Università degli studi di Napoli *Federico II*

DOTTORATO DI RICERCA IN

FISICA

XXX ciclo

Coordinatore: prof. Salvatore Capozziello

Machine Learning based Probability Density Functions of photometric redshifts and their application to cosmology in the era of dark Universe exploration

Settore scientifico disciplinare:FIS/05

Dottorando
Valeria Amaro

Tutor
Prof. Giuseppe Longo
Dr. Massimo Brescia

Anni 2014/2017

Abstract

The advent of wide, multiband multiepoch digital surveys of the sky has pushed astronomy in the big data era. Instruments, such as the Large Synoptic Survey Telescope or LSST which will become operational in 2020, are in fact capable to produce up to 30 Terabytes of data per night. Such data streams imply that data acquisition, data reduction, data analysis and data interpretation, cannot be performed with traditional methods and that automatic procedures need to be implemented. In other words, Astronomy, like many other sciences, needs the adoption of what has been defined the *fourth paradigm* of modern science: the so called "data driven" or "Knowledge Discovery in Databases - KDD" (after the three older paradigms: theory, experimentation and simulations). With the words "Knowledge Discovery" or "Data Mining" we mean the extraction of useful information from a very large amount of data using automatic or semi-automatic techniques based on Machine Learning, i.e. on algorithms built to teach machines how to perform specific tasks typical of the human brain.

This methodological revolution has led to the birth of the new discipline of Astroinformatics, which, besides the algorithms used to extract knowledge from data, covers also the proper acquisition and storage of the data, their pre-processing and analysis, as well as their distribution to the community of users.

This thesis takes place within the framework defined by this new discipline, since it is mainly concerned the implementation and the application of a new machine learning method to the evaluation of photometric redshifts for the large samples of galaxies produced by the ongoing and future digital surveys of the extragalactic sky. Photometric redshifts (described in Section 1.1) are in fact instrumental to deal with a huge variety of fundamental topics such as: fixing constraints to the dark matter and energy content of the Universe, mapping the galaxy color-redshift relationships, classifying astronomical sources, reconstructing the Large Scale Structure of the Universe through weak lensing, to quote just a few. Therefore, it comes as no surprise that in recent years a plethora of methods capable to calculate photo-z's has been implemented based either on template models fitting and/or on empirical explorations of the photometric parameter space. Among the latter, many are based on machine learning but only a few allow the characterization of the results in terms of a reliable Probability Density Function (PDF).

In fact, machine learning based techniques while on the one end are not explicitly dependent on physical priors and are capable to produce accurate photo-z estimations within the photometric ranges covered by the spectroscopic training set (see Chapter 1, Sec. 1.1.2), on the other are not easy to characterize in terms of PDF, due to the fact that the analytical relation mapping the photometric parameters onto the redshift space is virtually unknown.
In the course of my thesis I contributed to design, implement and test the innovative procedure METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts) capable to provide reliable PDFs of the error distribution for empirical techniques. METAPHOR was implemented as a modular workflow, whose internal engine for photo-z estimation may be either the MLPQNA neural network (Multi Layer Perceptron with Quasi Newton Algorithm) or any other machine learning model capable to perform for regression tasks. This kernel is completed by an algorithm which allows both the calculation of individual source and of stacked objects samples PDFs. More in detail, my work in this context has been: i) the creation of software modules providing some of the functionalities of the entire method and finalised to obtain and analyze the results on a variety of datasets (see the list of publications) and for the EUCLID contest (see below), ii) to provide the algorithms with some workflow facilities and, iii) the debugging of the whole procedure. The first application of METAPHOR was in the framework of the second internal Photo-z challenge of the Euclid[1] consortium: a contest among different European teams, aimed at establishing the best SED fitting and/or empirical methods, which will be included in the official data flow processing pipelines for the mission. This contest, which was the first in a series, lasted from September 2015 until the end of Jenuary 2016, and it was concluded with the releases of the results on the participants performances, in the middle of May 2016.
Finally, iv) I improved the original workflow by adding other statistical estimators needed to better quantify the significance of the results. Through a comparison of the results obtained by METAPHOR and by the SED template fitting method Le-Phare on the SDSS-DR9 (Sloan Digital Sky Survey - Data Release 9) and exploiting the already mentioned modularity of METAPHOR, we verified the reliability of our PDF estimates using three different self-adaptive techniques, namely: MLPQNA, Random Forest and the standard K-Nearest Neighbors models.
In order to further explore ways to improve the overall performances of photo-z methods, I also contributed to the implementation of an hybrid procedure based on the combination of SED template fitting estimates obtained with Le-Phare and of METAPHOR using as test data those extracted from the ESO (European Southern Observatory) KiDS (Kilo Degree

[1]The Euclid consortium is preparing for the launch in 2020 of the Euclid satellite, an ESA cornerstone mission aimed to study the dark component of the Universe.

Survey) Data Release 2[2].
Always in the context of the KiDS survey, I was involved in the creation of a catalogue of ML photo-z's and relative PDFs for the KiDS-DR3 (Data Release 3) survey, widely and exhaustively described in the third official data release of the KiDS data (de Jong et al., 2017). A further work on the KiDS-DR3 data, Amaro et al. (2017), has been submitted to the Montly Notices of the Royal Astronomical Society (MNRAS) and is detailed in Chapter 5. The main topic of this work was to achieve a deeper analysis of some systematics in photo-z PDFs, obtained using different methods, two machine learning models (METAPHOR and ANNz2) and one SED fitting technique (BPZ), through a direct comparison of both cumulative (*stacked*) and individual PDFs. The comparison was made by discriminating between *quantitative* and *qualitative* estimators and using a special *dummy* PDF as benchmark in order to assess their capability to measure the error and the invariance with respect to some types of error sources. In fact, in absence of systematics, there are several factors affecting the reliability of photo-z obtained with ML methods: photometric and internal errors of the methods as well as statistical biases induced by the properties of the training set. The main purpose of the paper was to present for the first time a ML method capable to deal with intrinsic photometric uncertainties. The results of this comparison, along with a discussion of the statistical estimators, have allowed us to conclude that, in order to assess the objective validity and quality of any photo-z PDF method, a combined set of statistical estimators is required.
The last part of my work focused on a natural application of photo-z PDFs to the measurements of Weak Lensing (WL), i.e. the weak distortion of the galaxies images due to the inhomogeneities of the Universe Large Scale Structure (LSS, made up of voids, filaments, halos) along the line of sight (see Chapter 6). The *shear*, or distortion, of the galaxy shapes (ellipticities) due to the presence of matter between the observer and the *lensed* sources, is evaluated through the *tangential* component of the shear. The Excess Surface Density (ESD, i.e. a measurement of the density distribution of the *lenses*), is proportional to the tangential shear, through a *geometrical* factor, which takes into account the angular diameter distances among observer, lens, and *lensed* galaxy source. Such distances in the geometrical factor are measured through photometric redshifts, or better through their full posterior probability distributions.
Such distributions have been measured, to the best of my knowledge, only with template fitting methods: our Machine Learning METAPHOR has therefore been used to make a preliminary comparative study on WL ESD obtained with ML methods, with respect to the

[2]This work is described in the paper of Cavuoti et al. (017a) of which I am a co-author since I provided the ML based redshifts. This work however is not included in this thesis since it was just an application of the photo-z I derived and therefore is outside of the mainstream of my work.

SED fitter results. Furthermore, a confrontation between the ESD estimates obtained by using both METAPHOR PDFs and photo-z punctual estimates has been performed. The outcomes of this preliminary work (which was started during the final months of my PhD) are very promising since we found that the use of punctual estimates and relative PDFs leads to indistinguishable results, at least within the limits imposed by the required accuracy. Most importantly, we found a similar trend for the ESD obtained with ML and SED template fitting methods, despite all the limits of Machine Learning techniques (incompleteness of the training dataset, low reliability for results extrapolated outside the knowledge base) which are usually assumed to be crucial in WL studies. These still preliminary results are outlined in Chapter 6.

Table of contents

List of figures

List of tables

Chapter 1

Photometric redshift and their probability density function

1.1 Photometric redshifts

Due to the large number of multiband photometric digital sky surveys either ongoing (e.g. Kilo-Degree Survey (KiDS), de Jong et al. 2015, 2017; Dark Energy Survey (DES[1]) ; Panoramic Survey Telescope and Rapid Response System (Pan-STARRS, Flewelling et al. 2016), or planned for the near future (e.g. Large Synoptic Survey Telescope (LSST, Ivezic 2009) and Euclid (Euclid 2011; Laureijs et al. 2014)), astronomy has entered the era of precision cosmology in which the observed properties for hundreds of millions to billions of galaxies are being systematically collected. However, for many if not all the scientific tasks, the exploitation of these large datasets requires an additional piece of information which is not easy to obtain: the distance. Distances of extragalactic objects are usually derived from the Hubble expansion law using the so called redshift (i.e. the displacement of a given spectral line from its rest frame position) defined as:

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}} \tag{1.1}$$

The cosmological redshift is related to the scale factor a which embeds the expansion of the universe:

$$1 + z = \frac{a_0}{a} \tag{1.2}$$

where a_0 is the value of the factor at $z = 0$ and is equal to 1. The cosmological redshift should not be interpreted as a galaxy recession velocity as in the case of special relativity

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

theory, except for very low redshifts $z < 0.13$. For these redshifts, the Hubble law, i.e. the proportionality between the galaxy recession velocity $v_{recession}$ and the source distance D via the Hubble constant H_0:

$$D = \frac{v_{recession}}{H_0} \tag{1.3}$$

is valid. The Hubble equation 1.3, combined with the relativistic redshift given in function of the radial component of the source recession velocity:

$$z = \sqrt{\frac{1 + v_r/c}{1 - v_r/c}} - 1 \tag{1.4}$$

in its first order expansion:

$$z = \frac{v_r}{c} \tag{1.5}$$

leads to the relation between the redshift and the Hubble constant:

$$D = \frac{cz}{H_0} \tag{1.6}$$

Equation 1.6 does not include the effects of the expanding, curved space-time of the Universe[2].

The most reliable way to derive the redshift of an object is through spectroscopy, by observing the displacement of specific emission of absorption lines. This approach, being too much time consuming in terms of precious observing time, fails to meet the needs of modern precision which calls for very large samples of objects. To quote just an example, even the Sloan Digital Sky Survey (SDSS; York 2000), which has covered almost one half of the sky and derived accurate photometric information for hundreds of millions sources, has obtained spectra only for a subsample of objects one hundred times smaller than the photometric one. This is why since the early ’980’s an alternative approach to the evaluation of redshift via broad band photometry has become increasingly popular: the so called "photometric redshift techniques" to which this thesis is devoted.

In spite of its relative novelty, this technique has already proven crucial to many tasks: in constraining dark matter and dark energy contents of the Universe through weak gravitational lensing (cf. Hildebrandt et al. 2017; Serjeant 2014), in reconstructing the Universe’s large scale structure (cf. Aragon-Calvo 2015); in identifying galaxies clusters and groups (cf. Capozzi et al. 2009); in classifying different types of astronomical sources (cf. Brescia et al. 2012), to quote just a few.

[2] An alternative source of redshift is the gravitational redshift predicted by the General Theory of Relativity which, however, is not described here since it is beyond the scopes of the present work.

Due to the specific needs of some of the above reported science cases and mainly of Weak Lensing (as we will see in Chapter 6), over the last few years particular focus had been put on the implementation of methods capable to compute a full Probability Density Function for both an individual galaxy as well as for an entire galaxy sample. In section 5.2 it will be shown that the single source PDF contains more information with respect to the simple estimate of a redshift together with its error, and it has been confirmed by the improvement in the accuracy of cosmological measurements (Mandelbaum et al., 2008).
In a broad over-simplification we can split photo-z in the classes: the template Spectral Energy Distribution (SED) fitting methods (hereafter, SED fitters, e.g., Benitez 2000; Bolzonella et al. 2000; Tanaka 2015) and the empirical methods (Bonnet 2013; Brescia et al. 014a; Carrasco and Brunner 013a; Cavuoti et al. 2015; Tagliaferri et al. 2002), both characterized by some advantages and many shortcomings. In order to understand how a probability density function can be obtained from these two methods, a short introduction is needed.

1.1.1 The template fitting methods for photometric redshift

Template SED fitting methods are based on a fit (generally, a χ^2 minimization) of the multi-band photometric measurements of a given object with theoretical or observed template SEDs taken at zero redshift, after these templates have been convolved with the transmission functions of the filters (for a specific survey) in order to create synthetic magnitudes as function of the redshift for each galaxy template. Therefore, these methods depend strictly on the library of template SED used and on the accuracy of the transmission curves of the filters used. The libraries usually consist of a set of few SED templates, covering different nuances of Elliptical, Spiral and Irregular galaxies, while the spectra for intermediate types are usually obtained by interpolating between this base of spectral templates.
The great advantage of SED methods is that they are capable to determine at once the photo-z, the spectral type and the PDF of the photo-z error distribution for a source. Furthermore SED fitters can be applied also to objects fainter than the so called "spectroscopic limit", i.e. to objects too faint to be observed spectroscopically. They, however, suffer of several cumbersome disadvantages, among which the greatest is the potential mismatch between the templates used for the fitting and the properties of the specific object or of a sample of galaxies for which one wants to estimate the redshifts (cf. Abdalla et al. 2011). Such problems become more cogent at high redshift, where not only galaxies are fainter and photometric error higher, but also there are few or no empirical spectra to enrich the template library.
The problem has been clearly posed by Benitez (2000):"...The statistical maximum likelihood approach allows to determine the redshift of a source, exclusively basing its choice on

the goodness of the fit between observed colors and templates: in case of redshift/color degeneracies the likelihood would have two or more approximately equally high maxima at different redshift...". However, additional information can help to solve these degeneracy issues.
In fact, in the context of a Bayesian inference approach, the use of priors, e.g. of additional information to be added to the data (for example, we could know that one of the possible redshift/type combinations is much more likely than any other, given the galaxy magnitude, angular size, shape etc.) can allow to disentangle to a certain degree the redshift/color degeneracy.
Template methods differ in their specific implementation but are basically the same in their nature. The description of one SED fitter, Le-Phare, will be given, in higher detail, in Sec. 4.3.2 (its conceptual framework is substantially equal to that of the most famous SED fitter, i.e. that of Benitez (2000), called BPZ). In that section I will also clarify how it is possible to obtain automatically a PDF for this family of methods.
We anticipate that BPZ makes use of Bayesian inference to quantify the relative probability that each template matches the galaxy input photometry and then determines a photo-z PDF by computing the posterior probability that a given galaxy is at a given redshift. Following the synthesis of such method, as presented in Carrasco and Brunner (2013), we can introduce the probability

$$P(z|\mathbf{x}) \tag{1.7}$$

for a specific template t, where $\mathbf{x}$ represents a given set of photometric features (magnitudes or colors). Marginalizing on the entire set of templates T, and using the Bayes theorem we have :

$$P(z|\mathbf{x}) = \sum_{t \in T} P(z,t|\mathbf{x}) \propto \sum_{t \in T} L(\mathbf{x}|z,t)P(z,t) \tag{1.8}$$

where $L(\mathbf{x}|z,t)$ is the likelihood that, for a given template t and redshift z, a galaxy has that set of magnitudes or colors. The probability $P(z,t)$ is a prior probability of a galaxy that, at redshift z, has the spectral type t. This prior probability can be calculated from a spectroscopic sample, if available, or from other type of information (as said before) such, for example, stellar population synthesis (SPS) models (cf. Tanaka 2015) based on the knowledge of the evolution of galaxy properties across cosmic times. The prior approach is the same, regardless the specific template method used: the priors are let to evolve with

redshift in order that the library templates, calibrated by the priors themselves, achieve an evolution compatible with the observations. The use of priors is not always necessary, but recommended also when we have deep and well-built data sets, in order to weaken, if not overcome, the redshift/color degeneracy.
In any case, focusing on equation 1.8, we can grasp, that, in the case of SED fitting methods, the photo-z PDF is either the posterior probability, defined above in the case of used priors, or the likelihood itself, if no prior is used. Therefore, PDF can be considered an automatic by-product for SED fitting methods. Among the different template fitters, besides the above mentioned BPZ and Le-Phare (Arnouts et al. 1999; Ilbert et al. 2006), we cite also ZEBRA (Feldmann et al., 2006).

1.1.2 Machine Learning Methods for photometric redshifts

Machine Learning methods (MLMs) can be sub-divided in two classes of approaches, supervised and unsupervised. The former are able to reconstruct not analytically the hidden relation between an input (in the case of photo-z multi-band photometry: fluxes, magnitudes and/or derived colors) and a desired output (the spectroscopic redshift, hereafter spec-z). This is possible due to the existence of a correlation unknown and highly non linear, between the photometric properties of a galaxy and its redshift. In supervised techniques, the learning process is controlled by the spectroscopic information. Instead, in the unsupervised approach, the spectroscopic information is not used in the training phase, but only at the end of the training phase, in order to validate the results and make predictions.
One of the greatest disadvantage of these methods (of both supervised and unsupervised approaches) is their incapability of extrapolating information outside the ranges of values covered by training data, so that, for instance, it is not possible to estimate a redshift for objects fainter than the spectroscopic limit. Possible ways to overcome this limitations have been recently proposed (cf. Hoyle et al. 2015) but their validity needs still to be tested.
Indeed, such methods are applicable only if accurate photometry is available for a large number of objects and if accurate spectroscopy is given for a statistically and representative sub-sample of the original sample itself. This subsample is usually called "knowledge base" and is used to train and test the network. One of the greatest advantages of MLMs with respect to SED fitters is the ease of incorporating additional information into the inference (with respect to the handling of the priors), like for example one could add, if known, the surface brightness of galaxies, which has a $(1+z)^{-4}$ redshift dependence (Sadeh et al., 2015), or galaxy profiles, concentration, angular sizes or environmental properties, and so on. Another great advantage, with respect to SED fitters is also that, as it has been widely discussed in the literature (cf. Cavuoti et al. 2017) they provide more accurate results than

SED fitting methods within the limits imposed by the spectroscopic knowledge base.
This statement can be substantiated by several proofs, spread out in the whole body of this thesis, e.g.:

- the outcome of the second Euclid Data Challenge for the Organization Unit for photo-z (main content of Chapter 3) assessed METAPHOR as the method endowed with the highest photo-z measurements precision score for all galaxies within the limits imposed by the knowledge base on which the network training has been performed;
- in Cavuoti et al. (2017), on which the content of Chapter 4 is based, we will show the better performance of either punctual photo-z estimates and relative PDFs of METAPHOR with respect to the SED fitter Le-Phare (Arnouts et al. 1999; Ilbert et al. 2006) on SDSS DR9 data;
- in de Jong et al. (2017), for which we provided a catalog for KiDS (cf. Sec. 1.1) DR3 with more than 8 million objects in which two Machine Learning methods (METAPHOR[3], and ANNz2, cf. respectively Chapter 2 and Secs. 1.3.1 and 5.4.2) and the SED fitter BPZ were tested. In that work one of the conclusion of the KiDS collaboration was that "*...which set of photo-z's is preferred will depend on the scientific use case. For relatively bright and nearby galaxies* $(r < 20.5; z < 0.5)$ *the MLPQNA catalog provides the most reliable redshifts. Moving to fainter sources the BPZ and ANNz2 results are strongly preferred, but caution has to be observed regarding biases that can be dependent on magnitude or redshift.*
  This is due to the fact that the SED fitters are able to extrapolate reliable redshifts without no limits in redshifts, while ML methods depend strictly on such limits imposed by the knowledge base. Moreover the ML method ANNz2 trained on a different dataset, deeper to that used by METAPHOR, and that is the reason why ANNz2 is able to perform better also in the faint region...". The interested reader can find all these information in Sec.4 of de Jong et al. (2017) of which I was a co-author.

Many machine learning algorithms have been used for the determination of photo-z: neural networks (Collister and Lahav 2004; Sadeh et al. 2015), boosted decision trees (Gerdes, 2010), random forests (Carrasco and Brunner, 013a), self organized maps (Carrasco Kind and Brunner, 014a), quasi Newton algorithms (Cavuoti et al., 2012), etc, just to quote some of them. For a complete review about the photo-z techniques, of both types (SED fitters and MLMs), a detailed explanation of their merits and shortcomings, the reader is referred to Hildebrandt et al. (2010), Abdalla et al. (2011) and Sánchez et al. (2014).

[3] in (de Jong et al., 2017) it is referred as MLPQNA

1.2 The photo-z probability density function

In order to understand what a probability density function is, we have to grasp the fact that it does not make much sense to ask what is the probability of taking one single exact value for a continuous random variable X (since it can be demonstrated that such probability is always equal to 0), and that the right question is instead to determine the probability that such variable is close to a given value, say x. The answer to this question is provided by the probability density function (usually denoted with $\rho(x)$). In general, in order to translate the density into a probability, given an interval I_X around the value x, in which the density $\rho(x)$ is continuous, we have to consider that the probability will depend both on the density and on the amplitude of the interval:

$$P_r(X \in I_X) \sim \rho(x) * Length\ of\ I_X \tag{1.9}$$

The approximation in Eq. 1.9 (note that there is not a symbol of equality, just because $\rho(x)$ can vary over the interval I_X) improves as the interval I_X shrinks around the value x, and $\rho(x)$ becomes increasingly closer to a constant within the small interval, converging to a probability zero when the interval reduces to the single point x. The information about X, is therefore contained in the rate of decrease of probability when the interval shrinks.
In general, to determine the probability of a given random variable X in any subset A of the real numbers, we simply integrate $\rho(x)$ over the set A. In other words the probability that X is in the interval A is:

$$P_r(x \in A) = \int_A \rho(x)\, dx \tag{1.10}$$

The function $\rho(x)$, in order to be a probability density function must satisfy two conditions. It must be non-negative, so the that integral in Eq. 1.10 is always non-negative, and it must integrate to 1:

$$\rho(x) \geq 0\ for\ all\ x \quad \int \rho(x)\, dx = 1 \tag{1.11}$$

where the integral in Eq. 1.11 in now considered on the entire R.
In the photo-z context, the quantity that we actually estimate, it is more properly addressable as a Distribution Function, normalizable to a Probability Distribution Function (PDF), i.e just the quantity defined in Eq. 1.10, in the case of a continuous random variable X and as a probability mass function in the case of a discrete random variable (that is the probability

that the variable assumes a given single value, provided that condition Eq. 1.11 are still valid in the discrete approximation).
It has to be remarked that, from a rigorous statistical point of view, a PDF is an intrinsic property of a certain phenomenon, regardless the measurement methods that allow to quantify the phenomenon itself. On the contrary, in the context of the photo-z estimation, a PDF is strictly dependent both on the measurements methods (and chosen internal parameters of the methods themselves) and on the physical assumptions done. In this sense, the definition of a PDF in the context of photo-z estimation, needs to be taken with some caution.
Finally, a PDF should provide a robust estimate of the reliability of an individual redshift. In absence of systematics, factors affecting such reliability are: photometric errors, internal errors of the methods, statistical biases. In fact, by imagining that one could reconstruct a perfect photometric redshift, the PDF would be given by a single number. Actually the redshift inference has intrinsic uncertainties due to the fact that the available observables cannot be perfectly mapped to the true redshift. The PDF can, therefore, be thought as a way to parametrize the uncertainty on the photo-z solution, by relaxing somehow the information contained in a single error estimate. In other words, the parametrization of a single error, through a probability, allows to span an entire redshift range with the chosen *bin* accuracy, thus leading to an augment of the information rate according to a precision degree useful for a given scientific topic (cf. Sec. 2.4).

1.3 Machine Learning supervised PDF methods

In the framework of Machine Learning, a series of methods have been developed over the past years, in order to determine a PDF, not only for every single source within a catalog, but also to estimate the cumulative PDF for a sample of galaxies, this latter being usually obtained by stacking the single source PDFs.
The cumulative or sample PDF describes the probability that a randomly sampled galaxy in the sample has a certain redshift and it is very important for many cosmological topics.
The PDF estimation, can be roughly divided in classification and regression methods for what concerns supervised approaches (Secs. 1.3.1, 1.3.2, 1.3.3) and unsupervised approaches (Sec. 1.4).

1.3.1 Supervised classifiers

Generally, in the case of a MLM used as classifier, the idea is to find the mapping function between the input parameters and an associated likelihood function spanning the entire

redshift region, properly binned in classes (regions or bins). Such likelihood is expected to peak in the region where the true redshift actually is, and to be flat in the regions where the uncertainty is high. The purpose of a classification MLM is to differentiate between so-called signal, belonging to a given bin, and background objects, not belonging to the bin (Sadeh et al., 2015).

Several authors apply this classification approach in order to determine both the PDFs of single sources within a given catalog and the "stacked" PDF for the whole sample of galaxies available.

The approach is more or less always the same regardless the specific MLM used. It foresees first of all, the binning of the spectroscopic range covered by the training data: each bin has an amplitude equal to the degree of resolution one wants to achieve and each bin is populated by a different number of data if the width of the bins is kept fixed (due to the sampling of the training data) or we can decide to keep an equal number of sources per bin allowing the bin width to change from one bin to the other.

In Bonnet (2013) (hereafter B13), a Multi Layer Perceptron (MLP) was used as a classifier. A MLP is a Neural Network (NN) consisting in several ordered layers of perceptrons, these latter being algorithms able to map the input vector of features (i.e properties of the objects such fluxes, magnitudes, or derived colors, or other photometric information) $\overrightarrow{x}$ to a scalar. MLP is used in B13 by taking as input features magnitudes and their errors, and as target N classes corresponding to as many bins as those which are obtained by binning the redshift range keeping fixed the bin width to a desired resolution:... "*Therefore, the classes consist of one number between* 0 *and* $n-1$*, and these NN n output values between* $[0,1]$*, one for each class which sum up to 1, can be interpreted as the probability that the galaxy resides in that class (or redshift bin in this case)...*"(B13).

