told me.

My hearty thanks go to Pietro (who exists and is unique), who delighted my Sunday afternoons. Together with him, I thank Beatrice and Davide. With the three of them, I shared many adventures and I wish to share much more in the future. I also thank my long-time friends Marta, who I also thank for the very nice back cover illustration, and Marcella, as well as Camilla. I thank my family, to whom I also dedicate this work, as I would not have been able to do any of it without them. Finally, as I am convinced that no thesis acknowledgements should be concluded without citing a pet, I thank Olly, my beloved cat. She slept on the notes of all my exams and studies, so I think I should share with her the merit of this work.