In this way, a PDF for every single source is obtained. Actually in B13 they provided only the stacked PDF, interpreted as the $N(photo-z)$ distribution of the whole data set under analysis, by taking as photo-z for every source the mode of the PDF for each source. Moreover, the only source of error considered in B13 was the one introduced by the used method itself[4]. In B13, 50 different PDFs were obtained by training and validating as many randomly initialized networks: the mean of such PDFs was then used as best PDF photo-z and the variance between the sets as error on the photo-z estimation. The performance of the method was finally evaluated by comparing the distribution of the true redshift $N(zspec)$ with the calculated $N(photo-z)$ through the calculation of the standard deviation of the differences of such distributions for the 50 validation samples, in 8 tomographic bins.

[4]Anticipating some results which will become apparent by reaching the detailed description in Sec. 2.3 we want to stress that these errors are induced by the intrinsically non deterministic nature of MLM which relies on a random initialization of the weights.

In Carrasco and Brunner (013a) (hereafter CK&B13a), two classes of prediction trees were implemented, classification and regression ones (that will be treated in Sec. 1.3.3), both recovered under a code called Trees for Photo-Z (TPZ).
Prediction trees are, among the non-linear and supervised MLMs to calculate photo-z's, the simplest and the most accurate: they are built by asking a sequence of questions that recursively split the data, frequently into two branches (but a splitting in more than two branches is always possible), until a terminal leaf (node) is created which fulfills a stopping criterion (e.g. minimum leaf size). The objects in the terminal nodes are characterized as having the same properties. According to the type of model used to make predictions about the objects in the leaves, we can have, as anticipated, classification or regression prediction trees. For what the classification prediction trees are concerned, they are designed to classify a discrete category from the data. Each terminal node containing therefore data that belong to one or more classes.
The prediction can be either a point prediction based on the mode of the classes inside that leaf or distributional by assigning probabilities for each category based on their empirically estimated relative frequencies. Without entering into the formal details about the construction of a classification tree, we only say that the splitting of an original node, encompassing all the training data, proceeds recursively, along the dimension (feature) that maximizes the information about the classes, through the maximization of a function called *Information Gain* which in its turn depends on the *Impurity Degree*, that as the name says, indicates the degree of impurity of the information about a given class. The classification obtainable from one classification tree is poor and surely does not give information on one of the sources of error to be accounted for in the calculation of a PDF, i.e. that induced by the specific MLM used. For this reason an ensemble learning technique, called *random forest*, is applied in order to build a forest of classification trees whose results are, at the end, combined all together. The idea is to use the bootstrap technique, *a method providing a direct computational way to asses uncertainty* (Hastie et al., 2001) on a given dataset. More precisely, if on the one end we want, as we will show in Chapter 5, to exclude the contribution to the error induced by the method, we need to find a way to evaluate it for every probed method.
A sketch of the random forest is given in Fig. 1.1. The bootstrap is based on the re-sampling of the training data through random replacement of the data themselves, a certain number of times (say B). In this way it is possible to obtain B replicates of the training set that have in common only the number of training samples. Indeed, the bootstrap proceeds by choosing randomly the objects from the original dataset, therefore, given a set of input parameters x_{ij}, with $i = 1, ..., N$ samples, and $j = 1, ..., m$ attributes, for each attribute, the x_i have a probability $1/N$ to be replicated, and each of the values x_i can figure more than one time

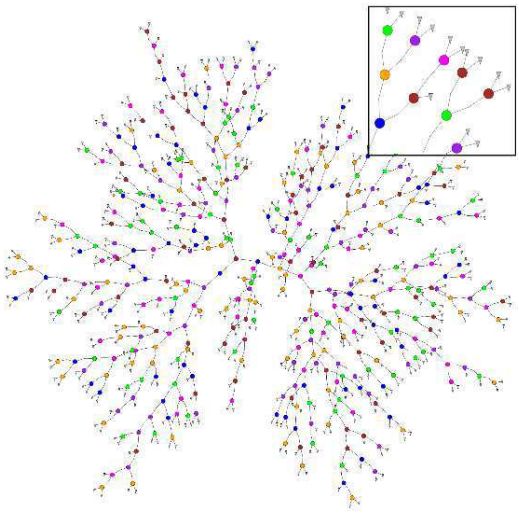

Fig. 1.1 Tree representation from Carrasco and Brunner (013a).

into the re-sampled bootstrapped datasets. Using these N_B replicates, an equal number of classification trees is generated. In order to introduce randomness in a controlled manner, in CK&B13a, N_R perturbed replicates of the pre-processed training set (removal of missing data, see also "proximity matrix" below) are generated by introducing the error on photometric attributes within the prediction trees. In this way it is possible to account for another source of error that the PDF has to reflect, i.e. the intrinsic uncertainties on photometry.
At this point, the random forest is created from the generation of N_B bootstrapped samples, for each perturbed sample N_R and the final PDF is constructed by leaving each galaxy evolving along each tree in the forest until the terminal leaves are reached. All the predictions are then combined together to give a final PDF for each source in the dataset.
The functionality of the TPZ code of CK&B13a, described until now, mostly in its classification mode, proceeds in the "classical" way of binning the spectroscopic redshift sample, with a fixed bin amplitude and a variable number of objects or viceversa. Within each bin a random forest is created, all these trees being able to classify an object as belonging or not to that bin: however the random forest is not trained on all the training set within each redshift bin, because by doing so, the performance of the method would be highly decreased due to large volume spanned on a little bin and to catastrophic errors since most of the data lie outside the considered bin. Therefore, Carrasco and Brunner (013a) follow the approach of Gerdes (2010) (we shall not enter into details): " *...Once all the forests are created for all the bins, the test data are run down on each tree of the forest, which assign the class to the test source inside or outside the bin: all the classes from the forest are combined in order to assign the probability for that source of belonging to that bin that is simply the*

number of times the source was assigned to that class divided for the number of trees. By repeating this process for each bin and re-normalizing the result, the PDF of the source is finally generated..."(CK&B13a).
Random forests are endowed by some techniques able at identifying zones where the photo-z prediction is either poor or weakly constrained by the training data, i.e where the representativeness of the photometric features in the training data is poor. These techniques are applied before computing the photo-z PDFs, and are:

- the "out of bag" data (OOB): in the growing of a forest, a part (about 1/3) of the data for the construction of each tree is not used and is considered as test sample for the tree. In this way, the error on such test samples is minimized, without resorting to the creation of a validation sample. The OOB data are also used to establish the feature importance, providing a way for removing attributes that not contribute significantly;

- the proximity matrix: is a symmetric, positive definite matrix that gives the fractions of trees in the forest in which two elements i and j fall in the same terminal node; this matrix is obtained processing all the data, also the OOB ones. When two galaxies are found in the same terminal leaf, their proximity is incremented by one and, at the end, all proximities are normalized using the number of trees. This matrix can then be used to identify the number of outliers within a dataset but also to substitute missing values within it by averaging the attributes of the k nearest galaxies.

Random forest classifiers are therefore able to address and reflect in their PDFs possible biases in the training data (sampling, completeness, degree of representativeness).
Finally, in Sadeh et al. (2015)(hereafter S15), the Artificial Neural Network publicly available algorithm ANNz2 (evolution of the previous version ANNz of Collister and Lahav (2004)) is used in order to determine both photo-z and their PDFs : it consists of several MLMs but the authors focus on three of them, and precisely, an Artificial Neural Network (a multi layer perceptron), a boosted decision tree (BDT) and k-nearest neighbours (KNNs). ANNz2 used as classifier is very similar to what has already been described for B13 and CK&B13a. It will be used also in Chapter 5 in the context of a comparison of PDF performances for Kilo Degree Survey data.

1.3.2 Ordinal classification methods

Rau et al. (2015) (hereafter R15), introduced an ordinal class PDF (OCP) algorithm to find sample (or cumulative) PDFs, according to which the bin membership, i.e the output of every classification method in treated as an ordinal variable: in other words, it makes use

of the information that the redshift bins are ordered. The fact that the classes are ordinal improves the classification itself. With respect to non ordinal classification methods (like those of Bonnet (2013) and Carrasco and Brunner (013a) in Sec. 1.3.1), an OCP trains a classifier that estimates the probability $p(z \geq z_i)$ that a new object has redshift z above a certain threshold z_i representing the edge of the relative redshift bin. In this work, the individual PDFs are determined as a kernel density estimate weighted in redshift space using as many kernel functions as the number of the training objects.
This ordinal classification leads to a Cumulative Distribution Function (it is worth to note that it is different from a *stacked* PDF), defined as:

$$CDF = \int_{-\infty}^{z} p(x)\,dx \tag{1.12}$$

which should be monotonically increasing under the hypothesis of perfectly PDF reconstruction by ordinal classifiers.

1.3.3 Regression Methods

The TPZ code of CK&B13a, described in section 1.3.1, has also the functionality of a regressor, according to which a fit is applied to the objects falling in the terminal leaves, due to the fact that now the variable to be predicted is continuous and not categorical. The construction of the tree is the same already given in section 1.3.1 for a classification tree.
An important difference between classification and regression trees lies in the procedure followed in selecting the best dimensions used to split the dataset. In the present case, it is based on the minimization of the sum of the squared errors, which for a node T is:

$$S(T) = \sum_{m \in values(M)} \sum_{t \in m} (z_i - \widehat{z}_m)^2 \tag{1.13}$$

where m are the possible values (bins) of the dimension M; z_i are the values of the target (i.e. spec-z) in each branch, and $\widehat{z}_m$ is the predictor model used (which is usually an arithmetic variable, e.g. the mean $\widehat{z}_m = \frac{1}{n_m}\sum_{i \in m} z_i$ on each branch m, n_m being the members of the branch m).
This procedure is repeated until some threshold in S is reached, or the fixed minimum of samples in the terminal leaves is achieved. It is applied to the whole training data in order to construct the regression trees, and the same procedure of perturbation and bootstrapping, described in Sec. 1.3.1 for the classification trees is followed to construct a random forest that encompasses all the redshift range. Finally, in order to estimate the *photo-z* and its error, we take the mean or the median of the few spectroscopic redshifts falling in the terminal

leaves, by further combining them with all the other means and medians from the other trees. In order to obtain a PDF, instead, for each test set source all the values in the leaves have to be kept and properly combined and normalized.

1.4 Machine Learning Unsupervised PDF methods

As specified in the introduction, unsupervised MLMs do not use the desired outputs (the spectroscopic redshift) during the training process in order to infer information but only the photometric data. The desired output is used only after the training is over to validate and understand the results.
A typical unsupervised method are Self Organizing Maps (SOM). SOM are a neural networks able to project a high-dimensional parameter space (PS) into a low dimensional space (in the most part of cases two dimensions are sufficient). In other words, a SOM is capable of performing a non-linear projection, by keeping intact the topology of the original PS.
In particular, in Carrasco Kind and Brunner (014a) (CK&B14, hereafter), a SOM is used as an ensemble technique, following the previous work done by the same authors on random forest (CK&B13a), as described in sections 1.3.1 and 1.3.3.
In practice, the bootstrapping approach has been used in order to construct N_B random different training data maps, all collected in a so-called random atlas, from which it is inferable a PDF. This kind of PDF, as that obtained by the method TPZ (see Sec. 1.3.3) keeps in consideration, the errors on the photometric features, incorporating them through a method of perturbation of photometry, by generating N_R perturbed replicates of each map starting from the known measurement errors. At the end N_B x N_R SOMs are generated, producing N_M bootstrapped maps for each perturbed sample of sources.
There exist several SOM versions, although all have the same procedure to train a map: the

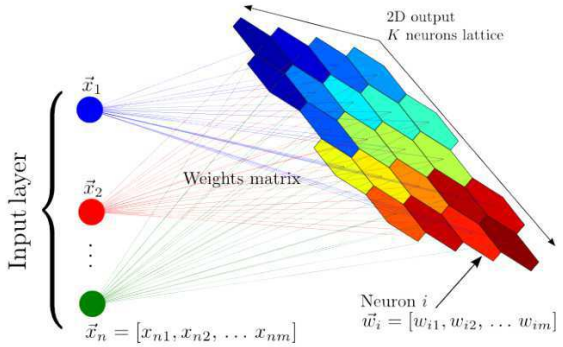

Fig. 1.2 SOM representation from Carrasco Kind and Brunner (014a).

differences among the methods arise in the way of updating the weight vector: see Bishop (2006).
Since the final aim of a SOM algorithm is the projection of a high dimensional PS into a two-dimensional space, in order to accomplish this task, a set of K weight vectors $\overrightarrow{w}_k \in R$ where $k = 1,2,...K$, corresponding to different neurons arranged in a given two-dimensional topology (rectangular, hexagonal, etc), are initially fixed at random by the algorithm, according to a certain distribution. By considering each galaxy within the dataset as a vector with m components that we can denote with $\overrightarrow{x} \in R^m$, in which each component represents an attribute or feature of the object (magnitude, color, and so on, except, we remember, the actual spectroscopic redshift), the SOM, at each training iteration, processes the n galaxies in the sample individually, by updating the weights at each step, i.e. after each galaxy has been processed.
This is the procedure that produces the self organization of the maps by keeping the topology of the parameter space. It is somehow as if the high-dimensional PS is bent onto a space with only two dimensions. During the processing of each training galaxy, the weight components of the best (in terms of reproduction of the galaxy features) neuron in the chosen topology, are updated, along with the components of the topologically closest neurons, in order to create regions of the map composed by neurons that are similar to each other, thus reproducing how similar galaxies tend to be co-located in the higher dimensional PS. This procedure of mapping is an approximation of the training probability distribution function.
Without entering in the details of the two techniques used and compared by CK&B14 in order to update the weights (on-line SOM, batch SOM), for which the interested reader is referred to the original paper, we shall spend only few words about the general SOM procedure.
The best neuron for each training galaxy is determined, by finding the smallest Euclidean distance between the feature vectors $\overrightarrow{x}$ and the neuron weights vectors $\overrightarrow{w}_k$ of the map. To update the weights at each iteration, the SOM uses a law depending on the learning-rate factor (which quantifies the correction for the regions of the map in function of time), and on a neighborhood function $H_{b,k}(t)$, that is a decreasing function of time and of the distance between two nodes (neurons): the best one, b, and another in the best node neighborhood, k. Such function defines the extent at which neurons near the best neuron, are updated to each time step. In this way neurons, closer to the best node are more strongly updated. The kernel function $H_{b,k}(t)$ is usually a Gaussian. Once the galaxies have been grouped according to their photometric features, and the map has been obtained, after the training, the desired attribute output is used at the end in order to visualize the map or to make an estimation of the photometric redshifts. For what it concerns the estimate of the PDF, the random atlas described above has to be constructed, (analogous to a random forest for supervised MLMs:

see sections 1.3.1 and 1.3.3).
Once all the weights for each map are recorded, the galaxies are processed again by using their associated weights in order to assign them to the corresponding (different) regions of the map, where each region represents galaxies that possess similar properties. In order to compute the photo-z for each galaxy, the test set has to be used (the one with objects having spectroscopic redshifts) and, again, which region in the map is the best representation the galaxy has to be fixed. This procedure is repeated for all the SOMs constructed as described above, and, at the end, all the SOMs photo-z results are combined together, after normalizing the result by the number of predictions, in order to have a PDF for each test galaxy. Among the authors using these techniques, we quote also the work of Speagle and Eisenstein (2015).

1.5 Ensemble learning techniques

In Carrasco and Brunner (014b), a combination of a modified version of the BPZ code of Benitez, and of the two methods TPZ (used in regression mode, see section 1.3.3) and of the SOMs (Sec. 1.4) is used in a successful attempt of determining an improvement of the PDF obtained in their previous works.
Such type of method combination is known as an Ensemble Learning (EL) technique. The idea of EL is to build a prediction model by combining the strengths of a collection of simpler base models. Ensemble learning can be broken down into two tasks: developing a population of base learners from the training data, and then combining them to form the composite predictor (Hastie et al., 2001). In the course of this PDF methods review, we already met, and properly named as ensemble techniques the random forest and the random atlas, since they actually can be addressable as such.
However, to combine different methods with different systematics through EL is a more difficult task that requires a special care. In any case, we will not enter into the details of all methods developed by Carrasco and Brunner (014b) (hereafter, CK&BEL) in order to combine the quoted methods, but we give only the general idea which is at the basis the EL method. In order to fix the framework, we can call the base learners (the models) **M**, and we know that the M_k models provide different photo-z PDFs or posterior probabilities. An EL photo-z PDF can be written as $P(z|\mathbf{x},\mathbf{D},\mathrm{M\_k})$ where **x** are the photometric attributes used to make predictions, **D** is the training set. These photo-z PDFs have to fulfill, for each model M, the following relation:

$$\int_{z_1}^{z_2} P(z|\mathbf{x},\mathbf{D},M_k) = 1 \tag{1.14}$$

where z_1 and z_2 are the extremes of the entire redshift range spanned by all the galaxies

$$P(z|\mathbf{x},\mathbf{M}) = \sum_k \omega_k P(z|\mathbf{x},M_k) \tag{1.15}$$

In CK&BEL, it is shown how the reliability of the combined PDFs through EL, outperforms that obtainable from each one of the combined method used individually, through the comparison of the cumulative $N(photo-z)$ PDF with the true one $N(spec-z)$ for the galaxy sample.

The ANNz2 code used by Sadeh et al. (2015), already quoted in Sec. 1.3.1, can be also used as "randomized regressor", i.e. another example of EL combination of the three different MLMs the authors use in their work (a NN, a BDT, and KNN, see Sec. 1.3.1). This allows to reflect the uncertainty of a given photo-z estimator as well as that on the specific combined training process itself.

The technique of the ensemble consists in organizing in a proper way the results obtained from several different methods: these latter differ in several ways, for example, in the case of a NN algorithm, by varying the number and types of neurons or by changing the arrangement of neurons in different layouts hidden layers. In the case of Boosted Decision Tree (BDT) the number of trees can be changed as well as the type of boosting/bagging algorithm and so on, or other more complex scenarios for which we refer the interested reader to the quoted paper. The photo-z distribution for each available galaxy is obtained, after the randomized MLM is initialized and trained on the entire training set. Then a selection on the ensemble of the answers is applied, by finding the best methods among those combined by the EL method. This is done according to certain criteria of minimization (based on statistical quantities like the bias or scatter averages): the MLMs with high values of average bias and scatter are rejected, and the remaining estimators are sorted on the basis of the quality of their bias and scatter averages.

Usually, the ensemble techniques proceed with the generation of several random weighting functions, according to which MLMs leading to the worse values of the statistical estimators of bias and scatter are ranked with lower weights and viceversa. In order to obtain the PDF of a galaxy, for each combination of weights, the weighted photo-z distribution is folded with the respective intrinsic scatter (that of the method). Once all the PDFs are determined, a fitting procedure is applied in order to extrapolate the best weighting scheme and then a cumulative PDF, defined as the integral between z_0 and z_{ref} of p(z_i) dz_i , this latter being the value of the PDF for a given redshift z_i, is derived.

The cumulative distribution is used for further constraining the reliability of the weighting scheme, ranking them by their compatibility with the true redshift distribution (that of spec-z).

1.6 Validation of the reliability of the PDF

Any method designed for the construction of PDFs should be tested to assess the reliability of the generated probability distribution function. A possible approach, used by almost all the authors quoted in this chapter, is the computation of the stacked PDF of the entire sample of galaxies at disposal, in order to compare this to the known distribution of the actual redshift (the spectroscopic one). This comparison assesses in a proper statistical way the differences. Sometimes a Kolmogov-Smirnov test is applied to the two distribution.
Another test could be the Kullback-Leibler divergence (used in the work on the ordinal classification by Rau et al. 2015), and finally, the credibility analysis, used also in this work, and described in Sec. 2.5.3, along with the Probability Integral Transform analysis.

Chapter 2

METAPHOR pipeline for photo-z and PDFs

(in part extracted from Amaro et al., 2017, MNRAS, submitted, and from S. Cavuoti, V. Amaro, M. Brescia, C. Vellucci, C. Tortora and G. Longo, "METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts", MNRAS, 2017, 465, 1959–1973)

2.1 The METAPHOR structure

As already mentioned in the previous Chapter, among empirical methods, those based on Machine Learning (ML) algorithms are the most frequently used. They infer (not analytically) the hidden relation between the input, mainly multi-band photometry (i.e. fluxes, magnitudes and/or derived colors) and the desired output (the spectroscopic redshift, hereafter spec-z). In the supervised ML techniques, explained in Sec. 1.3, the learning process is regulated by the spectroscopic information (i.e. redshift) available for a subsample of the objects, whereas in the unsupervised approach, the spectroscopic information is not used in the training phase, but only during the validation phase.
In the case of a supervised approach the construction of a proper Knowledge Base (KB) is a mandatory and crucial operation. The KB represents the dataset with which the empirical method can be trained and tested and it consists of a set of objects for which the true measurements (in this case, the spec-z) are available for a congruous and well-sampled number of objects.
METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts) is a python pipeline including functionalities which allow to obtain a PDF from any photo-z

prediction experiment done with interpolative methods.
Basically, it is a wrapper of the particular interpolative method chosen by the user to calculate the photometric redshifts (hereafter, photo-z).
The complete processing flow of the METAPHOR pipeline is laid in Fig. 2.1 and is based on the following functional macro phases:

- *Data preprocessing*. It includes data preparation, photometric evaluation and error estimation performed on the multiband catalog used as Knowledge Base (KB). This phase includes also the photometric perturbation of the KB which lays at the very base of METAPHOR method as we shall see in detail in Sec. 2.2.1;
- *Photo-z prediction*. It includes the training/test phase to be performed through the selected empirical method;
- *PDF estimation*. This phase is related to the method designed and implemented to furnish a PDF for the produced photo-z and to evaluate the statistical performance.

In the next paragraphs the details about these macro-phases and about some important micro-phases embedded within the method, will be given.

2.2 The Pre-processing phase

In the context of Machine Learning methods, the pre-processing phases are in geeneral always the same, independently on the particular spectroscopic/photometric catalogs at disposal.
The pre-processing aims at:

- Creating a suitable Knowledge Base in order to train our interpolative method (which will be explained in the next paragraph);
- Identifying and applying a proper law of photometric perturbation in order to allow the calculation of individual source PDFs, as we shall see in Sec. 2.4.

In the creation of the KB, we can recognize, the following micro-phases:

- *Application of the prescriptions to the multi-band photometric catalogs*: if there are flags which indicate the low quality of some photometric data in some bands, we apply them in order to train our algorithm on the best KB;

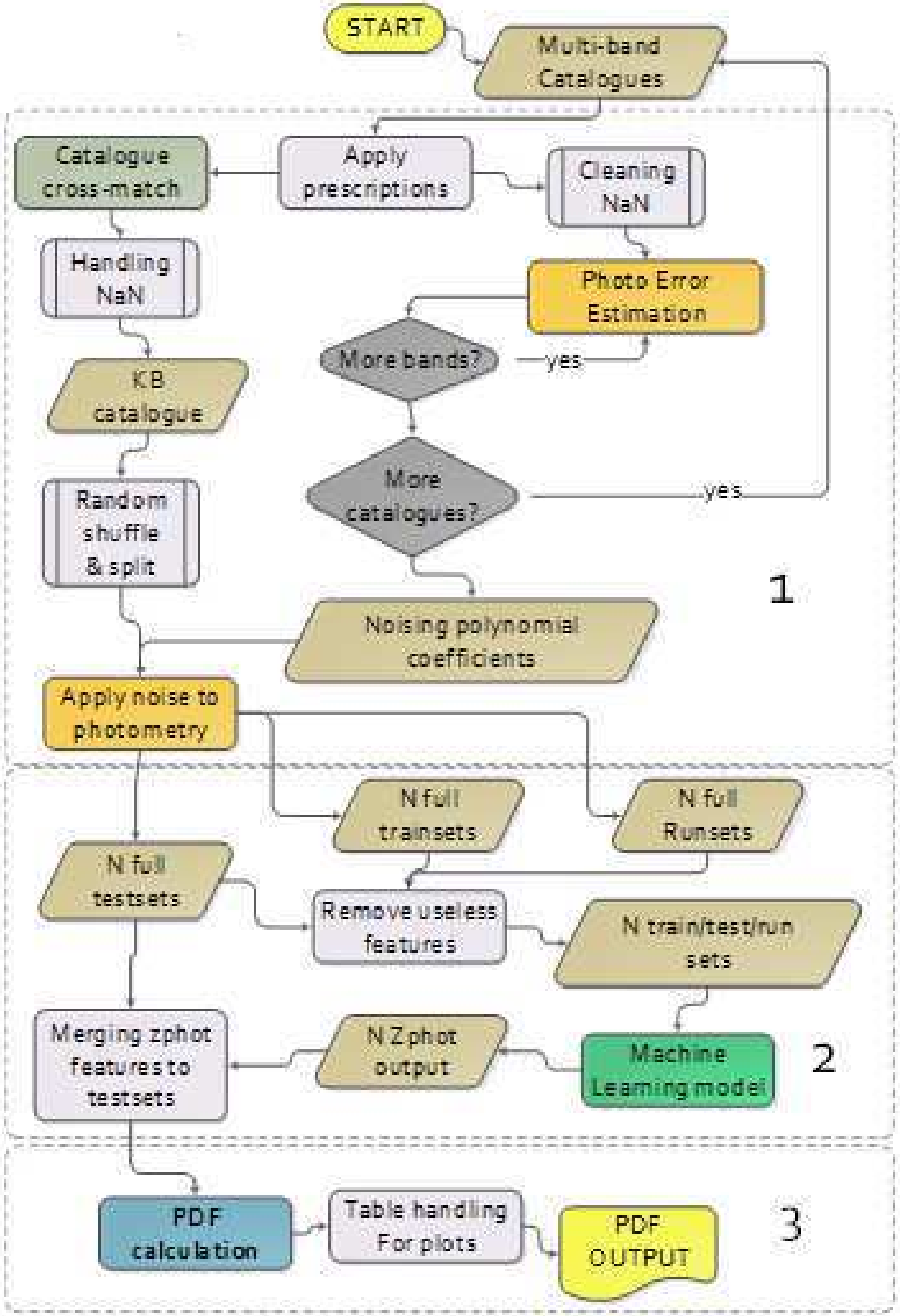

Fig. 2.1 Metaphor workflow.

- *Missing entries handling*: as known, astronomical multi-band catalogs are often affected by the presence of missing data. These may arise either from missing observations (in some bands) or bad reductions or from non-detections (objects too faint to be detected in a given band and translated into the presence of a NaN (Not-a-Number) symbol in place of the entry for a given object parameter). Furthermore, such symbol convention adopted to mark up missing values may be different among catalogs (especially when coming from different survey projects and/or data centers). The problem is that missing information has a negative impact on the photo-z prediction made with supervised machine learning methods, since the hidden correlation among band fluxes, found by training on these data, may be simply lost or strongly confused. Therefore particular care must be put in removing or at least minimizing the contamination induced by the presence of NaN, trying to reach the best trade-off between a sufficient amount of data, and, a minimal presence of missing entries in the KB;

- *Cross-Match (CM) between catalogs*: the fundamental rule for machine learning methods is to gather as much as possible information to infer the desired prediction and to maximize the estimation quality. Therefore both multiple bands and a sufficient amount of known samples (i.e. of spec-z) are required to increase the prediction performance on photo-z estimation. There are two modalities inside the Data Pre-Processing phase which are devoted, respectively, to cross-match spectroscopic catalogs (to increase the amount of available spec-z samples) and to cross-match photometric catalogs.
  The main difference between the two types of cross-matches is that, in the case of spectroscopic catalogs, an important rule of thumb is always to prefer the highest quality spec-z in the case of occurrence of multiple overlapping choices (hierarchical cross-match for spectroscopic catalogs). The CM phase leads to the creation of the *master* spectroscopic and photometric catalogs to be used along the pipeline process. Given two generic catalogs C_1 and C_2, both containing a list of objects with two columns representing, respectively RA and DEC coordinates, usually given in decimal degrees. The mathematical expression of the cross-matching is based on a distance calculated by the formula:

$$Distance(C_1, C_2) = \sqrt[2]{[RA_{C_1} - RA_{C_2}]\cos\left(\frac{dec_{C_1} + dec_{C_2}}{2}\right)^2 + (dec_{C_1} - dec_{C_2})^2} \tag{2.1}$$

  The formula 2.1 calculates a value of distance between two objects extracted from two generic sky catalogs. This value is generally expressed in decimal degrees. In order

to verify a match occurrence, the user must impose an initial max distance threshold, usually expressed in arcsec. Therefore, by taking care of any conversion between degrees and arcseconds, for each couple of objects, the match exists if:

$$Distance(C_1, C_2) \leq maxThreshold \tag{2.2}$$

In the case of two spectroscopic catalogs, as anticipated, the coordinates cross-match is a hierarchical dendrogram-like cross-match, in the sense that, given two spectroscopic catalogs, the merging of the two datasets will be maintained in order to grow the amount of objects, while for the common occurrences between the datasets, the highest quality spectroscopy catalog spec-z will be kept.

- *Photometry perturbation law application*: given the relevance of this topic to the work described in this thesis an entire subsection will be dedicated to it (Sec. 2.2.1);
- *Random Shuffle and Splitting*: the nature of supervised machine learning models imposes the splitting of a given KB into training and test datasets, to be used respectively, to train the model and to validate the learning performance. The random shuffle and split ensures the representativeness of the training set with respect to the parameter space covered by the test set as well as the homogeneity of the two data sets: a condition which is crucial to minimize systematics in the calibration of photometric redshifts.
- *Colors production*: this part of the pipeline performs the calculation of colors (which are given by the differences between magnitudes in different bands). This operation has to be done mandatorily after the application of the perturbation law, in order to avoid the propagation of the error already introduced by the photometric noise prescription.

2.2.1 The perturbation law

As anticipated, in the context of ML techniques, the determination of individual PDFs is a challenging task. This is because we would like to determine a PDF starting from several photo-z estimates, actually embedding the information on the photometric error uncertainties on those estimates. Therefore, we derived an analytical law to perturb the photometry by taking into account a realistic distribution of error on the magnitude derived by the photometric catalog itself.

The procedure to determine individual source PDFs consists of a single training of the MLPQNA model and by perturbing the photometry of the given test set to obtain an arbitrary

number N of test sets, characterized by a variable photometric noise contamination. The decision to perform a single train is mainly due to exclude the contribution of the method intrinsic error from the PDF calculation. We wish hower to stress that the internal error may easily be derived by naming N istances of training and then evaluating photo-z's on an unperturbed test set. Since repeated experiments have shown that this source of error is negligible with respect to the errors induced by the photometry, in what follows we shall not discuss the topic in any detail.

From a theoretical point of view, the characterization of photo-z predicted by empirical methods should be based on the real capability to evaluate the distribution of the photometric errors, to identify the correlation between photometric and spectroscopic error contributions and to disentangle the photometric uncertainty contribution from the one internal to the method itself.

Furthermore, the general approach is to perform the analysis through a binning of the parameter space in order to better focus the problem by keeping under control the photometry uncertainty at different distance regimes. The right choice of the bin size is however a common problem since it induces possible risks of information loss, varying between *aliasing* in the case of high density binning, to *masking* in the case of an under-sampling of the parameter space. We indeed tried to remain as much as possible bin size independent, at least in the first approximation of the design phase.

At the first order, the investigation could focus on the random perturbation of the photometry and the consequent estimation of its impact on the photo-z prediction. The main subject is therefore the final definition of a proper photometry perturbing function.

We start from the general idea based on finding a polynomial fitting of the mean error values which should be able to reproduce the intrinsic trend of the inner distribution, in order to derive a multiplicative factor for the Gaussian random seeds to be algebraically added to the fluxes values.

We parametrized the method, in order to ensure the flexibility and the possibility to adapt the method to arbitrary bands and photometric catalogs. In particular, the use of a different multiplicative constant for each band should be also considered, in order to customize the photometric error trend on the basis of the particular global photometry quality.

The impact of such approach was analyzed, reflecting the necessity to split the perturbation procedure in two steps: first, a preliminary statistical evaluation of the photometric error trend, in order to derive the perturbation coefficients of the polynomial noising function; and a second step to properly perturb the final cross-matched catalog.

Indeed, the pertubation law is:

$$\tilde{m}_{ij} = m_{ij} + \alpha_i F_{ij} u_{(\mu=0,\sigma=1)} \tag{2.3}$$

where *j* denotes the j-th object's magnitude and *i* the reference band; α_i is a multiplicative constant, chosen by the user (generally useful in order to take into account cases of heterogeneous photometry, i.e. derived from different surveys). The term $u_{(\mu=0,\sigma=1)}$ is a random value from a normal distribution; F_{ij} is the function used to perturb the magnitudes.
The function F_{ij}, can be chosen by the user among four different choises:

- *flat*: F_{ij} = constant weight, i.e. a floating number between 0 and 1, heuristically chosen;
- *individual*: for each object and band, F_{ij} coincides with the error on that feature, provided within the catalog;
- *polynomial*: F_{ij} is a polynomial fitting of the error means, say $p_i(m_{ij})$, after a proper binning of the bands in which to consider such means: this in order to reproduce the intrinsic trend of the errors;
- *bimodal*: composed by a constant function and a polynomial fitting of mean magnitude errors on the binned bands. The role of the constant function is as a threshold under which the polynomial term is too low to provide a significant noise contribution to the perturbation

In figure 2.2, we can see the types of perturbation that it is possible to build within METAPHOR.

2.2.2 The photometric error algorithm

Here we report the natural algorithm implemented to find the function F_{ij} in the Eq. 2.3:

1. Removal from the Knowledge Base of all the objects having errors in a specific photometric band higher than 1, thus creating a so-called *Reference*-catalog on which to proceed for the calculation of the polynomial fitting of the errors;
2. Binning of the photometric bands of the catalog just created, with a step equal to an arbitrarily chosen value (in the most part of our probed datasets this value has been set equal to 0.5) for all the available bands and fix the min and the max degree of the polynomial; note that the definition of a max degree for the polynomial is needed to avoid overfitting;

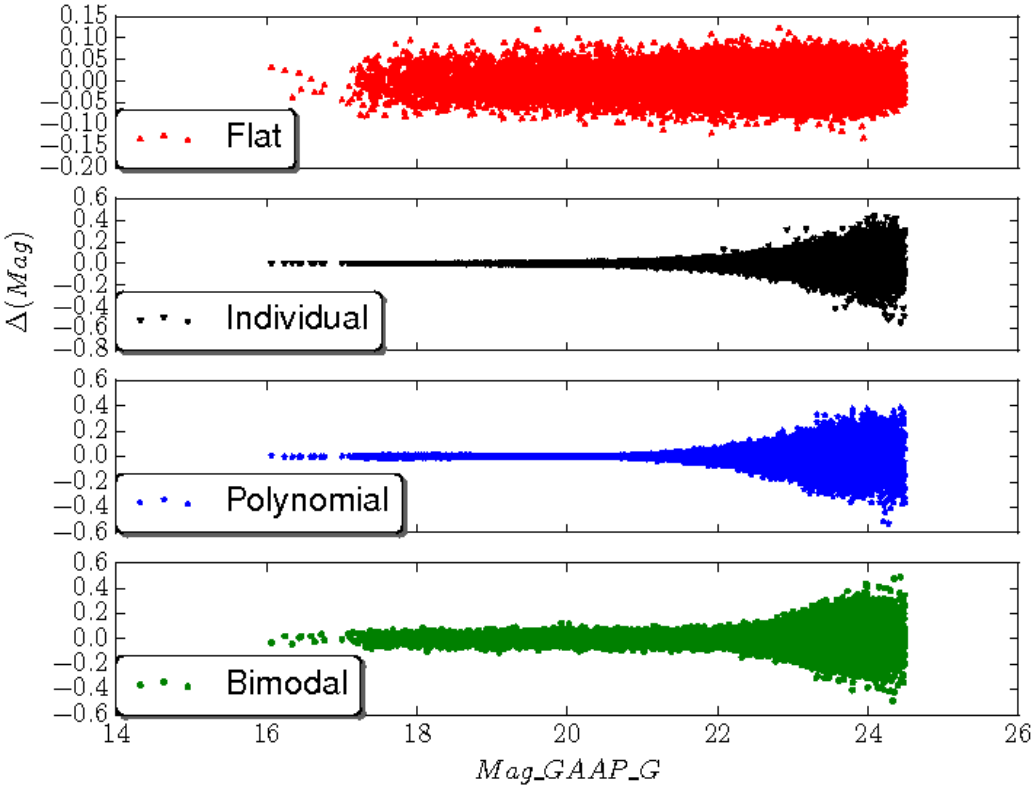

Fig. 2.2 Several types of function F_{ij}.

3. check the monotonicity and positivity of the polynomial fitting for each degree;
4. calculation in each bin, for each band, of the mean and the sigma (μ, σ) of the errors on magnitudes for the objects falling in that bin;
5. performance of a polynomial fitting of the error means $p_i(m_{ij})$;
6. compare the fit to σ to verify for each bin that the fitting error tolerance is within 1σ and register if this condition is fulfilled or not;
7. determine a fitness flag deriving from the truth table of the quoted features (monotonicity, positivity, condition on σ) starting from 0 for the worst condition to 7 for the best one;
8. among all the degrees of the polynomial, the best is the one corresponding to the highest fitness flag: in the case of equality of the flags among several polynomial degrees, the polynomial with the lower degree is chosen;
9. determination of a threshold in the polynomial value under which, the value of the function F_{ij} is constant and equal to such value. The value chosen for $p_i(m_{ij})$ can or cannot be the same for all the magnitudes used: under such threshold the value of F_{ij} is fixed to the threshold, above to the function given by the polynomial fitting $p_i(m_{ij})$.

Finally, as anticipated, by considering the whole perturbation law in 2.3, we can fix the values for α_i which can differ from one band to the other. In figure 2.3 it is possible to see the bimodal perturbation for the optical bands of the KiDS DR3 survey, which we shall describe in detail in Chapter 5.

2.3 The photometric calculation phase: MLPQNA

The current release of the METAPHOR pipeline includes the supervised machine learning model MLPQNA (Multi Layer Perceptron with Quasi Newton Algorithm), as method to produce the photo-z estimation and to perform classification of sky objects. However, as we shall see in Chapter 4, the user can choose the interpolative method he/she wants to use for the photo-z calculation. Indeed, as anticipated, the pipeline works like a wrapper of the embedded method.
Quasi-Newton Algorithms (QNAs) are variable metric methods for finding local maxima and minima of functions (Davidon, 1991). The model based on this learning rule and on the MLP network topology is then called MLPQNA.

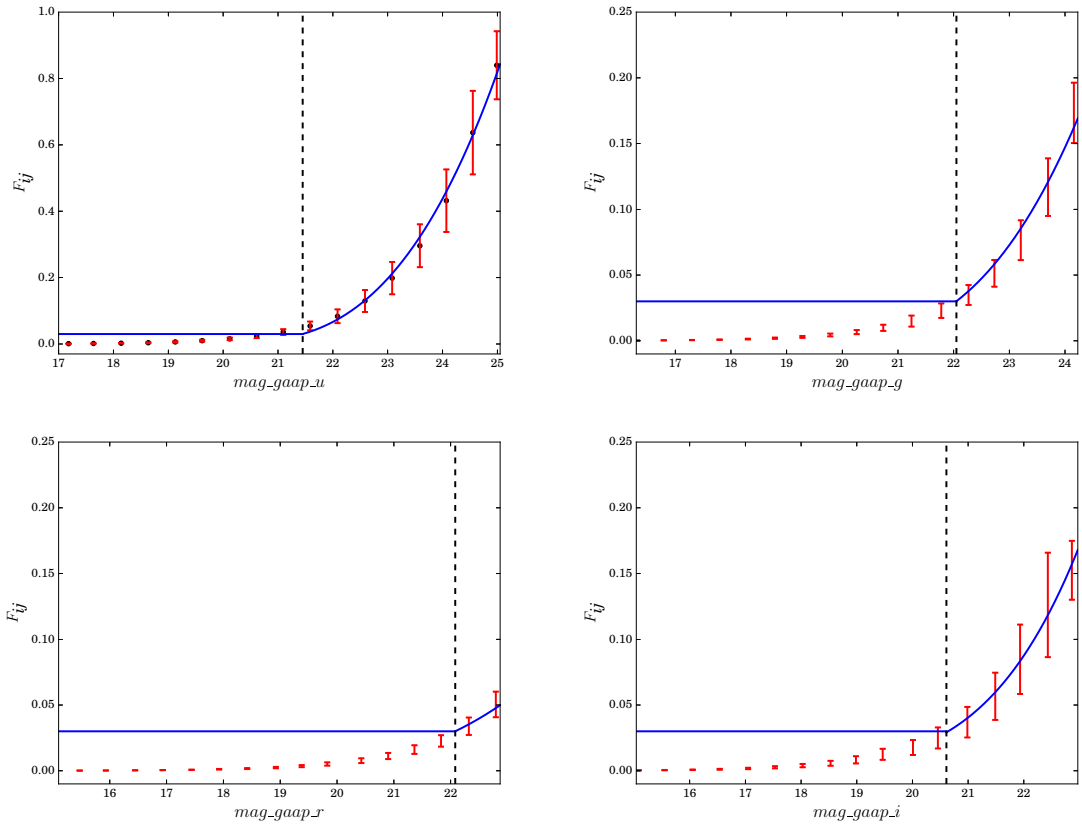

Fig. 2.3 Bimodal function F_{ij} in Eq. 2.3 for the GAaP magnitudes of the optical survey KiDS DR3 data, composed by a flat perturbation for magnitudes lower than a selected threshold (black dashed lines, chosen equal to 0.03) and a polynomial perturbation $p_i(m_{ij})$ for higher magnitude values. The switching thresholds between the two functions are, respectively, 21.45 in u band, 22.05 in g band, 22.08 in r and 20.61 in i band.

QNAs are based on Newton's method to find the stationary (i.e. the zero gradient) point of a function. Newton's method assumes that the function can be considered as quadratic in a narrow region around the optimum and uses the first and second derivatives (gradient and Hessian) to find the stationary point. In QNA, the Hessian matrix of second derivatives of the function to be minimized does not need to be computed and can be derived by analyzing successive gradient vectors. QNA is a generalization of the secant method to find the root of the first derivative for multidimensional problems.
In multiple dimensions, the secant equation is undetermined, and quasi-Newton methods differ in how they constrain the solution, typically by adding a simple low-rank update to the current estimate of the Hessian. The quasi-Newton method has been implemented by following the known L-BFGS (Limited memory–Broyden–Fletcher–Goldfarb–Shanno) algorithm (Byrd et al., 1994).
The QNA is an optimization of Newton-based learning rule, also because, as described below, the implementation is based on a statistical approximation of the Hessian by a cyclic gradient calculation, that is at the base of back propagation method. By using a local square approximation of the error function, we can obtain an expression for the minimum position. The gradient in every point w is in fact given by:

$$g = \Delta E = H \times (w - w_*) \tag{2.4}$$

Where $(w - w_*)$ corresponds to the minimum position. The gradient in every point w is in fact given by:

$$w_* = w - H^{-1} \times g \tag{2.5}$$

The factor $H^{-1} \times g$ is known as Newton direction and it is the base for a variety of optimization strategies, such as the QNA which instead of calculating the H matrix and then its inverse, uses a series of intermediate steps of lower computational cost to generate a sequence of matrices which are more accurate approximations of H^{-1} .
These matrices are computed using only information related to the first derivative of the error function. The Newton direction can be used in a line search (optimization problem) method when the Hessian matrix H is positive definite, because under such requirement it is a descent direction. When the Hessian is not positive definite, the Newton direction may not be defined, because its inverse matrix may not exist.
However, in addition, also when it is definite, it may not satisfy the descent trend. In particular, the main drawback of the Newton direction is the need for the exact Hessian matrix formulation.
As a matter of fact, this method was designed to optimize the functions of a number of argu-

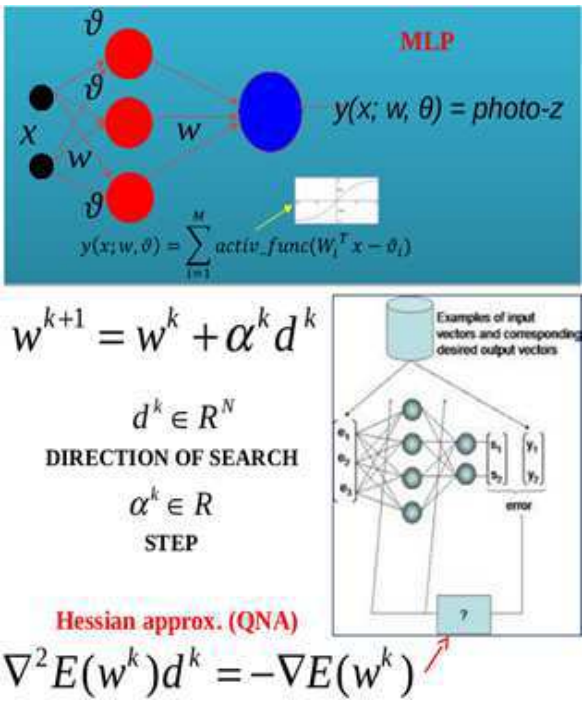

Fig. 2.4 General scheme of a MLP.

ments (hundreds to thousands), because in this case it is worth having an increased iteration number due to the lower approximation precision because the overheads become much lower. This is particularly useful in astrophysical DM problems, where usually the parameter space is dimensionally huge and is often affected by a low S/N. In terms of parameter setup, the model requires a proper heuristic choice of the following elements:

1. Input nodes: equivalent to the number N of features considered in the data set patterns;
2. 1_{st} layer hidden nodes: depending on the number of features considered in the data set patterns, this is the number of neurons of the first hidden layer (usually 2N+1 as rule of thumb);
3. 2_{nd} layer hidden nodes: depending on the number of features considered in the data set patterns, this is the number of neurons of the second hidden layer (usually N-1 as rule of thumb);
4. Activation functions: neuron function type, used to provide its output, by processing inputs;
5. Training mode: batch (weights update after each whole data set patterns calculation);
6. Training rule: Quasi Newton Algorithm;

7. Error loop threshold: one of the stopping criteria;
8. decay: error multiplicative regularization factor (see below);
9. Approximation steps: number of Hessian inverse matrix approximations to be done;
10. Number of iterations: one of the stopping criteria. Number of iterations for each approximation step.

In particular the regression error is based on Least Square error + Tikhonov regularization:

$$E = \sum_{i=1}^{N} \frac{(y_i - t_i)}{2} + \frac{\|W\|^2 \lambda}{2} \tag{2.6}$$

where y and t are respectively, output and target for each pattern, W is the weight matrix of MLP and λ represents the weight decay. Regularization of the weight decay is crucial issue. The implemented MLPQNA model uses Tikhonov regularization (AKA weight decay, λ). When the regularization factor is accurately chosen, then generalization error of the trained neural network can be improved, and training can be accelerated. Euristically, it is unknown what decay regularization value to choose (as usual), it is a good praxis to experiment values within the range of 0.001 (weak regularization) up to 100 (very strong regularization).
This can be done by starting with the minimum value and then increasing the decay value by a factor 3 to 10, while checking, using cross-validation, the network's generalization error. Optimization is performed from the initial point and until the successful stopping of the optimizer. Figure 2.5 shows a spectrum of neural networks trained with different values of decay, from zero value (no regularization) to infinitely large decay.
It can be seen that we control the tendency to overfit by continuously changing the decay. Zero decay corresponds to overfitted network. Infinitely large decay gives us underfitted network. Between these extreme values there is a range of networks which reproduce dataset with different degrees of precision and smoothness. Again, as it is shown, the perfect network is outside of this range. We can choose good neural network by heuristically tuning the weight decay coefficient.

2.4 The PDF calculation phase

Given a spectroscopic sample, randomly shuffled and split into training and test sets, a photometry perturbation algorithm (Sec. 2.2 and Sec. 2.2.1, respectively), and the selected photo-z estimation model, we proceed by perturbing the photometry of the given test set to

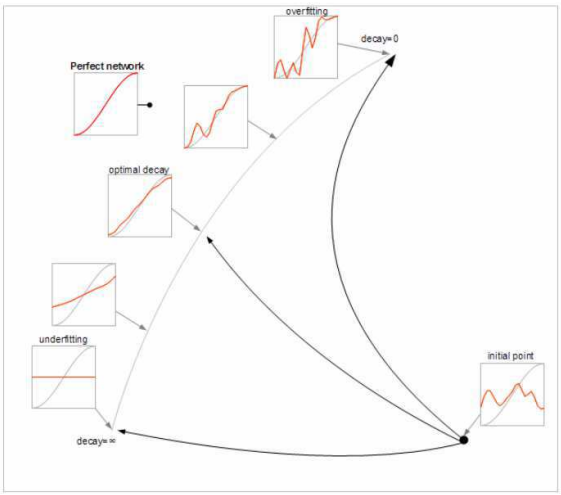

Fig. 2.5 QNA learning performance trend by varying the decay parameter.

obtain an arbitrary number N of test sets with a variable photometric noise contamination[1]. Although the multi-threading implementation of the interpolative method for photo-z estimation wrapped by the pipeline, foresees to apply perturbation also to the training set, we decided not to apply to it since, as we will explain several time in the course of this thesis, we would like to disentangle the error of the method from that arising from the uncertainties on photometry. This is obtained by taking into account the perturbation law (Sec. 2.2.1) only for the test set.

Indeed, we decided to proceed by training the model with the not-perturbed training set and to submit the $N+1$ test sets (i.e. N perturbed with noise sets and the original one) to the trained model, thus obtaining $N+1$ estimates of photo-z for each object within the available catalogs. The reason for not taking into account the performance of N perturbed train runs is just not to propagate the error depending from the method and due to the random inizialization of the weights in our MLPQNA model. In Chapter 5, where we compare the results obtained respectively with METAPHOR and another method in which the error information carried by the PDF is only the one induced by the method used to calculate the photo-z, we will show the contradictory results between the statistical estimators used to test such performances.

With these $N+1$ values, we perform a binning in photo-z, thus calculating for each one the probability that a given photo-z value belongs to each bin. The size of the binning is an arbitrary decision, to be made taking into account the specific requirements in terms of precision, related to the specific scientific topic to be addressed.

[1]We wish to stress that if one wants to estimate also the role played by the internal errors of the method, METAPHOR allows to do it by running N trainings on the same training set and then repeat the same sequence described here on each trained network. These errors however have been shown to be negligible.

We choose as default a binning step of 0.01 and adopted it for the experiments described in this thesis, otherwise specified[2].
The pseudo-algorithm, for a given photo-z binning step B, is the following:

- produce N photometric perturbations of the given test set, thus obtaining N additional test sets;
- perform 1 training (or N + 1 train) and N + 1 tests;
- derive and store the calculated N + 1 photo-z values;
- calculate the number of photo-z for each bin ($C_{B,i} \in [Z_i, Z_{i+B}Z[$);
- calculate, for each bin, the probability that the redshift be- longs to the bin: $P(Z_i \leq photo-z \leq Z_{i+B}) = \frac{C_{B,i}}{N+1}$;
- derive the resulting PDF as the set of all probabilities obtained at the previous step;
- calculate the statistics.

In figure 2.6, we summarize the general idea on which METAPHOR is based.
Concerning the photo-z production, the *best-estimate* photo-z values are not always corresponding to the given unperturbed catalog estimate of photo-z (hereafter photo-z_0), as calculated by MLPQNA. In particular it coincides with photo-z_0 if this measurement falls into the interval (or *bin*) representing the *peak* (maximum) of the PDF; otherwise, it corresponds to the one closest to photo-z_0 and falling in the *peak* of the PDF.

2.5 The statistical estimators calculated by METAPHOR

METAPHOR allows to calculate a whole series of statistical estimators which may be useful to assess the characteristics of the individual as well as of the stacked PDFs, besides the statistics on photo-z punctual estimates. We will summarize all these quantities in the next sections. These quantities have been used in all the investigations conducted on several datasets provided by several sky surveys on which METAPHOR has been tested and that will be described in the next chapters.

[2]This step is somehow very small but as it will be shown in Chapter 6, a small bin size can be easily rescaled, if needed, to larger values.

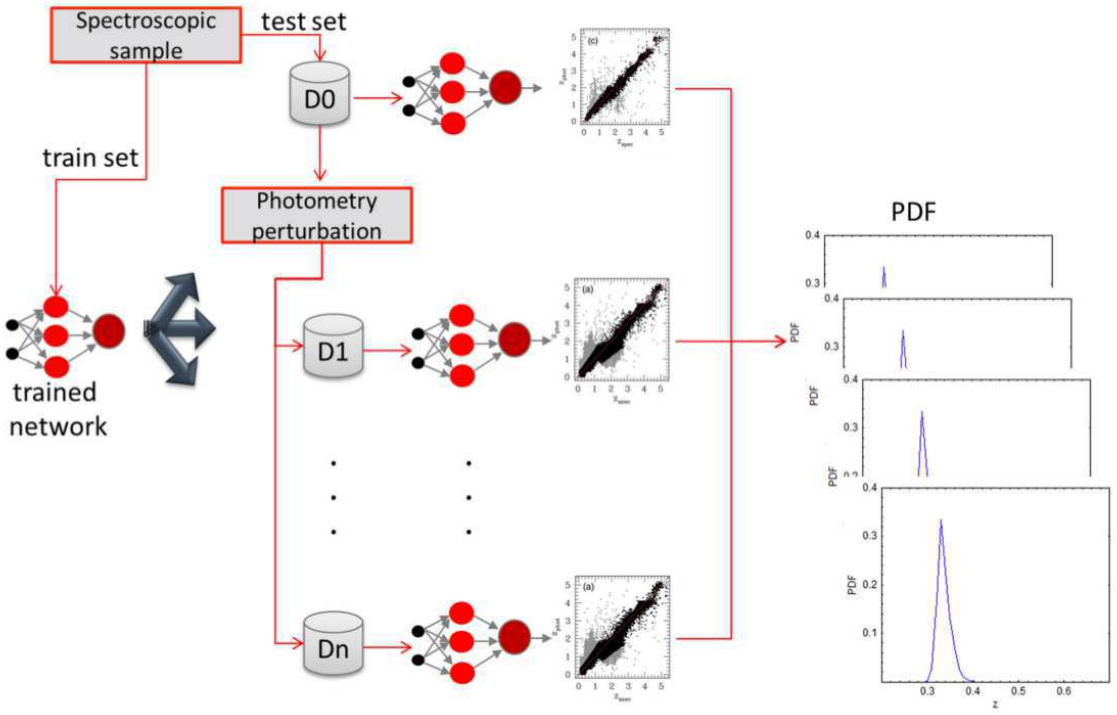

Fig. 2.6 Basic scheme behind the idea of the METAPHOR method. The photo-z estimation method shown here is the MLPQNA neural network model, although it could be replaced by an arbitrary interpolation technique.

2.5.1 The punctual photo-z statistical estimators

Usually, the results of the photo-z calculations were evaluated using a standard set of statistical estimators for the quantity:

$$\Delta z = (z_{phot} - z_{spec})/(1 + z_{spec}) \tag{2.7}$$

on the objects in the blind test set, as listed in the following:

(i) bias: defined as the mean value of the residuals Δz;

(ii) σ: the standard deviation of the residuals;

(iii) σ_{68} : the radius of the region that includes 68 per cent of the residuals close to 0;

(iv) *NMAD*: the normalized median absolute deviation of the residuals, defined as $NMAD(z)$ = 1.48 × Median ($|\Delta z|$);

(iv) fraction of outliers with $|\Delta z| > 0.15$;

(vi) skewness: measurement of the asymmetry of the probability distribution of a real-valued random variable around its mean;

(vii) kurtosis: gives information about the shape of a distribution tails.

The quoted indicators are calculated for all the $N+1$ estimates, and a final average, along with a standard deviation for all the estimates is provided. Usually, however, in order to evaluate the performances of the punctual photo-z estimation, we give such indicators calculated for the photo-z values obtained for the non perturbed test set.

2.5.2 The individual PDF statistical estimators

The quality of the individual PDFs is evaluated with respect to the whole spectroscopic data set, by defining five categories of occurrences:

- *zspecClass* = 0: the spec-z is within the *bin* containing the peak of the PDF;
- *zspecClass* = 1: the spec-z falls in one bin from the peak of the PDF;
- *zspecClass* = 2: the spec-z falls into the PDF, e.g. in a bin in which the PDF is different from zero;
- *zspecClass* = 3: the spec-z falls in the first bin outside the limits of the PDF;

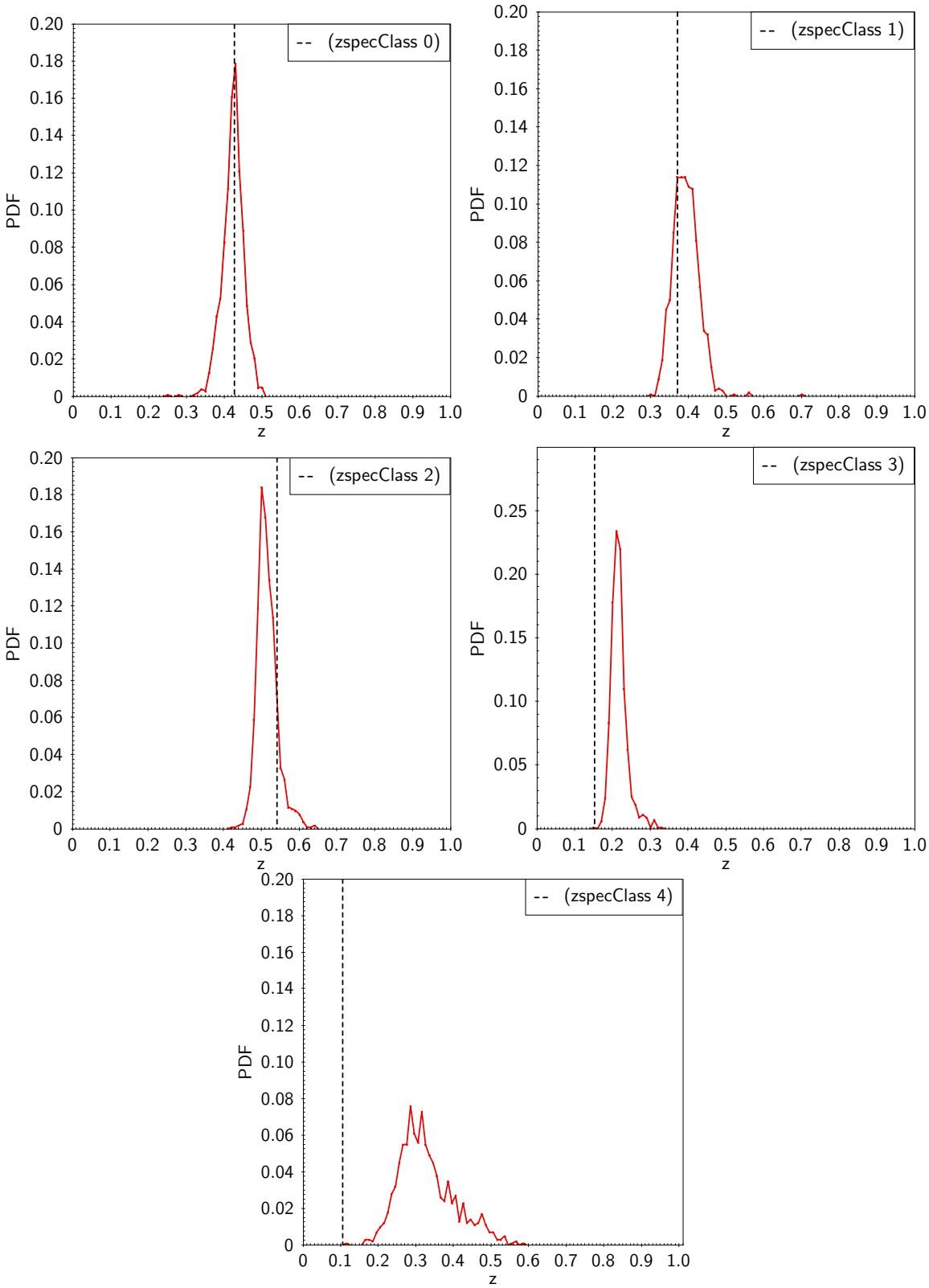

Fig. 2.7 Examples of several *zspecClass* PDFs.

- *zspecClass* = 4: the spec-z falls out of the first bin outside the limits of the PDF.

In figure 2.7, example plots for the aforementioned *zspecClass* individual PDF types. By definition, the *zspecClass* term depends on the chosen bin amplitude, which also determines the accuracy level of PDFs. The quality evaluation of the entire PDF can be hence measured in terms of fractions of occurrences of these five categories within the test data set.
Moreover, in the context of the Euclid Data Challenge 2, subject of the next Chapter 3, we calculate, for individual PDFs, some quantities which give information on the shape of the PDFs and therefore on their intrinsic quality. They are:

- *pdfWidth*: the width in redshift of the PDF;
- *pdfNBins*: the total number of bins (of chosen amplitude, that, as said, defines the accuracy of the PDF itself), in which the PDF has a value different from 0;
- *pdfPeakHeight*: the amplitude of the peak of the PDF, i.e. the value of the maximum probability of the PDF;
- *pdfNearPeakWidth*: the amplitude of the PDF near the peak, i.e the distance between the latest bin with PDF $\neq 0$ higher than the peak bin, and the latest bin with PDF $\neq 0$ lower than the peak bin.

2.5.3 The stacked PDF statistical estimators

In order to evaluate the cumulative performance of the PDF, we computed the following three estimators on the stacked residuals of the PDFs:

i) $f_{0.05}$: the percentage of residuals Δz within ± 0.05;

ii) $f_{0.15}$: the percentage of residuals Δz within ± 0.15;

iii) $\langle \Delta z \rangle$: the average of all the residuals Δz of the stacked PDFs.

2.6 Qualitative estimators for stacked PDFs

Finally, we adopted two more graphical diagnostics to analyze the *cumulative* performance of the PDFs, respectively, the credibility analysis presented in Wittman et al. (2016) and the Probability Integral Transform (hereafter PIT), described in Gneiting et al. (2007).
The credibility test should assess if PDFs have the correct *width* or, in other words, it is a test of the *overconfidence* of any method used to calculate the PDFs. In particular, the method is

considered overconfident if the produced PDFs result too narrow, i.e. too sharply peaked; underconfident otherwise. In order to measure the credibility, rather than the Confidence Intervals (hereafter CI), the Highest Probability Density Confidence Intervals (hereafter HPDCI) are used, since it is considered one of the best statistical ways to perform such measurement (Wittman et al., 2016).

The implementation of the credibility method is very straightforward, and is reached by computing the threshold credibility c_i for the i-th galaxy with

$$c_i = \sum_{z \in p_i \geq p_i(z_{s,i})} p_i(z) \tag{2.8}$$

where p_i is the normalized PDF for the i-th galaxy.
The credibility is then tested by calculating the cumulative distribution *F(c)*, which should be equal to c. *F(c)* resembles a q-q plot, (a typical quantile-quantile plot used for comparing two distributions), in which F is expected to match c, i.e it follows the bisector in the F and c ranges equal to [0,1]. Therefore, the *overconfidence* corresponds to *F(c)* falling below the bisector (implying that too few galaxies have spec-z with a given CI), otherwise the *underconfidence* occurs. In both cases this method indicates the inaccuracy of the error budget (Wittman et al., 2016). See figure 2.8 in which cases of *overconfidence* and *underconfidence* are shown. The figure is taken from Wittman et al. (2016).

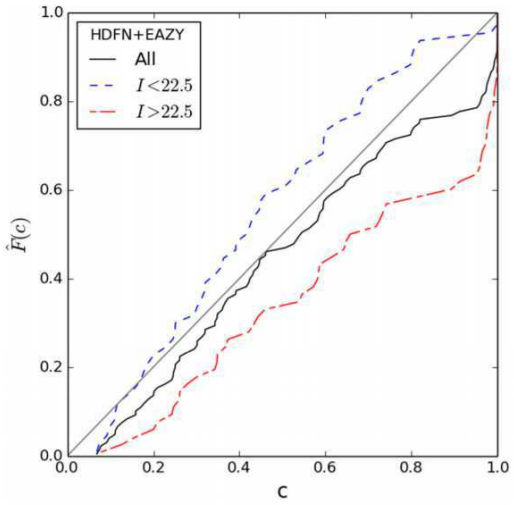

Fig. 2.8 Wittman credibility analysis examples of *overconfidence* (for the curves below the bisector of the plot $F(c)$ *vs* c, and of *underconfidence*, indicated by the curves above the same bisector. The figure is taken from the paper of Wittman et al. (2016).

The PIT takes the form of a PDF or of a predictive Cumulative Distribution Function (CDF) and measures the predictive capability of a forecast, which is generally probabilistic for continuous or mixed discrete-continuous random variables (Gneiting et al., 2007). We can define the PIT as

$$p_i = F_i(x_i) \tag{2.9}$$

Ideal forecasts produce continuous F_i and PIT with a uniform distribution on the interval (0,1). In other words, we can check for an ideal forecast by investigating the uniformity of the PIT: the closer the histogram to the uniform distribution, the better the calibration, i.e. the statistical consistency between the predictive distributions and the validating observations (Baran and Lerch, 2016). Nevertheless, it is possible to show that the uniformity of a PIT is a necessary but not sufficient condition for having an ideal forecast (Gneiting et al., 2007).
A strongly U-shaped PIT histogram (e.g. the bottom right panel of Fig. 5.6) indicates a highly *underdispersive* character of the predictive distribution (Baran and Lerch, 2016).

Chapter 3

METAPHOR and the Euclid Data Challenge 2

3.1 Introduction

Euclid (Euclid, 2011) is an ESA mission aimed at understanding the nature of dark matter and dark energy by means of weak lensing and baryon acoustic oscillations. The launch of the satellite is foreseen for 2020. This mission will observe galaxies and galaxy clusters up to $z \sim 2$, and will cover $15,000$ deg^2 for the wide extra-galactic survey, plus a deep survey covering 40 deg^2. The on board instruments will be two: an imager in the visible domain (VIS) and an imager-spectrometer (NISP) covering the near-infrared. The Euclid Consortium (EC), is formed by over 110 institutes spread in 15 countries and it provides the data to be processed by the Euclid Science Ground Segment (SGS) formed in its turn by the Science Operations Centre (SOC) operated by ESA and nine Science Data Centres (SDCs). In figure 3.1, a scheme of the interactions among the eleven Processing Functions of Euclid, and of the data flow. These functions are eleven and are:

- LE1: is in charge of telemetry processing;
- VIS, NIR, EXT: production of fully calibrated photometric exposures from Euclid and ground-based surveys;
- SIR: production of fully calibrated 1D spectra extracted from the NISP spectroscopy;
- MER: production of a source catalog containing consistent photometric and spectroscopic measurements;
- PHZ: production of photometric redshifts for all sources within the catalogs;

- SPE: production of spectroscopic redshifts for all sources with spectra;
- SHE: measurements of galaxy shapes;
- LE3: production of all high-level science;
- SIM: production of all the simulated data necessary to validate the data processing stages and to calibrate observational or method biases.

In this Chapter, we report an application of METAPHOR to the data released by the ESA EUCLID Consortium Euclid Data for the internal 2th OU-PHZ (Organization Unit for Photo-z) Challenge (hereafter, EDC2), with the aim to establish the best SED fitting and/or empirical methods which will be included in the official data flow processing pipelines for the mission. This contest lasted from September 2015 until the end of Jenuary 2016, and ended with the releases of the results on the participants performances, in the middle of May 2016.
For us, the final goal of the challenge was to obtain the probability density functions for the photometric redshifts, obtained by the application of the machine learning method MLPQNA (Brescia et al. 2012, Brescia et al. 2013).

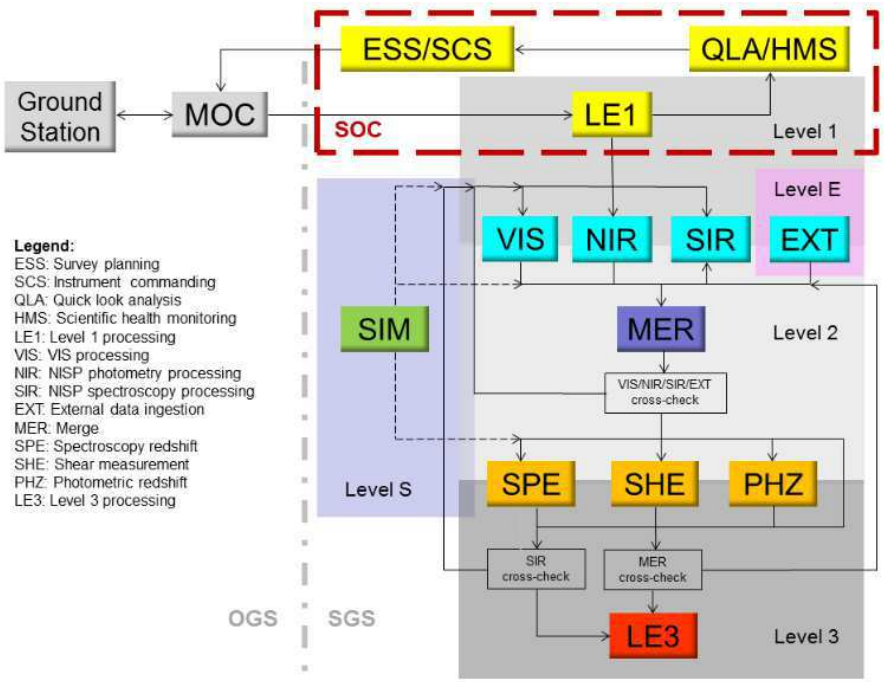

Fig. 3.1 Euclid SGS Processing Functions interactions and data flow.

3.2 The Challenge data

The calibration catalog (hereafter “calib”, with #190,508 objects) used for the experiments shown in this Chapter, was

euclid_cosmos_DC2_2fwhm_S2_v2_DESnoise_calib.fits

Such catalog was recommended by the Euclid OU-MER (Organization Unit-Merging), i.e. the organization which realizes the merging of all information produced by other Processing Functions (VIS, NIR, EXT, SIR) shown in figure 3.1 and listed in the previous paragraph. That is OU-MER is in charge of providing stacked images and source catalogs where all multi-wavelength data (photometric and spectroscopic) are aggregated in flux measurements, as well as the VIS-like Hubble Space Telescope (HST)-Advanced Camera for Surveys (ACS) image and of providing the algorithmic definition of the processing to be implemented by the Science Data Centers and validating the implementation.
OU-MER organization obtained photometry within an aperture of $2fwhm$[1], by using homogenized PSF images. Note that, catalogs calib and validation/run (hereafter, “verif”) are split in RA: calib has objects with RA > 150.125; verif lower. The calib catalog has spectra but no coordinates, opposite to verif catalog. This is in order to obtain a "truly blind" challenge.

3.2.1 Prescriptions applied to the calibration and run catalogs

The prescription applied to calib catalog were:

- Cleaning of all available magnitudes from NaN entries (mag_x, with $x = g,r,i,VIS,z,Y,J,H$, column numbers #22 – #29);
- Application of the reliability flag (reliable_$S15$, column number #54, the spec-z is reliable if the flag is =1) according to the scheme of Salvato et al. 2015 (S15, hereafter) with the additional removal of the objects with a quality flag $Q\_f\_S15 = 6$ (Q_f_S15, column number #52);
- Elimination of stars and conservation of the AGN : we defined a new column (column number #59) within the calibration catalog, by means of the columns #57 (“STAR” flag, if star, flag=1) and #58 (“ AGN” flag, if AGN, flag=1), using the expression, #59 defined as:

$$(\#57 = 1 \;\&\; \#58 = 0)?1 : 0 \tag{3.1}$$

[1] FWHM= Full Width Half Maximum

and we kept all the objects with #59 = 0. Actually, equation 3.1, by fixing the condition to be a star and not an AGN, with the value for the column number #59 = 1, allows, by requiring #59 = 0 that the probed objects were AGNs. However, a certain number of objects were classified as ambiguous (either stars or AGNs), but through the evaluation of the X-ray flag ("$flag\_X\_ray\_s15$", column number #55) we found that all these objects were actually X-ray emitters, and then AGNs. Moreover, the restriction of the spectroscopic redshift range, removing objects with spec-z=0 (see following point), ensured that such objects were reliably classified as AGNs;

- Restriction of the redshift (z_spec_S15, column number #51) range to the interval $]0, 4.5[$;
- Application of a prescription through the Sextractor Flags ("FLAGS_DETECT", column number #44). We remember the meaning of these flags:

  1. 1: an object that has neighbors bright and close enough to significantly bias the photometry, or bad pixels (more than 10% of the integrated area affected);
  2. 2 : the object was originally blended with another one;
  3. $3 : 3 = 2 + 1$
  4. 4 : at least one pixel of the object is saturated (or very close to);
  5. 8: the object is truncated (too close to an image boundary);
  6. 16 : object's aperture data are incomplete or corrupted;
  7. 32 : objectc's isophotal data are incomplete or corrupted;
  8. 64 : a memory overflow occurred during deblending;
  9. 128: a memory overflow occurred during extraction.

In order to avoid highly problematic photometry without losing too many objects, we agreed, on the basis of the SEXtractor flags meaning, and of the calculated number of objects per SEXtractor flag value, that the problematic photometry to be removed is the one flagged with numbers higher than or equal to 4, i.e. from flag = 4 up. Therefore a prescription was applied to the $FLAGS\_DETECT$ column, by removing all samples with flags higher than 3, using the condition for keeping samples with

$$\#44 < 4$$

All these actions can be summarized in the following string:

$$\begin{aligned} \#22 > 0 \ \&\ \#23 > 0 \ \&\ \#24 > 0 \ \&\ \#25 > 0 \ \&\ \#26 > 0 \ \&\ \#27 > 0 \ \& \\ \#28 > 0 \ \&\ \#29 > 0 \ \&\ \#44 < 4 \ \&\ \#54 = 1 \ \&\ \#52 \neq 6 \ \& \\ \#59 = 0 \ \&\ \#51 > 0 \ \&\ \#51 < 4.5 \end{aligned} \tag{3.2}$$

The application of prescription in 3.2 brought the number of objects of the calib catalog to #13,789. As regards the verif catalog (with #190,462 samples):

euclid_cosmos_DC2_2fwhm_S2_v2_DESnoise_valid.fits

the application of:

- cleaning of the available magnitudes;
- the SExtractor flag condition quoted above for the calibration catalog;

determined a dataset with a number of objects equal to #140,828. This led to the creation of catalogs that we shall call in the following with the prefix "Ref", and are useful to calculate the photometry perturbation law (See Eq. 2.3).
However, the final results required by challenge, had to be returned for the whole "verif" catalog (i.e for all the samples within) in a "fits" file containing the following quantities :

- #1: REDSHIFT: redshift point estimate (for tomographic bins definition), i.e. the one that we called *best-estimate* photo-z (See Sec. 2.4) in our algorithm to find the PDF;
- #2:USE: 1 if redshift is reliable (e.g. belongs to a color space with enough spectra), 0 otherwise;
- #3: STAR: 1 if star (0 otherwise) or "-99" if we do not perform classification experiments;
- #4: AGN: 1 if AGN (0 otherwise) or "-99" if we do not perform classification experiments;
- #5: END: PDF(z), min=0.0, max=6.0, dz = 0.02 with PDF columns that must correspond to the bin center.

3.2.2 Some other photometric prescriptions

At this point, in order to remove some problematic photometry (too faint and under-sampled objects and/or samples with error on magnitude $>$ or $\gg 1$), we decided to conduct two experiments:

- one creating for both the calib and verif catalogs two relative subsets imposing the condition for the cut of samples with magnitudes deeper than those with depths within 5 sigma, provided in the "readme" file, delivered by the Consortium together with the catalogs. The magnitude depths within 5σ are:

  1. g: 24.95 ± 0.01
  2. r: 24.60 ± 0.01
  3. i: 23.72 ± 0.01
  4. ACS: 24.82 ± 0.01
  5. z: 23.21 ± 0.01
  6. Y: 24.57 ± 0.02
  7. J: 24.35 ± 0.02
  8. H: 23.89 ± 0.02

  The catalogs obtained were saved with the names:

  - calib_depth_mag.csv (#11,545 objects, note that the cut on mag depths removed all calib samples with mag errors > 1)
  - verif_depth_mag.csv (#33,355 objects, brought to #33,348, removing further 7 objects with error on mag > 1)

- for both calib and verif catalogs two relative subsets imposing only a cut for the objects with mag err (columns from #30 to #37) >1 were created; these catalogs were saved with the names:

  - calib_error_cut.csv (#13,302 objects);
  - verif_error_cut.csv (#83,585 objects).

For the quoted experiments the condition in Eq. 3.2 in section 3.2, becomes, for the calib_depth_mag.csv and calib_error_cut.csv catalogs, respectively:

$$\begin{aligned} &\#22 > 0 \ \&\ \#23 > 0 \ \&\ \#24 > 0 \ \&\ \#25 > 0 \ \&\ \#26 > 0 \ \& \\ &\#27 > 0 \ \&\ \#28 > 0 \ \&\ \#29 > 0 \ \&\ \#44 < 4 \ \& \\ &\#54 = 1 \ \&\ \#52 \neq 6 \ \&\ \#59 = 0 \ \&\ \#51 > 0 \ \&\ \#51 < 4.5 \ \& \\ &\#22 < 24.95 \ \&\ \#23 < 24.60 \ \&\ \#24 < 23.72 \ \&\ \#25 < 24.82 \ \&\ \#26 < 23.21 \\ &\& \ \#27 < 24.57 \ \&\ \#28 < 24.35 \ \&\ \#29 < 23.89 \end{aligned} \tag{3.3}$$

$$\begin{aligned} &\#22 > 0 \ \&\ \#23 > 0 \ \&\ \#24 > 0 \ \&\ \#25 > 0 \ \&\ \#26 > 0 \ \& \\ &\#27 > 0 \ \&\ \#28 > 0 \ \&\ \#29 > 0 \ \&\ \#44 < 4 \ \&\ \#54 = 1 \ \& \\ &\#52 \neq 6 \ \&\ \#59 = 0 \ \&\ \#51 > 0 \ \&\ \#51 < 4.5 \ \&\ \#30 < 1 \ \& \\ &\#31 < 1 \ \&\ \#32 < 1 \ \&\ \#33 < 1 \ \&\ \#34 < 1 \ \&\ \#35 < 1 \\ &\& \ \#36 < 1 \ \&\ \#37 < 1 \end{aligned} \tag{3.4}$$

The training sets, for the two experiments, have been randomly shuffled and split in a train set (with 80% of the training set samples) and a test set (with 20% of the training set objects). The perturbation of photometry foresaw two phases, as described in Sec. 2.2.1 and 2.2.2:

- the first phase in which we calculated weighted third degree polynomial coefficients for each band, by dividing the magnitude ranges in bins of Δmag=0.5, and by finding the mean of errors for the samples falling within the bins. Lastly, the quoted fit is performed on these error means, according to what has been described in Sec. 2.2.2;
- The second phase, in which the magnitudes of the test set were perturbed 100 times with the law of perturbation in Eq. 2.3;
- In order to perform the first phase, a merging of the calib catalogs and of the verif catalog (with all the prescriptions quoted in this section and in previous one, applied), has been saved with the name:

  Refcatalog_error_cut.csv (#96,887 objects)

Such "Ref" catalog has been created by the merging of the catalogs calib/verif_error_cut.csv, in order to maximize information about sky objects: indeed the catalog calib_dept_mag.csv was contained in that calib_error_cut, and verif_error_cut contained many more objects than that verif_depth_mag.

3.3 Determination of the photo-z with MLPQNA

Two MLPQNA experiments have been conducted with one train set, 101 test sets, 101 run sets, where 100 test and run sets have been created using the perturbation law explained in section 2.2.1. The parameter space (PS) features used were the 8 available magnitudes:

g,r,i,VIS, z, Y, J, H

The parameters of the network used for both experiments (couple calib/verif_depth_mag and couple calib/verif_error_cut catalogs) are summarized in table 3.1 while the results of the experiments are in table 3.2.

Table 3.1 Network parameters for two experiments performed using the PS 8 available magnitudes as features.

Network parameter	
input Neurons	8
# Hidden layers	2
# neurons 1^{th} hidden layer	17
# neurons 2^{th} hidden layer	7
restarts	70
epochs	10,000
threshold	0.001
decay	0.01

Table 3.2 Statistics of the results of the experiments. All quantities are reported for Δz=|$photo - z_0$-spec-z|/(1+spec-z), where $photo - z_0$ is the estimated MLPQNA photo-z for the non perturbed test set.

Estimator	calib_depth_mag (8 magnitudes)	calib_error_cut (8 magnitudes)
bias	0.0099	0.017
σ	0.132	0.162
NMAD	0.047	0.059
σ_{68}	0.051	0.064
σ_{95}	0.183	0.239
outliers	6.74%	11.06%
train/test	9,278/2,267	10,617/2,685

Table 3.3 *zspecClass* occurrences for the experiments quoted in Tab. 3.2

zspecClass	calib_depth_mag	calib_error_cut
0	231(10%)	211(8%)
1	392(17%)	394(15%)
2	1,523(67%)	1,746(66%)
3	121(%)	334(12%)

3.4 Determination of the probability density functions

The PDF algorithm, contained in the last part of the METAPHOR pipeline, allows to calculate all the individual PDFs on the test set objects along with the calculation of the number of occurrences of the estimator *zspecClass* explained in Sec. 2.5. Moreover, the calculation of the quantitative estimators for the overall stacked PDF performances are also given.
For what the PDF statistics of the *zspecClass* estimator concerns, we have for the two experiments in table 3.2 and described in the previous section, the results reported in table 3.3.

3.5 A further experiment

To understand how difficult it is to assess the best parameter space (known as feature selection in the context on Machine Learning methods), in this paragraph we report the results about another expriment done on the EDC2 data.
Due to the better performance of the calibration catalog on which the condition of the cut on the magnitude depths had been applied, as it is inferable from tables 3.2 and 3.3, we decided to conduct another experiment, using 17 features: besides the 8 available magnitudes, the 9 associated colors:

g-r,r-i,i-z,z-Y,Y-J,J-H, VIS-Y, VIS-J, VIS-H

This was made for the training set only with the cut on the magnitude depths. This time, however, the condition on the confidence class “6” at column #52 was not been applied, maintaining only that on the reliability of spec-z fixed by the scheme S15. For the verif catalog, the conditions applied are those applied also for the previous experiments.
Therefore, we had the catalogs

- calib_depth_mag: with the new conditions, the number of objects is #11,730;
- verif_depth_mag: with the old conditions, the number of objectss is #33,348.

Table 3.4 Network parameters for the new experiment with the PS made up of 8 available mag as features plus 7 derived colors.

Network parameter	
input Neurons	17
# Hidden layers	2
# neurons 1^{th} hidden layer	35
# neurons 2^{th} hidden layer	16
restarts	80
epochs	10,000
threshold	0.001
decay	0.1

Table 3.5 Statistics test results of the new experiment described in this section; all the statistical indicators are based on Δz=|$photo - z_0$-spec-z|/(1+spec-z), where $photo - z_0$ is the estimated MLPQNA photo-z for the non perturbed test set.

Estimator	calib_depth_mag (8magnitudes+9cols)
bias	0.012
σ	0.145
NMAD	0.044
σ_{68}	0.048
σ_{95}	0.220
outliers	8.17%
train/test	8,218/3,512

As Refcatalog, in order to calculate the polynomial coefficients for the error function we used the merging of the two catalogs quoted above, with #$45,078$ samples. The topological and training parameters of the network are in table 3.4 and the results in table 3.5. We have to note that the determination of the colors had to be done after the perturbation of the magnitudes, in order not to propagate the error on the photometry. Note, moreover that in performing this new experiment, the training set was randomly shuffled and split with new percentages for train and test set, respectively, to 70% and 30%, with respect to the previous experiments (Sec. 3.3), and that also the "decay" training parameter changed from 0.01 for the old experiments to 0.1 for the one here described.

For what it concerns the PDF statistics for spec-z position with respect to individual PDF, in table 3.7, we report the fraction of the test set *zspecClass* occurrences. As it is visible from a comparison of the results for the calib_depth_mag.csv catalog, in tables 3.1, 3.2, 3.3, and

Table 3.6 *zspecClass* occurrences for the experiment quoted in Tab. 3.7.

zspecClass	calib_depth_mag
0	407 (12%)
1	761 (22%)
2	1,938 (54%)
3	406 (12%)

tables 3.4, 3.5, 3.7, the new parameter space including magnitudes and colors, the new split percentages together with the new decay parameter, fixed at 0.1, led to better performances not only in terms of σ_{68} but also, as the spec-z statistics on the PDF concerns, in terms of the percentages of samples that had a spec-z falling within 1 bin from the PDF peak, which increased from 27% to 34%.

3.6 An attempt to infer some useful cuts of the outliers objects

This section describes a series of actions aimed to gain a deeper insight into the features of the PDF of the test set for which the spectroscopic information is available. The approach was to divide the samples of the test set in outliers and non-outliers, this time using the definition of *best-estimate* photo-z as calculated by the PDF algorithm (cf. Sec. 2.4). Remember that such value does not necessarily correspond to $photo - z_0$ (i.e. to the estimate of the photo-z for the non perturbed test set), according to the well know normalized conditions:

- |*best-estimate* photo-z- spec-z| /(1+spec-z)>0.15 for outliers that are:
  #326/3512 ($\sim 9\%$)
- |*best-estimate* photo-z- spec-z|/(1+spec-z)<0.15[2] for non-outliers that are:
  #3186/3512 ($\sim 91\%$)

The quoted division of samples between outliers and non-outliers, has been combined together to the definition of some PDF features (such as *PdfWidth*, *PdfNBins*, and so on) which have been already described in Sec. 2.5.2 in order to look for correlation and/or specific trends among outliers, non-outliers, and such features.
All this was also in order to look for useful cuts to be applied to the data in order to remove the most part of outliers for which the estimate of the PDF is unreliable, by preserving, at

[2]The 0.15 value is commonly adopted in the photo-z literature to define outliers and was initially derived from simulations.

Table 3.7 Number of objects per *zspecClass* in the two subsets of outliers and non-outliers for the test set of the calib_depth_mag catalog. *The first percentages are on the test subsets (outliers/non-outliers), the second percentages on the whole test set.

zspecClass	Outliers (#326/3512)	Non-Outliers (#3186/3512)
0	0	407 (13%-11%)*
1	0	761 (24%-22%)*
2	218 (67%-6%)*	1,720(54%- 49%)*
3	108 (33%-3%)*	298 (9%- 8%)*

Table 3.8 Statistics of the 4 indicators of the PDF features quoted above.

PDF features	Outliers				Non-Outliers			
	mean	SD	MIN	MAX	mean	SD	MIN	MAX
pdfWidth	1.63	1.11	0.04	4.34	0.61	0.69	0.02	4.44
pdfNBins	34.41	18	3	78	18.27	12.21	2	75
pdfPeakHeight	0.13	0.14	0.03	0.84	0.23	0.4	0.03	0.93
pdfNearPeakWidth	0.24	0.17	0	0.78	0.23	0.12	0	0.72

same time, a remarkable number of non-outliers.
The conditions/cuts that determined objects with reliable PDF (i.e. with appropriate values of PDF features in order to minimize the outliers) will be used to flag the "verif" samples (for which, we remember we do not have spectroscopic information) as "useful" i.e. endowed with a reliable PDF: these samples will be those with PDF features fulfilling the same conditions/cuts found for the test set PDFs. A realiable PDF for the verif catalog, to be returned for the challenge, is flagged with a 1 in the column "USE"(see Sec. 3.2.1).
First of all, in table 3.7, we reported the numbers of objects per *zspecClass* for outliers and non-outliers objects.

As expected, we had no outliers of *zspecClass* 0 and 1. As anticipated below, for what the reliability of the PDF is concerned, we decided to introduce some new indicators (PDF features) among which we expected a certain degree of correlation to exists: *pdfWidth*, *pdfNBins*, *pdfPeakHeight* and *pdfNearPeakWidth* (already defined in Sec. 2.5.2). The statistics of these 4 parameters is given in table 3.8 for outliers and non-outliers.
Looking at the mean values of PDF features in table 3.8, we can see the differences for the two populations: outliers have wider PDFs, with a higher number of bins (intervals of amplitude Δz=0.02, in which the PDF is not null) and shallower peaks with respect to those of the non-outliers samples, as expected.

3.6.1 Outlier cuts

In figures 3.2, 3.3, 3.5 are given the scatter plots for the following couples of parameters used for the evaluation of the reliability of the PDF (see previous section).
Precisely:

1. Figure 3.2: scatter plot of *pdfNBins* vs *pdfWidth*. We expected a strong correlation between these two parameters, with a collocation of the outliers in the region with higher values of pdfWidth as well as pdfNBins, and, actually, up to a certain extent, this is visible in figure 3.2. Several trials of different cuts have been done in order to remove the most part of outliers, and to see if a recalculation of the normalized statistical parameters on the test set sample led to an improvement. In figure 3.2, the red straight line corresponds to:

$$pdfNBins - 35 \times pdfWidth + 35 = 0 \tag{3.5}$$

   By keeping the objects with Eq. 3.5 ≥ 0, we removed 27% of outliers (2% of objects with respect to the whole sample);

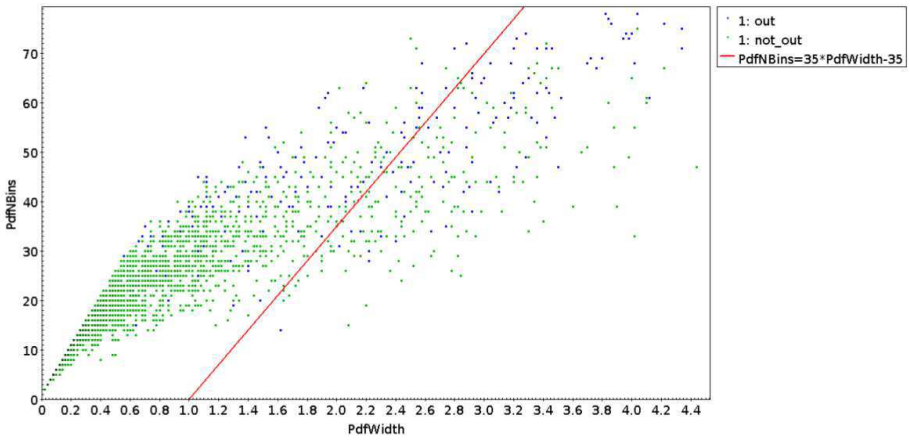

Fig. 3.2 PdfNbins VS pdfWidth.

2. Figure 3.3: scatter plot of *pdfNearPeakWidth* vs *pdfWidth*: we can note that $\sim 39\%$ of the outliers (4% of the whole sample) are under the parabolic branch, defined equation is:

$$pdfNearPeakWidth - 0.199 \times \sqrt{pdfWidth} = 0 \tag{3.6}$$

   therefore with the condition Eq. 3.6≥ 0 we kept a congruous number of non-outliers, removing the quoted fraction of outliers; moreover, the removal of samples on the left

of the vertical straight line

$$pdfWidth = 2 \tag{3.7}$$

with the condition Eq. 3.7 < 2, allowed the removal of 35% of outliers (3% on the whole sample); finally, the removal of the samples above the orizontal line

$$pdfNearPeakWidth = 0.44 \tag{3.8}$$

with the condition Eq. 3.8 < 0.44, allowed the removal of 15% of outliers (1% on the whole sample). These last cuts, discussed in equations 3.7 and 3.8 are better visible in the distributions of the relative parameters, shown for both outliers and non-outliers, in figure 3.4.

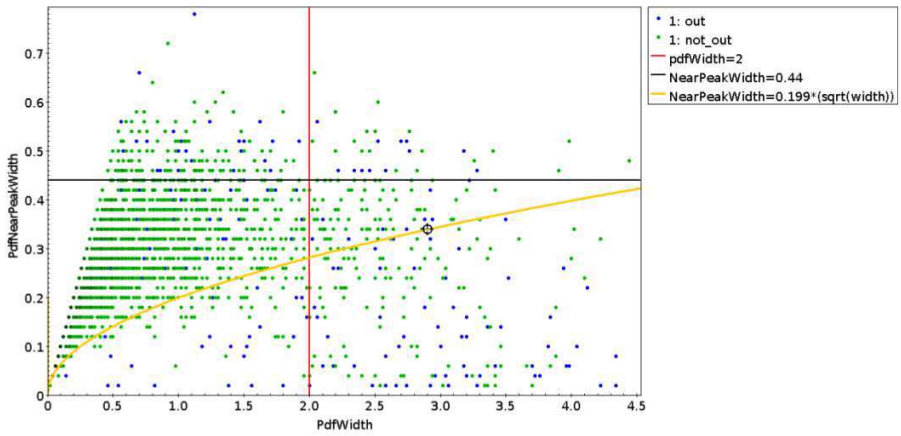

Fig. 3.3 PdfNearPeakWidth VS pdfWidth.

3. Figure 3.5: scatter plot of *pdfPeakHeight* vs *pdfWidth*: we expected a strong anti-correlation between these two parameters, although it becomes not visible from a certain width threshold up, in any case the selection of the region between the black straight line

$$pdfPeakHeight = 0.09 \tag{3.9}$$

and the hyperbolic branch

$$pdfPeakHeight - (0.13/pdfWidth) - 0.11 = 0 \tag{3.10}$$

with the conditions Eq. 3.9 > 0.09 and Eq. 3.10 ≤ 0, allowed to remove the great part of the outliers ($\sim$ 59%, 5% on the whole sample).

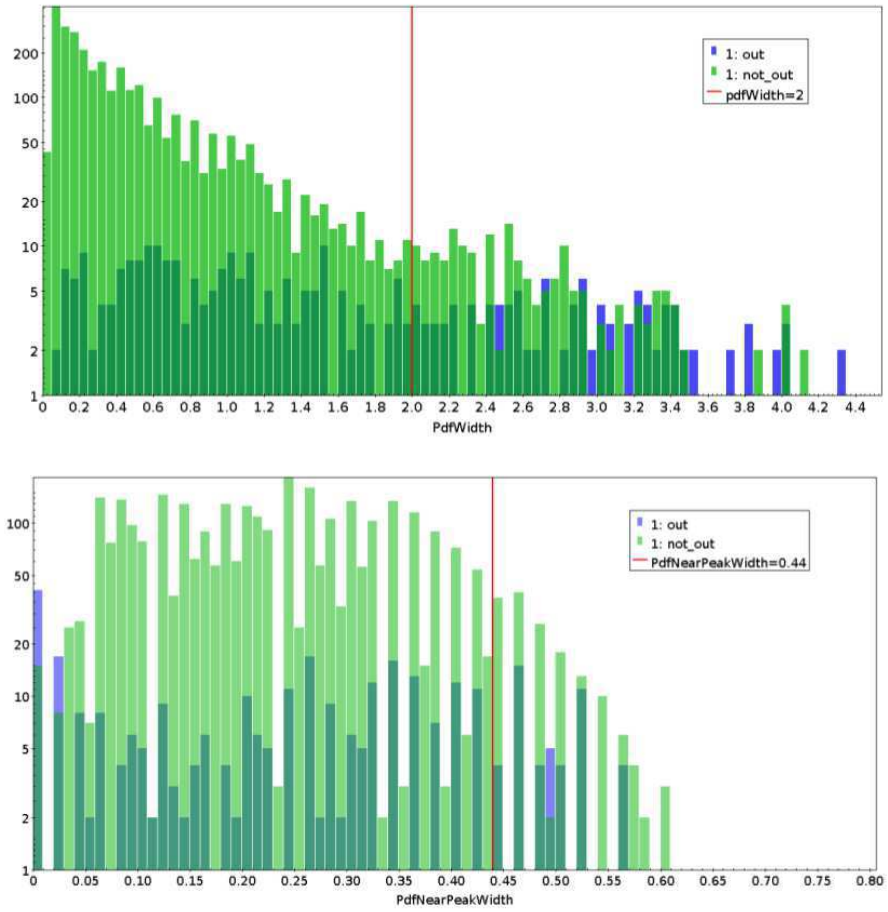

Fig. 3.4 PdfWidth distribution (top panel) and of pdfNearPeakWidth (bottom panel) for outlier and non-outliers objects in the test set: the cut of samples with pdfWidth higher than 2 and pdfNearPeakWidth higher than 0.44 ensures the compromise between leaving a congruous number of non-outliers, removing the most part of outliers.

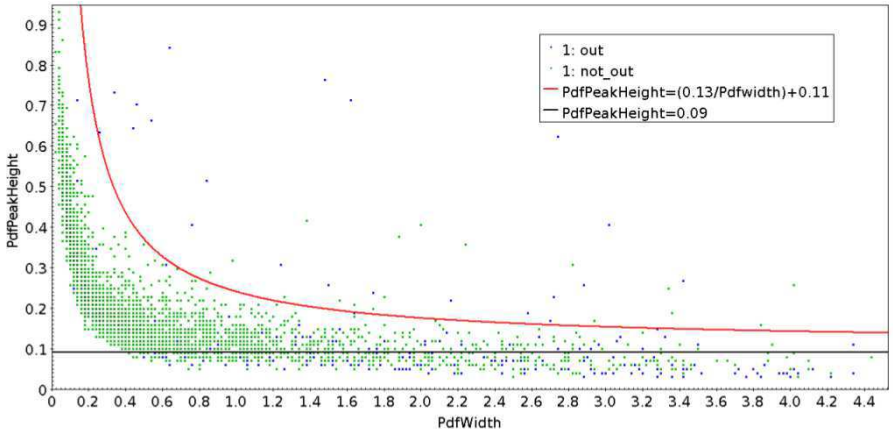

Fig. 3.5 PdfPeakHeight VS pdfWidth.

All the cuts on the PDF features will be combined in several ways, and the statistics on the test set recalculated in the following section, in order to meet the requirements of the Euclid Data Challenge 2.

3.7 The requirements of the Euclid data Challenge 2

We recall here the requirements of the EDC2. For what concerns the regression statistics between photo-z (our *best-estimate*, calculated by the PDF algorithm) and spec-z, for the samples of the verif catalog with "USE" flag set to 1, i.e. those with reliable PDFs, the Euclid requirements were:

- σ_{68}=0.05;
- % outliers $\leq$ 10%.

While, on the normalized cumulative PDF, i.e. PDF(spec-z - *best-estimate* photo-z)/(1+spec-z):

- $f_{0.05}$ equal to 68%;
- $f_{0.15}$ equal to 90%;
- $\langle \Delta z \rangle$ known better than 0.002.

For the stacked PDF we will report in table 12 also the Δz bias for samples with residuals falling within the intervals (centered on Δz=0) of amplitudes 0.05 and 0.15, respectively. In such table, the results obtained for the statistics quoted above and for several combinations of

cuts (described in the previous section) will be given. More in detail, the evaluated statistics were for the cut combinations:

1. $pdfBinsNumb \geq 35 \times pdfWidth - 35 \& pdfNearPeakWidth < 0.44$ & $pdfPeakHeight > 0.09$ & $best\text{-}estimatephoto - z < 2$ & $pdfWidth < 2$ & $0.2 \leq zspec \leq 2$;
2. $0.09 < pdfPeakHeight <= (0.13/pdfWidth) + 0.11 \& pdfWidth < 2$ & $best\text{-}estimatephoto - z < 2$ & $0.2 \leq zspec \leq 2$;
3. condition 1 without the cut in spec-z;
4. condition 2 without the cut in spec-z;
5. condition 1 plus $0.09 < pdfPeakHeight \leq (0.13/pdfWidth) + 0.11$;
6. $0.09 < pdfPeakHeight \leq (0.13/pdfWidth) + 0.11$ & $0.199 \times \sqrt{pdfWidth} \leq pdfNearPeakWidth < 0.44$ & $best\text{-}estimatephoto - z < 2$ & $0.2 \leq zspec \leq 2$;
7. $0.199 \times \sqrt{pdfWidth} \leq pdfNearPeakWidth < 0.44$ & $best\text{-}estimatephoto - z < 2$ & $0.2 \leq zspec \leq 2$ & $pdfBinsNumb \geq 35 \times pdfWidth - 35$ & $pdfWidth < 2$ & $0.09 < pdfPeakHeight$.

We note that the condition on the restriction of the spec-z to the interval $[0.2, 2]$ removed only 9% of samples from the test set (#3,182/#3,512 of the whole test set). In table 3.12, in the following section, we will report, for the seven cuts just quoted, for this experiment and for the one which will be described in the following section, a table containing the features of the individual and of the cumulative PDFs.

3.8 Other MLPQNA experiments

We decided to perform three other experiments on the calib_depth_mag.csv catalog, trained also this time, by using 17 features (8 magnitudes and 9 colors), changing the value of the decay parameter and leaving all the others topological and training parameters as fixed in table 3.4.
All the statistics on the non perturbed test set, are shown in table 3.9. For a comparison the statistical values for the decay 0.1, already shown in table 3.5, are reported in the same table. As the PDF spec-z statistics is concerned, it is reported in table 3.10 for the new three values of the decay parameter: the values, already given in table 3.7 are reported for a comparison.

Table 3.9 Statistical results for the test sets in three experiments with the features used for training, indicated in the table header.

Estimator	net1(decay 0.01)	net2(decay 0.05)	net3(decay 0.1)	net4(decay 0.15)
|bias|	0.018	0.011	0.012	0.012
σ	0.180	0.140	0.145	0.151
NMAD	0.047	0.044	0.044	0.046
σ_{68}	0.052	0.048	0.049	0.049
σ_{68}	0.268	0.231	0.220	0.223
outliers	9.11%	8.88%	8.17%	8.26%

Table 3.10 PDF *zspecClass* statistics for the test of the the three new experiment (decays 0.01-0.05-0.15):the data in table 3.7 the decay 0.1 are reported for a comparison.

zspecClass	net1(decay 0.01)	net2(decay 0.05)	net3(decay 0.1)	net4(decay 0.15)
0	403 (11%)	446(13%)	407 (12%)	414 (12%)
1	737 (21%)	803 (23%)	761 (22%)	807 (23%)
2	1,958 (56 %)	1,858 (53 %)	1,938 (54 %)	1,862(53%)
3	414(12%)	405 (12%)	406 (12%)	423 (12%)

From table 3.10, we can see that the best performance in terms of numbers of samples with a spec-z within 1 bin from the PDF peak was given for the decay value 0.05, with a 36% of samples fulfilling such condition. The old experiment, with decay value 0.1, has been outperformed also by the experiment done with decay equal to 0.15 with an increase of 1% samples falling within 1 bin from PDF peak. In table 3.11, we give the number of objects per *zspecClass* in the two subsets of outliers and non-outliers for the test set of the calib_depth_mag catalog for the three experiments performed with decay parameter equal to 0.01, 0.05 and 0.15 (in analogy to table 3.8 for decay=0.1).

Table 3.11 PDF *zspecClass* statistics for the test of the the three new experiment (decays 0.01-0.05-0.15):the data in table 3.7 the decay 0.1 are reported for a comparison.

	decay 0.01		decay 0.05		decay 0.1	
zspecClass	out	no out	out	no out	out	no out
0	0	403 (12.6%-11%)*	0	446(13.9%-13%)*	0	414 (12.9%-12%)*
1	0	737 (23%-21%)*	0	803 (25%-23%)	0	807 (25.1%-23%)*
2	219 (68%-6%)*	1,739 (54.5%-50%)*	214 (67.5%-6%)*	1,644 (51.4%-47%)*	175 (59.7%-5%)*	1,693 (52.6%-48%)*
3	101(31.6%-3%)	313 (9.8%-9%)*	103(32.5%-3%)*	302 (9.4%-9%)*	118(40.3%-3%)*	305 (9.5%-9%)*

Table 3.12 EDC2 requirement values for seven cuts of the outliers (specified above in section 3.7) and for four different decays (dec, in the first column) values for training: 0.1, 0.01, 0.05, 0.15.

#Comb-decay	#Comb-%objc	σ_{68}	outliers	$f_{0.05}$	$\langle \Delta z \rangle$ in ± 0.05	$f_{0.15}$	$\langle \Delta z \rangle$ in ± 0.15	overall $\langle \Delta z \rangle$
1 dec 0.1	#2538/72%	0.038	3.19%	67.81%	0.00085	94.09%	0.0015	-0.0018
1 dec 0.05	#2437/69%	0.038	2.75%	67.21%	0.0023	94.13%	0.0045	0.00055
1 dec 0.01	#2480/71%	0.037	3.02%	68.18%	0.0012	94.54%	0.0014	-0.0015
1 dec 0.15	#2574/73%	0.037	2.80%	67.68%	0.00099	94.99%	0.0012	-0.0011
2 dec 0.1	#2607/74%	0.038	2.76%	67.22%	0.00081	94.09%	0.0013	-0.0037
2 dec 0.01	#2430/69%	0.039	2.84%	66.24%	0.0023	93.88%	0.0045	-0.00036
2 dec 0.05	#2555/73%	0.037	3.05%	67.39%	0.0012	94.18%	0.0016	-0.0017
2 dec 0.15	#2642/75%	0.037	2.80%	67.68%	0.0010	94.82%	0.0014	-0.00090
3 dec 0.1	#2786/79%	0.040	4.23%	66.42%	0.00055	93.10%	-0.00032	-0.0063
3 dec 0.1	#2657/76%	0.039	3.46%	66.45%	0.0019	93.50%	0.0021	-0.0036
3 dec 0.01	#2728/78%	0.038	3.85%	67.35%	0.00067	93.60%	-0.00067	-0.0070
3 dec 0.15	#2810/80%	0.038	3.52%	66.88%	0.00058	94.17%	0.00054	-0.0043
4 dec 0.1	#2852/81%	0.040	3.85%	65.94%	0.00051	93.10%	-0.00044	-0.0082
4 dec 0.1	#2651/75%	0.040	3.50%	65.50%	0.0018	93.27%	0.0020	-0.0045
4 dec 0.01	#2806/80%	0.038	3.95%	66.56%	0.00070	93.18%	-0.00045	-0.0075
4 dec 0.15	#2878/82%	0.039	3.54%	66.31%	0.00059	93.99%	0.00039	-0.0051
5 dec 0.1	#2526/72%	0.038	2.77%	68.08%	0.00085	94.45%	0.0014	-0.0031
5 dec 0.01	#2347/67%	0.038	2.60%	67.12%	0.0023	94.31%	0.0046	0.00055
5 dec 0.05	#2467/70%	0.037	2.78%	68.34%	0.0012	94.71%	0.0013	-0.0025
5 dec 0.15	#2566/73%	0.037	2.61%	67.83%	0.0010	95.17%	0.0012	-0.00090
6 dec 0.1	#2581/73%	0.038	2.52%	67.13%	0.00077	94.15%	0.0014	-0.0040
6 dec 0.01	#2325/66%	0.039	2.49%	66.56%	0.0023	94.25%	0.0044	-0.0010
6 dec 0.05	#2455/70%	0.037	2.69%	67.92%	0.0011	94.67%	0.0013	-0.0030
6 dec 0.15	#2576/73%	0.037	2.60%	67.64%	0.0010	95.10%	0.0011	-0.0015
7 dec 0.1	#2473/70%	0.038	2.63%	68.23%	0.00082	94.56%	0.0014	-0.0025
7 dec 0.01	#2292/65%	0.038	2.36%	67.36%	0.0022	94.66%	0.0045	0.00025
7 dec 0.05	#2419/69%	0.037	2.73%	68.31%	0.0011	94.82%	0.0012	-0.0028
7 dec 0.15	#2556/73%	0.037	2.62%	67.86%	0.0010	95.17%	0.0012	-0.00069

With reference to the cut (4) for network 1 (decay 0.01), network 2 (decay 0.05) and network 4 (decay 0.15) we removed, respectively 41% , 36.6% and 24% of outliers (4%, 3% and 2% of the whole test set).
For what cut (5) concerned, for the networks 1, 2 and 4 we removed 49%, 42%, 32% of the outliers (4%, 4%, 3% of the whole test set), respectively.
For the cut in (6), we removed the 47%, 42% and 35% (4%, 4%, 3% of the whole sample) for respectively networks 1, 2 and 4. For the cut in (7) we removed for networks 1, 2, 4, respectively, 10% , 13%, and 17% of outliers (1%, 1%, 1% on the whole sample).
We then calculated for the seven cut conditions reported in Sec. 3.7, and for the four values of decay network parameter, both the individual photo-z estimate statistics, as well as the statistics regarding the whole performances of the PDF, by obtaining (table 3.12):

- the worst performance in terms of EDC2 requirements ($f_{0.05}$, $f_{0.15}$) and the highest values of $< \Delta z >$ bias within 0.05 and 0.15 (see table 3.12), was given for the training phase performed with decay 0.01, as it was visible also in Sec. 3.8 for what concerned the results about the regression statistics on the non-perturbed test set. Despite this, the worst performance, for all the conditions, for the EDC2 requirement on the overall $< \Delta z >$ bias was given just in correspondence of this decay. Therefore it has been discarded for the preparation of the validation catalog to be returned for the EDC2;

- the best performances in terms of lower $< \Delta z >$ bias values and higher percentages of no-cut samples, have been obtained in correspondence of decay 0.1 and 0.15 with respect to decay 0.05, although to this latter decay value corresponds the highest values of $f_{0.05}$, for all conditions (combinations of cuts) and moreover for at least one useful (compared also to the other requirements) value of overall $< \Delta z >$ bias (condition 1); to the conditions 3 and 4, that were the same of 1 and 2 without the cut in spec-z (see Sec. 3.7). We can see, as expected by the increased number of remaining sources (we remember that the restriction on spec-z inteval $[0.2, 2]$ removed 9% samples), that both the performance on the individual and the stacked PDFs are worse with respect to the conditions 1 and 2 in which such cut has been applied. Since the absolute values of $< \Delta z >$ bias, both within 0.05 and 0.15, are almost comparable for decay 0.1 and 0.15, for all the conditions, and slightly lower than those for decay 0.05, the final choice of the best condition to be applied had to be fixed on the basis of the best compromise between EDC2 requirements and the number of remaining objects, by making a choice if we wanted to keep also the $< \Delta z >$ bias values within 0.05 and 0.15 or not.

- In the first case, looking to all the quoted condiderations, the best choices seemed to be conditions 1 and 7 for the decay 0.1: however, between the two, the best one appears

to be condition 1 since, despite $f_{0.05}$= 67.81% we have a 2% of samples more, and half the overall $< \Delta z >$ bias with respect to the other. Instead, the best condition for decay=0.15, is the 7 since the very low overall $< \Delta z >$ bias, the 73% of remaining objects, and a $f_{0.05}$= 67.86% that is very satisfactory. For this case, condition 7 with decay 0.15 seemed the best.
However, the condition 2, with a likewise satisfactory value of $f_{0.05}$= 67.68%, allowed to gain a 2% of remaining samples. At the end, in this case, the best choice is condition 2 with decay=0.15.

- In the second case, the best conditions are 1 and 5 with decay=0.05, even though condition 5 corresponds to a very low number of remaining objects (69%). At the end, the best choice turned out to be condition 1 with decay=0.05.

3.9 Last actions to create the validation catalogs to be returned for the challenge

Among all the experiments shown until now, the best (in terms of number of remaining objects, removal of outliers and relative improvement of the performance on individual and stacked PDFs statistics) couples "best training parameters configuration+ best cut conditions chosen on the test set" were two:

- (A) training configuration: features = (8 magnitudes+9 colors), decay= (0.05), split pecentages train/test sets (70-30%)+ cuts condition number 1;
- (B) training configuration: features = (8 magnitudes+9 colors), decay= (0.15), split pecentages train/test sets (70-30%)+ cuts condition number 2.

Finally, the PDF algorithm was applied to the MLPQNA outputs for the verif_depth_mag catalog, only devoid of NaN entries (#140,944/#190,462 of the original catalog), obtained for the two quoted values of decay parameters.
All the prescriptions applied to the calibration catalog, and the cuts condition number 1 and 2 applied to the test set, have been "translated" in appropriate flags on the "verif" catalogs PDFs. In particular, were flagged "0" in the "USE" flag column (see below), all the samples of the validation catalog with:
For couple training "configuration+cuts condition (A)":

- mag values deeper of the depth mag cut values within 5 sigma, applied to the calibration catalog;

- all the samples with the FLAG_DETECT ≥ 4
- all the samples with mag errors >1
- $pdfBinsNumb < 35 \times pdfWidth - 35$
- $pdfNearPeakWidth \geq 0.44$
- $pdfPeakHeight \leq 0.09$
- *best-estimate* photo-z >2
- $pdfWidth \geq 2$

For couple "training configuration+cuts condition (B)"s:

- mag values deeper of the depth mag cut values within 5 sigma, applied to the calibration catalog;
- all the samples with the FLAG_DETECT ≥ 4
- all the samples with mag errors >1
- $pdfPeakHeight < 0.09 | pdfPeakHeight \geq (0.13/pdfWidth) + 0.11$
- $pdfWidth > 2$
- *best-estimate* photo-z >2

Finally, for all the NaN mag entries of the validation catalog, the “REDSHIFT” column (see below) is fixed to “-99” , the PDF identically =0 and the “USE” column to 0. At the end, the fits files:

euclid_cosmos_DC2_2fwhm_S2_v2_DESnoise_results_A.fits
euclid_cosmos_DC2_2fwhm_S2_v2_DESnoise_results_B.fits

were created in correspondence of the above quoted couples (A) and (B). They contain a number of rows equal to all the rows of the validation catalog (#190,462) and the following columns:

- a first column “REDSHIFT” containing *best-estimate* photo-z;
- a second column “USE” on the reliability of the redshift to be put to 1 if reliable, 0 otherwise;

- a third column “STAR” fixed at -99 since we did not perform classification experiments;
- a fourth column “AGN” fixed at -99 since we did not perform classification experiments;
- from the fifth to the final column there is the PDF for each bin from redshift 0 to 6 with a Δz=0.02. The PDF is referred to the mean of each bin.

At the end we obtained a reliable redshift for:

- couple (A): #20,424 objects, that represent the 14% of the objects of the validation catalog, devoid of NaN in the photometry;
- couple (B): #22,744 objects, that represent the 16% of the objects of the validation catalog, devoid of NaN in the photometry.

3.10 The true definitive EDC2 catalogs

The 8^{th} of Jenuary 2016, we received a new improved version of catalogs. Calib and verif catalog were always split in RA (calib containing objects with RA > 150.125; verif lower). The main changes between the old couple of catalogs calib and verif (from which the PDF results catalogs, shown in the previous section have been prepared), and the new couple of catalogs, were:

- a higher number of objects both in the calib and verif catalogs: exactly, #198,435 “new calib” objects vs #190,508 “old calib” (about 8k objects more), and #192,864 “new verif” objects vs #190,462 of “old verif” (more than 2k objects added): in total we had more than 10k objects added to the Euclid Data Challenge field;
- the magnitudes and fluxes were not more corrected for Galactic extinction, however the correction factors for each flux and objects were provided, for the photometry of both catalogs, in order to proceed to the correction;
- a higher number of spectroscopic redshifts added;
- the introduction of a new photometric flag, named “FLAG_PHOT” (equal to 0 for good photometry), whose application was recommended;
- a rule to select, and then exclude true stars, based on the two flags “STAR” and “reliable_S15”, i.e. : to select "true stars", set reliable star is STAR = 1 & reliable_S15 = 1;

Table 3.13 Number of galaxies (third column) classified as no-AGNs and AGNs (fourth and third row, respectively) , and of those objects classified either as stars and AGNs (first row), and the corresponding number of objects which are X-rays emitters (fourth column).

STAR FLAG	AGN FLAG	#objs	X-Ray FLAG
1	1	334	334
1	0	11,931	0
0	1	967	967
0	0	185,203	0

- the photometry previously split in as many catalogs as the apertures in which the it has been calculated, i.e., 1fwhm, 2fwhm and 3fwhm, had been merged into one (FLUX_X_1, FLUX_X_2 etc).

To take into account all these changes, the actions we applied were:

- to correct the fluxes and magnitudes for Galactic extinction;
- to ascertain again the validity of our method to exclude true stars on the new calib catalog. Indeed, we applied a Xor condition (as already done for the preliminary catalogs, see Sec. 3.2.1) to the flags “STAR” (col #122) and “AGN” (col #123), we removed objects with spec-z=0, and we compared such objects with the X-ray emitters objects, by means of the X-ray flag (col #120). Also for the new calib catalog, such method ensured the removal of the stars alone.

Indeed, by looking at the numbers in table 3.13, the classification of the X-ray emission, appears very reliable. Therefore we decided to keep the 334 objects in the first row of table 3.13, by considering reliable for them the classification as AGNs.

3.11 The results delivered by the Consortium

Composed the new delivered "verif" catalogs according to the rules specified in the end Sec. 3.2.1, and applying the conditions (A) and (B) described in the previous section, after the blind test performed by the Consortium we obtained the following plots as results of photo-z point estimate statistics, as well as in terms of whole stacked PDF statistics performance, calculated in tomographic bins of redshift.
The photo-z individual estimation are shown in figures 3.6 and 3.7.

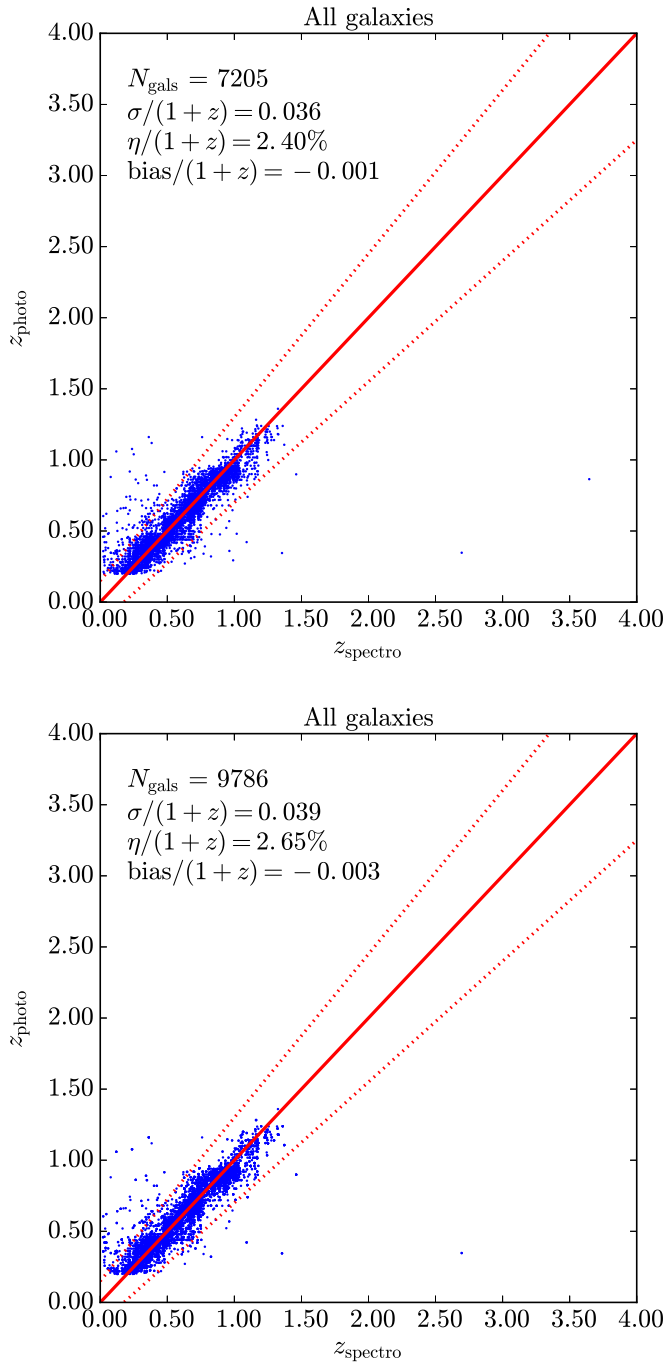

Fig. 3.6 Photo-z vs spec-z plots for the "verif" catalog fulfilling condition (A). Plots courtesy of J. Coupon.

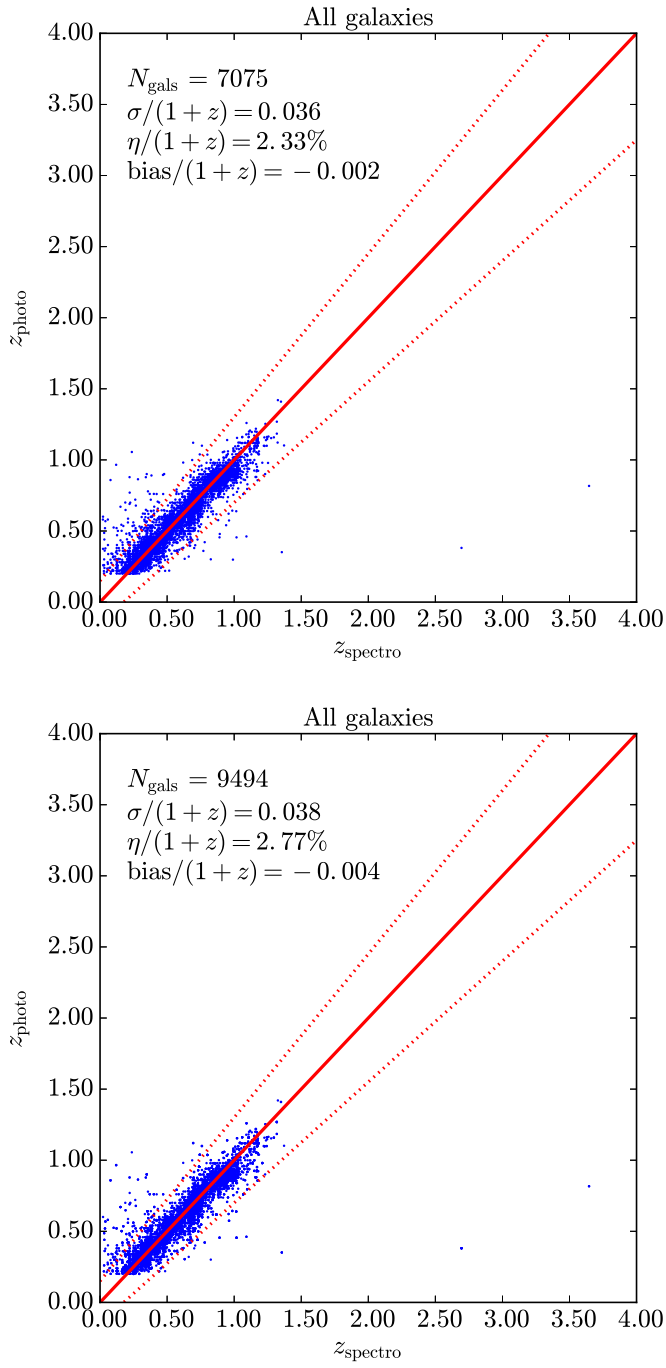

Fig. 3.7 Photo-z vs spec-z plots for the "verif" catalog fulfilling condition (B). Plots courtesy of J. Coupon.

For only one of the catalog delivered, we obtained also the results about the performances of the whole stacked PDF in tomographic bins of redshifts. We report them in figures 3.8 and 3.9. In such figures also the quantities $f_{0.05}$, $f_{0.15}$ and the total average $\langle \Delta z \rangle$, useful to evaluate the overall stacked PDF performances are reported.
We fulfill the Euclid requirements for $f_{0.15}$ in all the tomographic bins except the last one. A fact which could be expected due to the small number of training points.
For what the fraction $f_{0.05}$ is concerned, we fulfill the Euclid requirements for bins, 3, 4, 5, 6, i.e. for a redshift range from 0.55 to 0.9. The condition on $f_{0.05}$ is not fulfilled in the brighter and fainter parts of this tomographic analysis. If the reason why the condition fails in the fainter part could be explained with the low number of sources, it is more difficult to explain this behavior for brighter objects. It is likely could be due to several peculiar effects in the parameter space that the network is not able to generalize in a suitable manner.

3.12 Euclid Data Challenge 2 outcome and Conclusions

The Euclid Data Challenge 2 has been performed by seven competing teams. We report in this Section the final outcome in terms of photo-z σ and fraction of outliers along with the used methods to calculate photo-z's and relative PDFs. It is important to stress that the METAPHOR results shown in figures 3.6, 3.7, 3.8, 3.9 are calculated for the reliable galaxies provided for the Challenge (those having the flag $USE = 1$ as described in Sec. 3.2.1.)
Analogous plots have been produced for all the participants, and obviously it is not possible to report all of them in this thesis. However the main results about the performances of all the participants (indicated by name) are shown in table 3.2. These results are given considering all the galaxies provided by the participants (not only the reliable ones flagged with $USE = 1$, see above). Moreover, a summary of the scatter plots for the photo-z estimation is given in figure 3.10.

From table 3.14 and figure 3.10 it is visible that the best precision was achieved by our group ("Brescia") with MLPQNA whereas the highest completeness was obtained with the Multi Layer Perceptron used by "Hoyle". However no method fulfills all the Euclid requirements (see Sec. 3.7) for photo-z precision and outlier fraction.
It is mandatory to highlight that a fair comparison among the Machine Learning participant performances is not trivial in the light of all the procedure needed to obtain our results. Indeed, in the previous sections, a lot of criteria have been applied in order to favor the accuracy of our results, rather than the completeness. This has led to the removal of the faintest sources, thus leading to better performances in terms of precision on the remaining dataset. This must

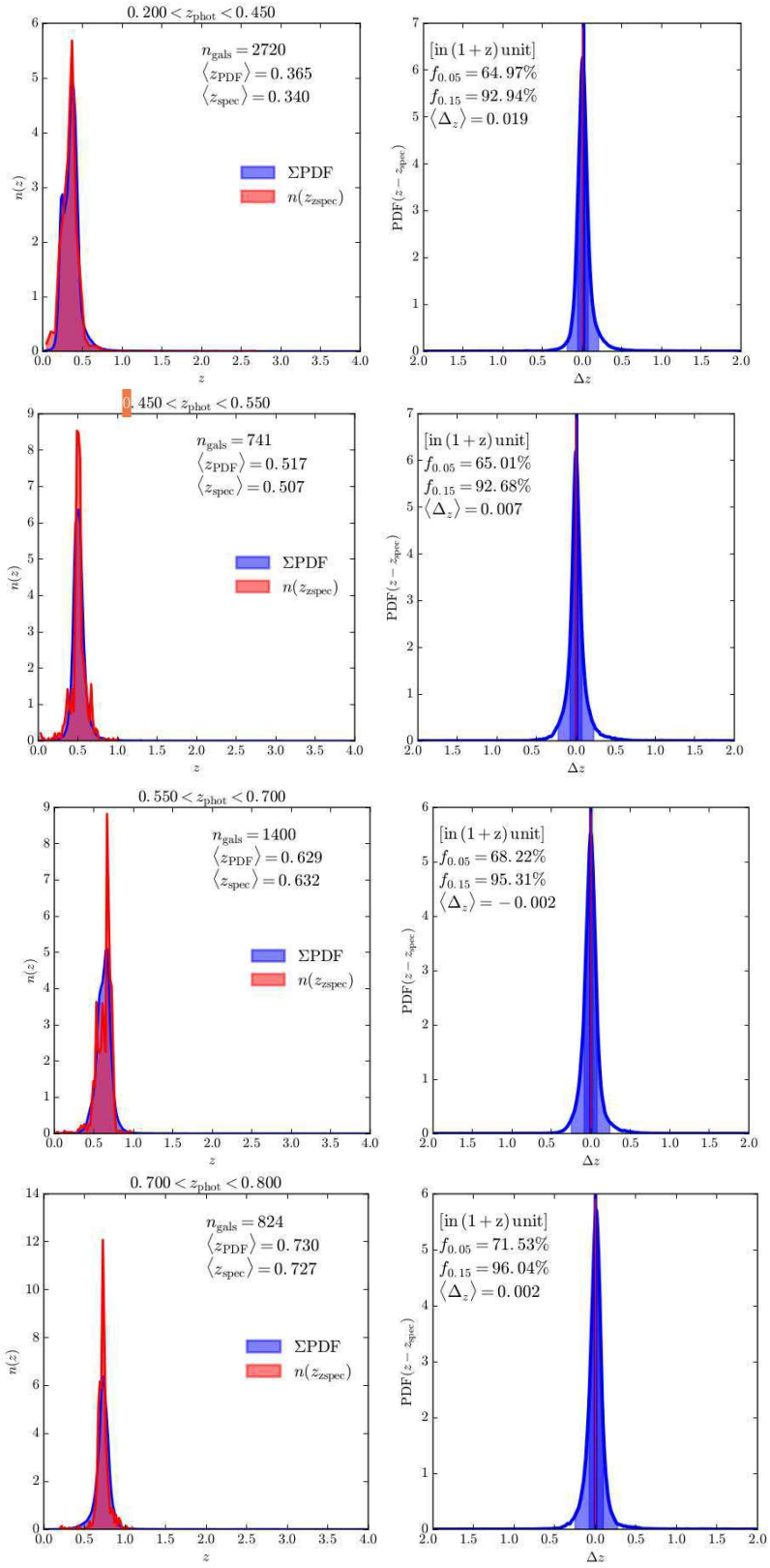

Fig. 3.8 Photo-z vs spec-z plots in their stacked representation (left panels) and stacked representation of the residuals Δz (right panels) for the "verif" catalog fulfilling condition (A) for four tomographic bins of redshift ranging from 0.4 to 0.8. In the plots are reported also the fractions $f_{0.05}$,$f_{0.15}$ and the total average $\langle \Delta z \rangle$. Plots courtesy of J. Coupon.

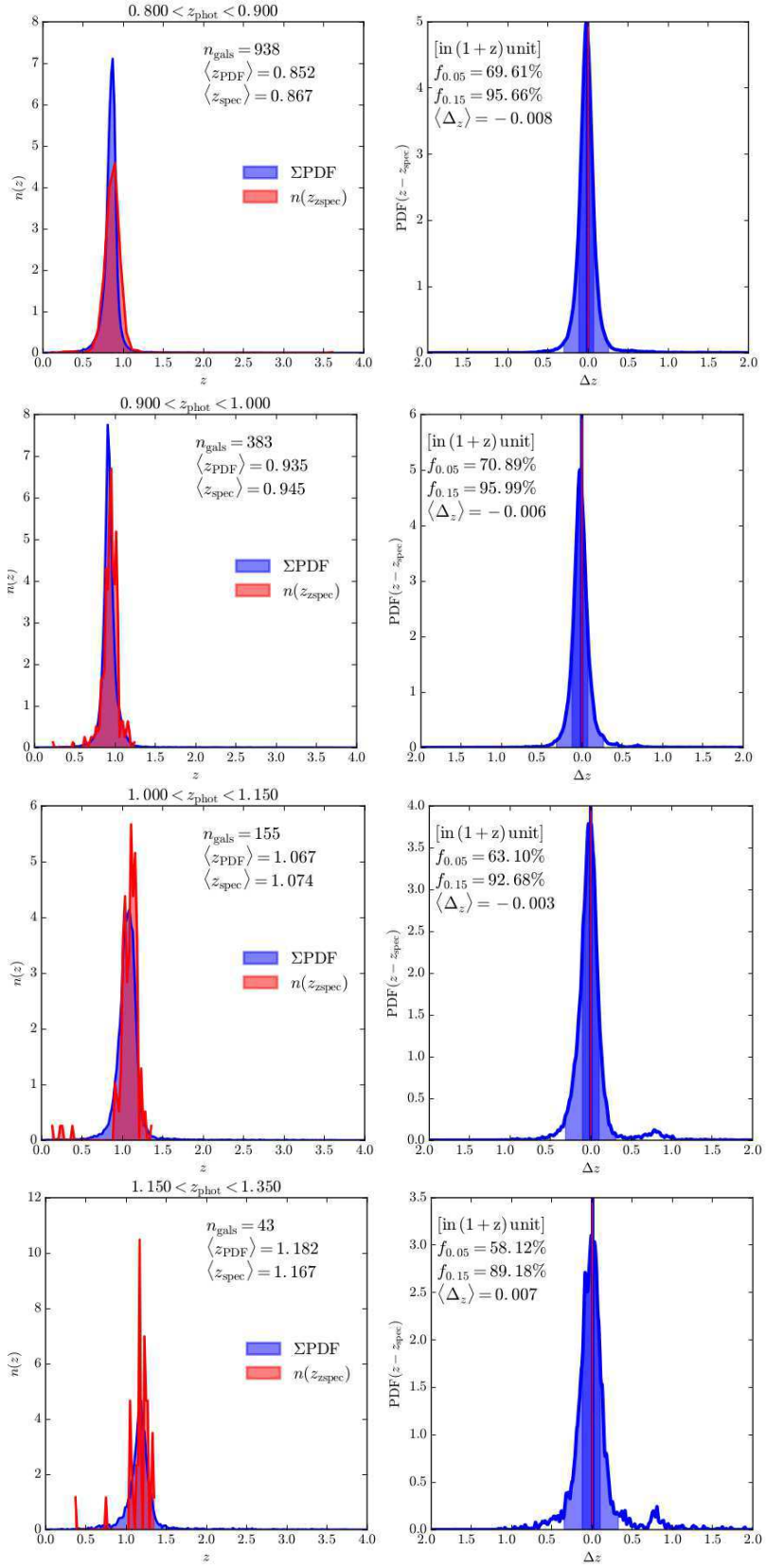

Fig. 3.9 Photo-z vs spec-z plots in their stacked representation (left panels) and stacked representation of the residuals Δz (right panels) for the "verif" catalog fulfilling condition (A) for four tomographic bins of redshift ranging from 0.8 to 1.35. In the plots are reported also the fractions $f_{0.05}$,$f_{0.15}$ and the total average $\langle \Delta z \rangle$. Plots courtesy of J. Coupon.

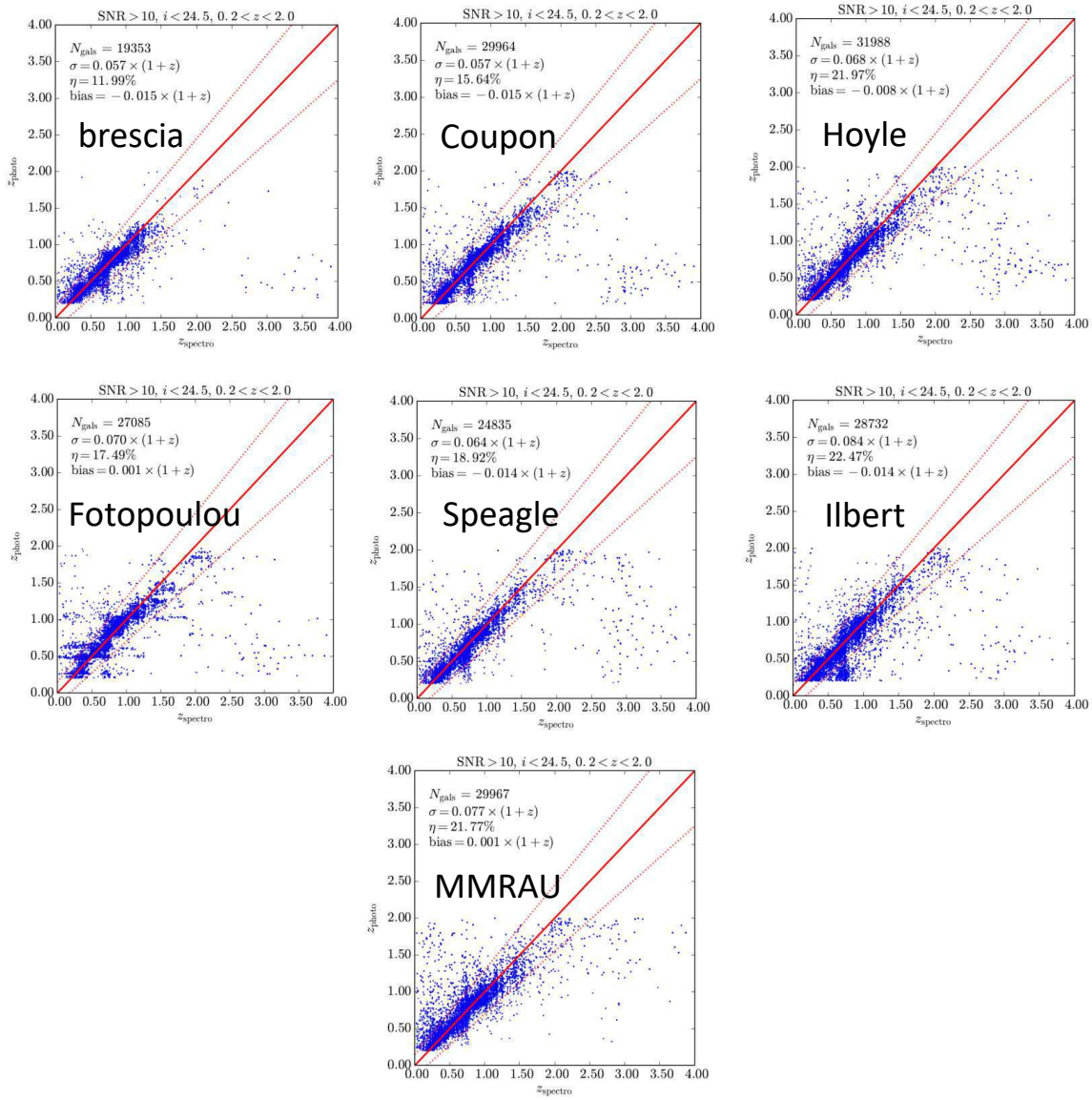

Fig. 3.10 Photo-z vs spec-z plots for the "verif" catalog fulfilling condition for all the Challenge participants. Plots courtesy of J. Coupon.

Table 3.14 Euclid Data Challenge 2 main outcome for all the galaxies (not only those flagged as reliable). First Column: Partecipant name; Second Column: Used code acronym (see Chapter 1 for their explanation); Third Column: σ; Fourth Column: Outlier Fraction; Fifth Column: fraction of galaxies in the range $0.2 < z < 2.0$ relative to the highest score ("Hoyle"). Results courtesy of J. Coupon.

Name	Code Name	σ	Outlier Fraction (%)	Relative Fraction
Brescia	MLPQNA	0.057	11.99	0.60
Hoyle	ML	0.068	21.97	1.0
Fotopoulou	LePhare	0.070	17.49	0.85
Rau	ANNz	0.077	21.77	0.94
Speagle	SOM+RF	0.064	18.92	0.78
Coupon	LePhare+ColorPrior	0.057	15.6	0.94
Ilbert	LePhare	0.084	22.47	0.90

be considered in the evaluation of a comparison among all the ML techniques results shown in table 3.14: such confrontation cannot be interpreted straightforwardly.

Chapter 4

METAPHOR for SDSS DR9

(extracted from S. Cavuoti, V. Amaro, M. Brescia, C. Vellucci, C. Tortora and G. Longo, "METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts", MNRAS, 2017, 465, 1959–1973)

4.1 Introduction

In this Chapter, we present a summary of results obtained applying METAPHOR to the Sloan Digital Sky Survey-Data Release 9 (SDSS DR9, hereafter) galaxy data, and a direct comparison with the PDFs obtained using the Le-Phare spectral energy distribution template fitting.
We will show that METAPHOR is capable to estimate the precision and reliability of photometric redshifts obtained with three different self-adaptive techniques, i.e. MLPQNA, Random Forest and the standard K-Nearest Neighbors models.
We presented in Chapter 2 METAPHOR, which tries to account in a coherent manner for the uncertainties in the photometric data to find a perturbation law of the photometry, which could include not only a special procedure for a fitting of the errors on the attribute themselves, but also a level of randomness to be added to the information obtained from the errors.
This in order to perform the perturbation of the attributes that have those errors, in a controlled, not biased by systematics, way. A proper error fitting, accounting for the attribute errors, allows us to constrain the perturbation of photometry on the biases of the measurements.
We remember that from a theoretical point of view, the characterization of photo-z predicted by empirical methods should disentangle the photometric uncertainties from those intrinsic to the method itself.
The perturbation law, described in Sec. 2.2.1 foresees four different type of function for the

Table 4.1 The *psfMag*-type magnitude cuts derived in each band during the KB definition.

Band	brighter limit	fainter limit
u	17.0	26.8
g	16.0	24.9
r	15.4	22.9
i	15.0	23.3
z	14.5	23.0

Eq. 2.3, that we repeat here for convenience:

$$\tilde{m}_{ij} = m_{ij} + \alpha_i F_{ij} u_{(\mu=0,\sigma=1)} \tag{4.1}$$

i.e. F_{ij} can be either flat, or individual, or polynomial or bimodal (cf. Sec. 2.2.1.

In this Chapter we show an analysis of the performances for all the four choices we have at disposal to perturb photometry through the function F_{ij}, both for what regards the statistics on the punctual photo-z estimates and on the individual as well as *stacked* PDFs, using almost all the statistical indicators described in Sec. 2.5.

4.2 SDSS Data

In order to evaluate the performance of the METAPHOR processing flow, we used a galaxy spectroscopic catalog extracted from the SDSS DR9 (York, 2000).
The SDSS combines multiband photometry and fiber-based spectroscopy, providing all information required to constrain the fit of a function mapping the photometry into the spectroscopic redshift space. The KB for the presented experiment is composed of objects with *specClass* galaxy together with their photometry ($psfMag$-type magnitudes) and rejecting all objects with non-detected information in any of the five SDSS photometric bands (the original query is in Cavuoti et al. (2017)). From the original query, we extracted $\sim 50,000$ objects to be used as train set and $\sim 100,000$ objects to be used for the blind test set. The redshift distributions for the train and test sets are shown in Fig. 4.1. The train and test sets are drawn from the same population distribution in order to minimize the occurrences of biases/mismatch between train and test samples, which could induce degeneracies in the predicted photo-z. The ranges in terms of magnitudes are reported in Table 4.1 and detailed in Brescia et al. (014b).

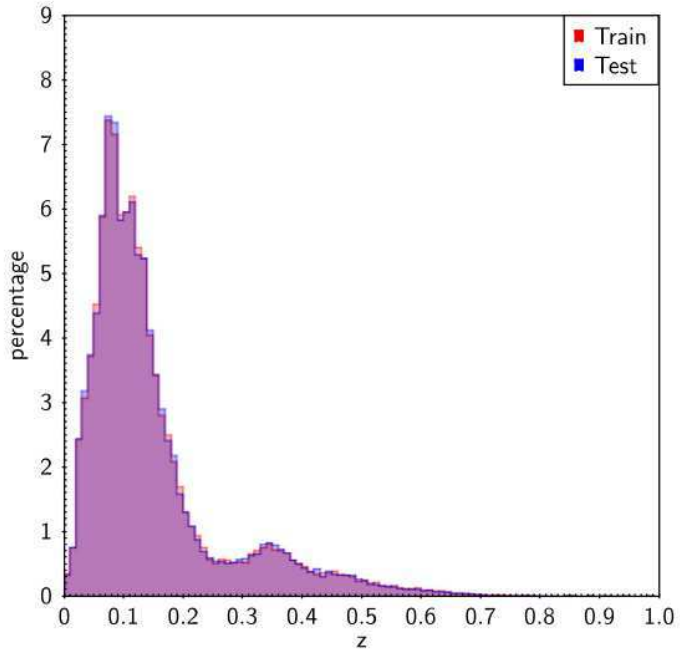

Fig. 4.1 Distribution of SDSS DR9 spectroscopic redshifts used as a KB for the PDF experiments. In blue, the blind test set, and in red, the training set. The values are expressed in percentage, after normalizing the two distributions to the total number of objects.

4.3 A comparison among photo-z estimation models

As already mentioned several times, the METAPHOR procedure can be, in principle, applied by making use of any arbitrary empirical photo-z estimation model. Moreover, as it was introduced in Sec. 1.1, the alternative category of photo-z estimation methods, based on SED template fitting, intrinsically provides PDFs. Therefore, since METAPHOR is a wrapper of the particular interpolative method chosen by the user, we experimented the METAPHOR procedure with three different empirical methods, for instance, Multi Layer Perceptron with Quasi Newton learning rule (MLPQNA) neural network, K-Nearest Neighbours (KNN) and Random Forest (RF), and compared their results with the Le-Phare SED template fitting technique.

In particular, the use of different empirical models has been carried out in order to verify the universality of the procedure with respect to different empirical models. It must also be pointed out that, aside from the selection of the RF model, the choice of the KNN method has been driven by its extreme simplicity with respect to the wide family of interpolation techniques.

Therefore, by validating the METAPHOR procedure and PDF statistical performance with

KNN, it would empirically demonstrate its general applicability to any other empirical method. All these methods are briefly described in the following sections. According to the traditional supervised paradigm of ML, the KB used is split into different subsets, dedicated to training and test steps, respectively. The training set is used to learn the hidden relationship between photometric and spectroscopic information, while the blind test set allows the evaluation and validation of the trained model on objects never submitted before to the network. In order to analyze the results on the test objects, a series of statistical estimators is then derived (see Sec. 2.5).
For what the interpolative methods are concerned, since METAPHOR has been exaustively discussed in Chapter 2, and Random Forest (RF) has been presented in Chapter 1, just KNN will be described in the next section, along with the description of Le-Phare discussed in Sec. 4.3.2.

4.3.1 KNN

In a KNN model (Cover and Hart, 1967), the input consists of the K closest training examples in the parameter space. A photo-z is estimated by averaging the targets of its neighbours. The KNN method is based on the selection of the N training objects closest to the object currently analyzed. Here, closest has to be intended in terms of Euclidean distance among all photometric features of the objects. Our implementation makes use of the public library SCIKIT - LEARN (Pedregosa, 2011).

4.3.2 Le-Phare SED fitting

To test the METAPHOR workflow against to SED fitting model, we used the Le-Phare (Arnouts et al. 1999, Ilbert et al. 2006) code as a benchmark. SDSS observed magnitudes were matched with those predicted from a set of SEDs. Each SED template was redshifted in steps of Δz= 0.01 and convolved with the five SDSS filter transmission curves. The following merit function was then minimized:

$$\chi^2(z,T,A) = \sum_{i=1}^{N_f} \left(\frac{F^f_{\rm obs} - A \times F^f_{\rm pred}(z,T)}{\sigma^f_{obs}} \right)^2 \tag{4.2}$$

where $F^f_{\rm pred}(z,T)$ is the flux predicted for a template T at redshift z. $F^f_{\rm obs}$ is the observed flux and σ^f_{obs} the associated error derived from the observed magnitudes and errors. The index f refers to the considered filter and N_f is the number of filters. The photometric redshift is determined from the minimization of $\chi^2(z,T,A)$ varying the three free parameters: the

photometric redshift, z = z_{phot}, the galaxy spectral type T, and the normalization factor A. For the SED fitting experiments with Le-Phare , we used the SDSS *Modelmag* magnitudes in the u, g, r, i and z bands (and related 1σ uncertainties), corrected for galactic extinction using the reddering map in Schlafly and Finkbeiner (2011).
As a reference template set, we adopted the 31 SED models used for the COSMOS photo-z (Ilbert, 2009). The basic COSMOS library is composed of galaxy templates from (Polletta, 2007), which includes three SEDs of elliptical galaxies (E) and five templates of spiral galaxies (S0, Sa, Sb, Sc, Sd). These models are generated using the code GRASIL (Silva et al., 1998), providing a better joining of ultraviolet and mid-infrared than those by Coleman and Weedman (1980) used in Ilbert et al. (2006). Moreover, to reproduce very blue colors not accounted for by the Polletta (2007) models, 12 additional templates using Bruzual and Charlot (2003) models with starburst ages ranging from 3 to 0.03 Gyr have been added. In order to improve the sampling of the redshift–color space and therefore the accuracy of the redshift measurements, the final set of 31 spectra was obtained by linearly interpolating the original templates. We have finally imposed the flat prior on absolute magnitudes, by forcing the galaxies to have absolute i-band magnitudes in the range of $(-10, -26)$.
Le-Phare, as it is usual in the case of SED template-fitting techniques, provides the PDF for the estimated photo-z through the χ^2 distribution, which is defined as:

$$PDF(z) \propto exp\left(-\frac{\chi^2(z) - \chi^2_{min}}{2}\right), \tag{4.3}$$

where $\chi^2_{min}$is the minimum of χ^2, corresponding to the best-fitting redshift.
We wish to stress that our main interest was to check the consistency of our ML-based results with PDFs from standard SED-fitting procedures, without running any competition among different methods. For this reason, we used a basic implementation of the Le-Phare code, not taking into account the systematics in the templates, data sets and optimizations (Brammer et al. 2008, Ilbert 2009, Tanaka 2015), and only imposing a flat prior on the absolute magnitudes. In literature, most of such systematics are taken into account introducing zero-point offsets and a template error function.
Zero-point offsets in the photometric bands due to a bad calibration and uncertainties in the model templates (e.g. stellar tracks, extinction law and other features not included in the spectra) can produce shifts between the predictions and real data. These average shifts are usually determined by means of an iterative process which minimizes the χ^2 for the spectroscopic sample with the redshift set to the spec-z value. Then, these shifts were applied to the magnitudes and used for the redshift determination (Ilbert, 2009). We have done some tests, and except for the more uncertain u band, for which the shift can also reach

values of 0.1 mag or more, for the other bands the shifts are less than 0.01 mag; thus, for the sample under analysis and for the main objectives of the this analysis, the contribution from zero-point shifts was negligible.
Since no template is immune to these systematics, in general it is also possible to introduce an error budget in the χ^2 minimization to account for them. However, this error budget would be less than ~ 0.05 and varies a little across the wavelengths probed by SDSS bands (see e.g. Brammer et al. (2008)). Tanaka (2015) generalized the error function in Brammer et al. (2008), adding a systematic flux stretch to the random flux uncertainty, used to reduce the mismatch between data and models. Both the terms account for systematics at a few per cent level in the optical wavelengths. The calculation of this error function could be coupled with zero-point shifts.

4.4 Results and discussion

The stacked PDF has been obtained by considering bin by bin the average values of the single PDFs. The cumulative statistics used to evaluate the stacked PDF quality have been derived by calculating the stacked PDF of the residuals Δz. In this way, aside from the evaluation of PDFs for single objects (a sub-sample is shown in Fig. 4.2), it is possible to obtain a cumulative evaluation within the most interesting regions of the error distribution.

In order to compare the different perturbation laws described in Sec. 2.2.1 and repeated here in Eq. 4.1, we performed a variety of experiments with MLPQNA using 100 photometric perturbations. Results are summarized in Table 4.2. The most performing experiment turns out to be number 8, where we made use of a bimodal perturbation law with threshold 0.05 and a multiplicative constant $\alpha = 0.9$ (see equation 4.1). This experiment leads to a stacked PDF with ~ 92 per cent within $[-0.05, 0.05]$, $\sigma_{68} = 0.019$, ~ 21 per cent of the objects falling within the peak of the PDF, ~ 53 per cent falling within one bin from the peak and ~ 82 per cent falling within the PDF. We therefore run an additional experiment using the same configuration as in number 8 but improving the error representation using 1000 perturbations. This experiment led to an increase in the performances: $\sigma_{68} = 0.018$ and ~ 21.8 per cent within the peak of the PDF, ~ 54.4 per cent within one bin from the peak and ~ 89.6 per cent inside the PDF.
In order to verify the universality of the procedure with respect to the multitude of methods that could be used to estimate photo-z, the use of three different empirical models (for instance, MLPQNA, RF and KNN) has been carried out. We also derived PDFs with the Le-Phare method, in order to evaluate the quality of the produced PDFs using a classical

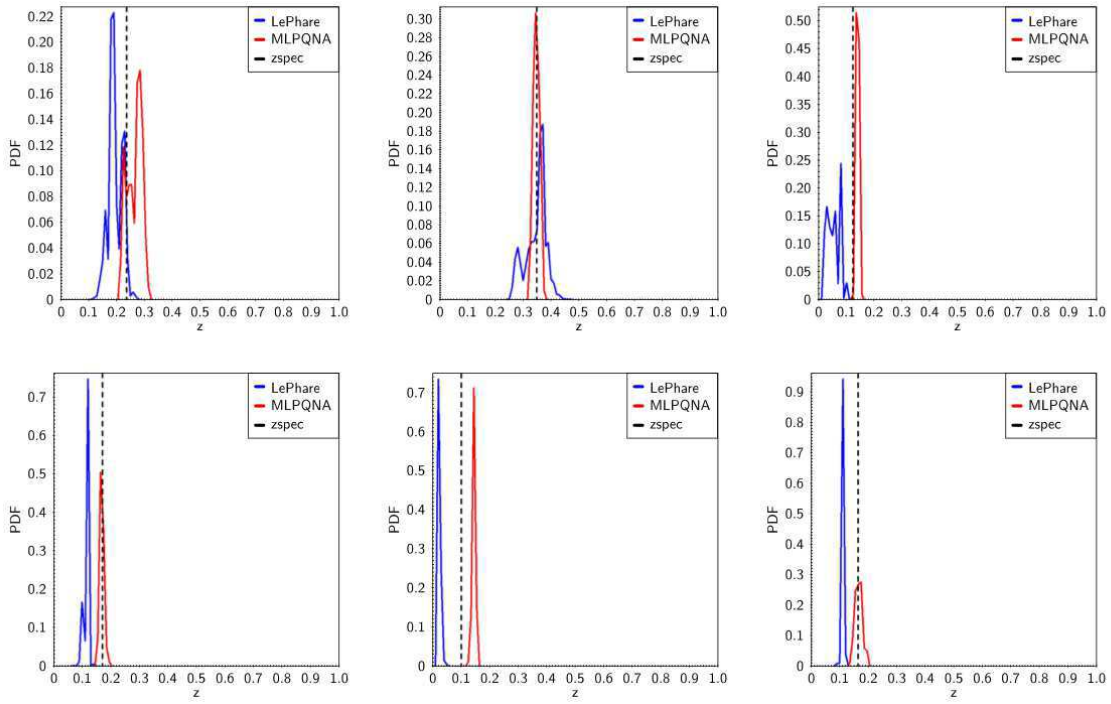

Fig. 4.2 Some examples of photo-z PDF for single objects taken from the test set, obtained by MLPQNA (red) and Le-Phare (blue). The related spectroscopic redshift is indicated by the dotted vertical line. In some cases, the PDF peak appears lowered, due to an effect of a spread over a larger range of the peak (panel in the lower right-hand corner).

SED template fitting model as a benchmark.
In Table 4.3, we report the results in terms of the standard set of statistical estimators used to evaluate the quality of predicted photo-z for all methods. The results about the statistics of the stacked PDFs are shown in Table 4.4.

Table 4.2 Results for the various experiments obtained with MLQPNA. Column 1: identification of the experiment; column 2: type of error perturbation; column 3: threshold for the flat component; columns 4 - 10: $f_{0.05}$, $f_{0.15}$, z, bias, σ, σ_{68} , NMAD (see Sec. 2.5; column 11: fraction of outliers outside the 0.15 range; column 12: skewness of the z; columns 13 - 16: fraction of objects having spectroscopic redshift falling within the peak of the PDF, within 1 bin from the peak, inside the remaining parts of the PDF and outside the PDF, respectively.

ID	Type	Threshold	$f_{0.05}$	$f_{0.15}$	Δz	|bias|	σ	σ_{68}	NMAD	$\%of\ outliers$	skewness	$\%peak$	$\%one\ bin$	$\%in\ PDF$	$\%out\ PDF$
1	flat	0.05	92.3	99.8	-2.0E-4	0.0	0.024	0.018	0.017	0.12	-0.12	21.3	32.4	26.9	19.3
2	flat	0.1	87.3	99.7	7.7E-4	0.0	0.019	0.019	0.018	0.11	-0.2	18.0	30.0	44.0	7.0
3	flat	0.2	73.8	98.4	6.5E-4	0.0	0.024	0.024	0.023	0.14	-0.35	14.0	24.0	59.0	2.0
4	flat	0.3	61.4	95.4	-0.0045	0.0	0.03	0.03	0.03	0.17	-0.37	12.0	21.0	66.0	2.0
5	flat	0.4	51.7	90.8	-0.014	0.0	0.039	0.039	0.038	0.31	-0.24	10.0	18.0	69.0	2.0
6	poly	no	92.9	99.8	-0.0011	0.0	0.024	0.018	0.017	0.11	-0.16	22.1	30.3	13.5	34.1
7	indiv	no	92.4	99.7	-0.001	0.0	0.024	0.018	0.017	0.12	-0.21	22.0	15.0	31.0	31.0
8	bimod	0.05	91.8	99.8	-6.1E-4	0.0	0.024	0.019	0.017	0.11	-0.17	21.0	32.0	29.0	18.0
9	bimod	0.1	87.1	99.6	5.4E-4	0.0	0.025	0.019	0.018	0.11	-0.23	18.0	31.0	44.0	7.0
10	bimod	0.15	80.6	99.2	0.0012	0.0	0.026	0.021	0.02	0.12	-0.32	16.0	27.0	54.0	3.0
11	bimod	0.2	73.8	98.4	5.8E-4	0.0	0.022	0.023	0.023	0.13	-0.39	14.0	11.0	73.0	2.0

Table 4.3 Statistics of photo-z estimation performed by the MLPQNA, RF, KNN and Le-Phare models.

Estimator	MLPQNA	KNN	RF	Le-Phare
bias	0.0006	0.0029	0.0035	0.0009
σ	0.024	0.026	0.025	0.060
σ_{68}	0.018	0.020	0.015	0.035
NMAD	0.017	0.018	0.018	0.030
skewness	-0.17	0.330	0.015	-18.08
outliers	0.11%	0.15%	0.15%	0.69%

Table 4.4 Statistics of the *stacked* PDF obtained by Le-Phare and by the three empirical models MLPQNA, KNN and RF through METAPHOR.

Estimator	MLPQNA	KNN	RF	Le-Phare
$f_{0.05}$	91.7%	92.0%	92.1%	71.2%
$f_{0.05}$	99.8%	99.8%	99.7%	99.1%
$< \Delta z >$	-0.0006	-0.0018	-0.0016	0.0131

4.4.1 Comparison between METAPHOR and SED template fitting

Although there is a great difference in terms of performances between Le-Phare and MLPQNA, as it can be seen from Table 4.3 and the first three panels of Fig. 4.3, the results of the PDFs in terms of $f_{0.05}$ are comparable (see Table 4.4 and the right-hand panel in the lower row of Fig. 4.3).

But the greater efficiency of MLPQNA induces an improvement in the range within $f_{0.05}$, where we find ~ 92 per cent of the objects against the ~ 72 per cent for Le-Phare . Both individual and *stacked* PDFs are more symmetric in the case of empirical methods presented here than for Le-Phare .

This is particularly evident by observing the skewness (see Table 4.3), which is ~ 100 times greater for the SED template fitting method; this can also be seen by looking at panels in the lower row of Fig. 4.3.

4.4.2 METAPHOR as general provider of PDF for empirical models

The model KNN performs slightly worse than MLPQNA in terms of σ and outliers rate (Table 4.3), as it can be seen by looking at the first three panels of Fig. 4.4, while RF obtains results which pose this model between KNN and MLPQNA in terms of statistical performance, as visible from Table 4.3 and panels of Fig. 4.5. The higher accuracy of MLPQNA

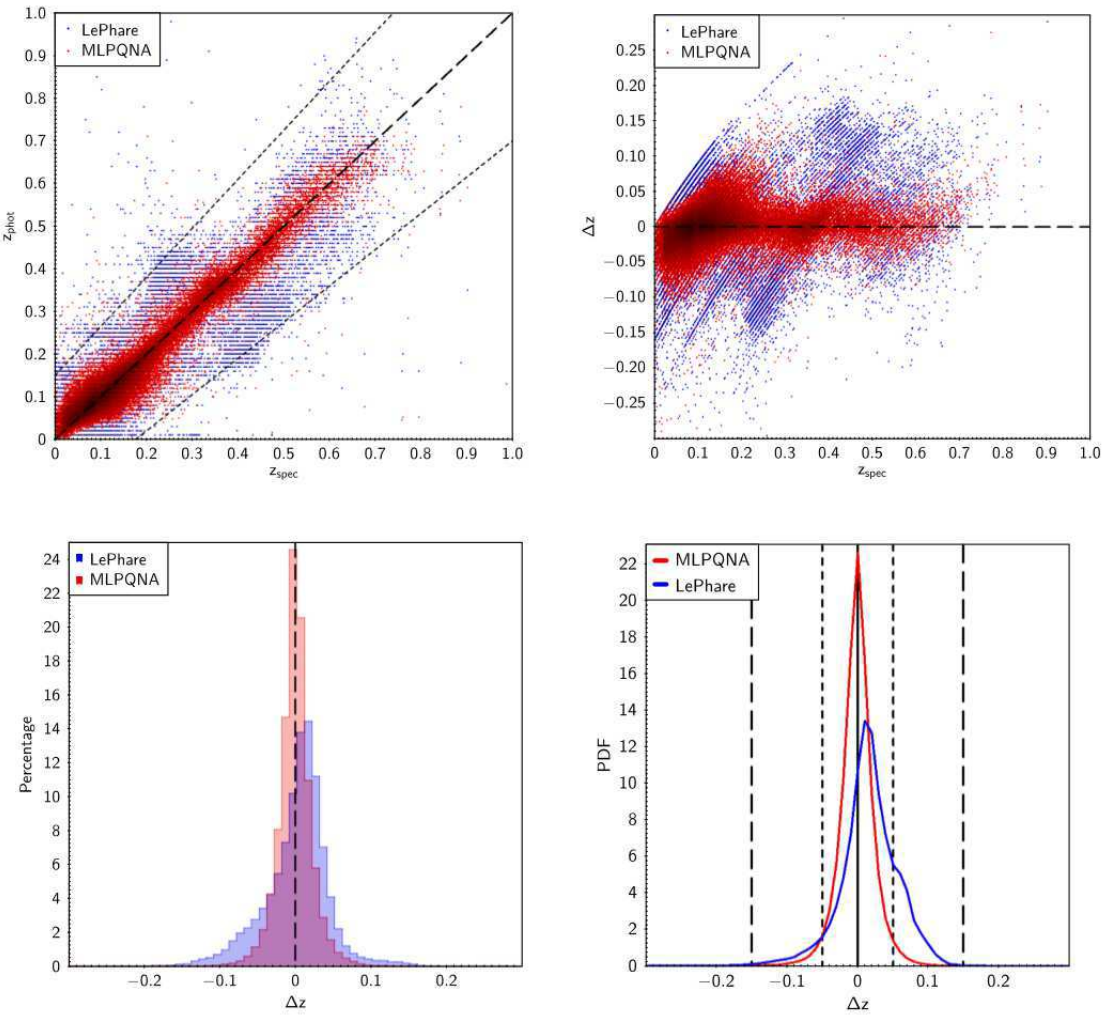

Fig. 4.3 Comparison between MLPQNA (red) and Le-Phare (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as a function of spectroscopic redshifts (z_{spec} versus z_{phot}); right-hand panel of upper row: scatter plot of residuals as a function of spectroscopic redshifts (z_{spec} versus Δz); left-hand panel of lower row: histograms of residuals (Δz); right-hand panel of lower row: stacked representation of residuals of the PDFs (the redshift binning is 0.01).

causes a better performance of PDFs in terms of $< \Delta z >$. However, also in the case of KNN and RF, METAPHOR is capable to produce reliable PDFs, comparable with those produced using MLPQNA (see Table 4.4 and right-hand panel in the lower row of Figs 4.4 and 4.5). This confirms the capability of METAPHOR to work efficiently with different empirical methods, regardless of their nature since even a very simple empirical model like KNN is able to produce high-quality PDFs. It also confirms that, provided a suitable large KB, all ML methods lead to similar accuracies.

The efficiency of the METAPHOR with the three empirical methods becomes clear by looking at Fig. 4.6, where we show the stacked PDF and the estimated photo-z distributions obtained by METAPHOR with each of the three models, superposed on the distribution of spectroscopic redshifts. The stacked distribution of PDFs, derived with the three empirical methods, results almost indistinguishable from the distribution of spectroscopic redshifts, with the exception of two regions: one in the peak of the distribution at around $z \simeq 0.1$ and the other at $z \simeq 0.4$. The first one can be understood in terms of a mild overfitting induced by the uneven distribution of objects in the training set. In fact around $z \simeq 0.1$ there is a large number of objects in the training set which induces a bias causing a small reduction in the generalization capability. The second one ($z \simeq 0.4$) can be explained by the fact that the break observed in the spectra of most galaxies at 4000 Å enters in the *r* band at this redshift thus inducing an edge effect in the parameter space, which leads our methods to generate predictions biased away from the edges. However, biases in color- space (averaging over/between degeneracies) specific to the SDSS filters clearly play a role as well.

By analyzing the relation between the spectroscopic redshift and the produced PDFs, we find that about ~ 22 per cent of z spec falls in the bin PDF peak, but we emphasize that a further ~ 33 per cent of spec-z falls one bin far from the peak (in our exercise, this means a distance of 0.01 from the peak). Finally, ~ 10 per cent of the spec-z falls outside the PDF. We analyzed the results in a tomographic way in order to verify whether there is a different behavior in different regions. This has been done by cutting the output in bins of photo-z (the best guess of our method) and deriving the whole statistics bin by bin. Results are shown in Table 4.5 and in Figs. 4.7 - 4.14.

In order to analyse the level of confidence of our PDFs, we performed a test using the credibility analysis presented in Wittman et al. (2016). The diagram shown in Fig. 4.15 indicates an overconfidence of our method. We notice, however, that this test is more suitable for continuous distribution functions and in our case is likely to introduce some artefacts in the low-credibility region.

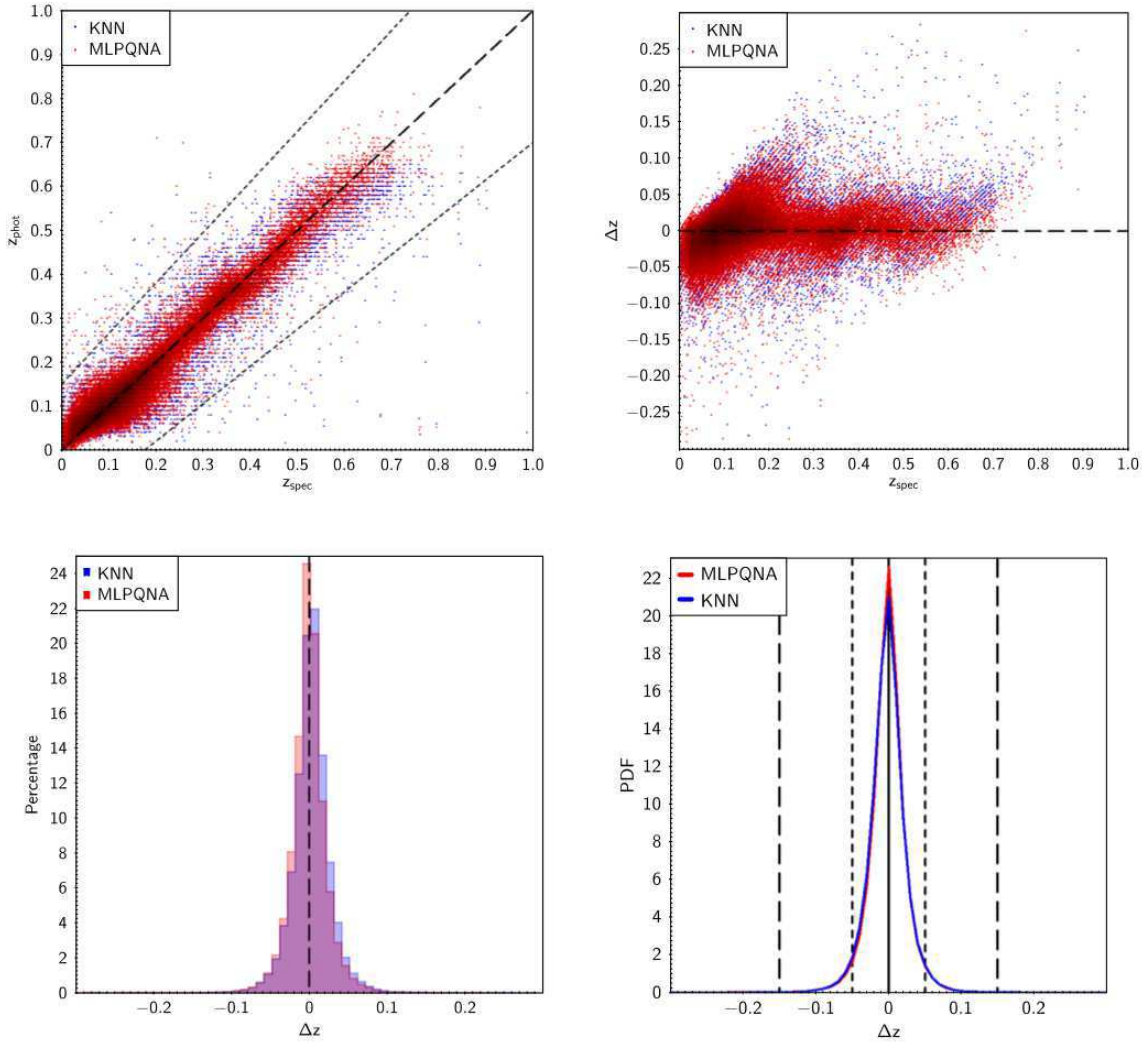

Fig. 4.4 Comparison between MLPQNA (red) and KNN (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as a function of spectroscopic redshifts (z_{spec} versus z_{phot}); right-hand panel of upper row: scatter plot of residuals as a function of spectroscopic redshifts (z_{spec} versus Δz); left-hand panel of lower row: histograms of residuals (Δz); right-hand panel of lower row: stacked representation of residuals of the PDFs (the redshift binning is 0.01).

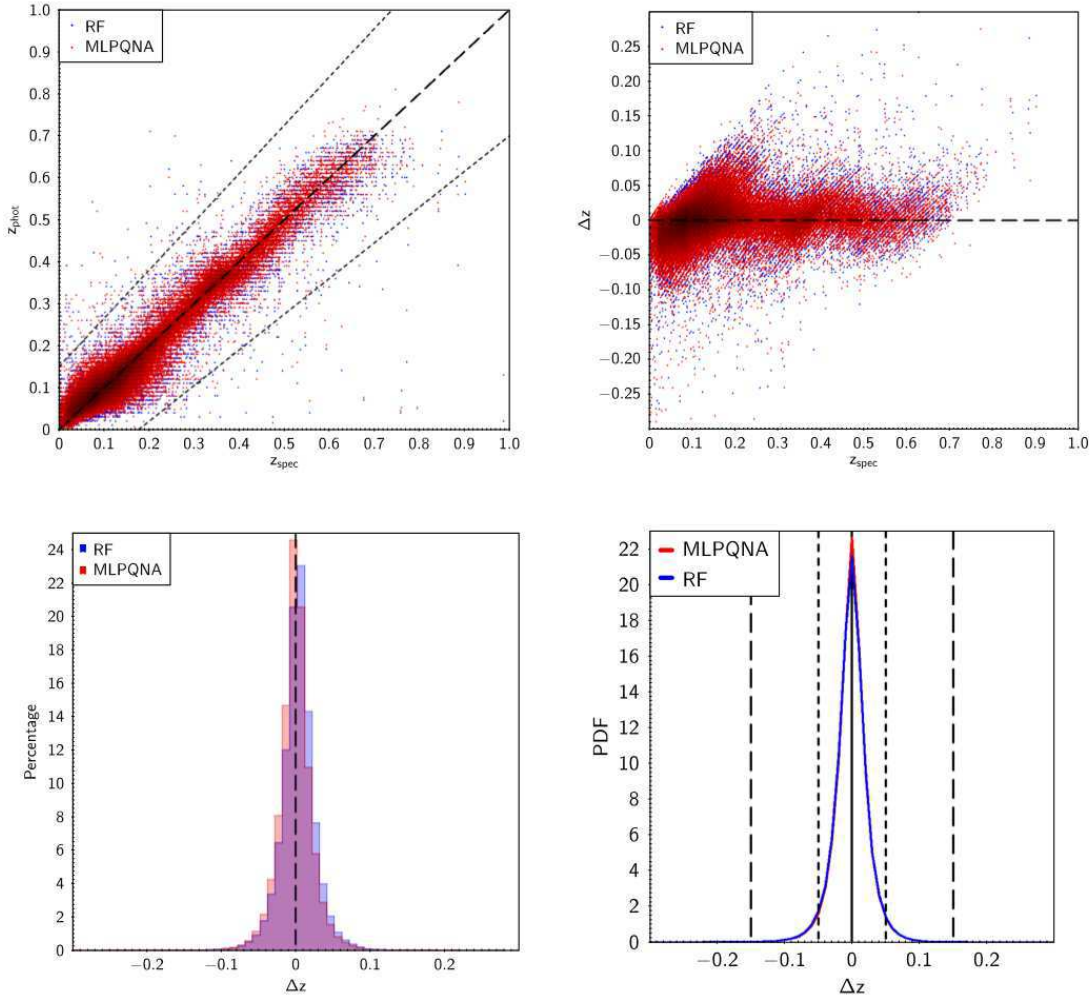

Fig. 4.5 Comparison between MLPQNA (red) and RF (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as a function of spectroscopic redshifts (z_{spec} versus z_{phot}); right-hand panel of upper row: scatter plot of residuals as a function of spectroscopic redshifts (z_{spec} versus Δz); left-hand panel of lower row: histograms of residuals (Δz); right-hand panel of lower row: stacked representation of residuals of the PDFs (the redshift binning is 0.01).

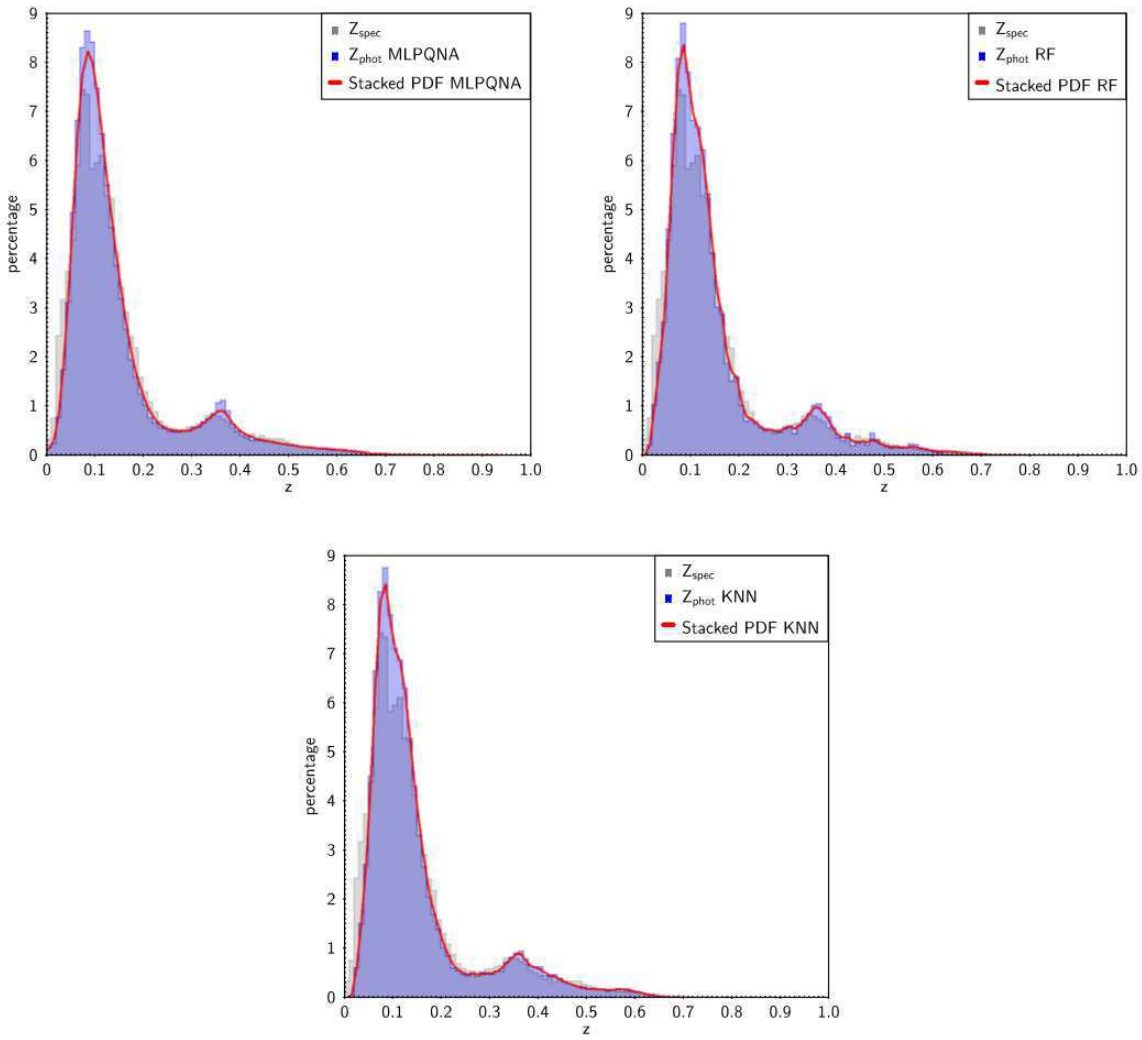

Fig. 4.6 Superposition of the stacked PDF (red) and estimated photo-z (blue) distributions obtained by METAPHOR with, respectively, MLPQNA, RF and KNN on the z_{spec} distribution (in grey) of the blind test set.

Table 4.5 Tomographic analysis of the photo-z estimation performed by the MLPQNA on the blind test set.

Estimator	Overall	$]0,0.1]$	$]0.1,0.2]$	$]0.2,0.3]$	$]0.3,0.4]$	$]0.4,0.5]$	$]0.5,0.6]$	$]0.6,1]$
bias	-0.0006	-0.0002	-0.0002	-0.0008	-0.0010	0.0017	-0.0028	-0.0054
σ	0.024	0.022	0.024	0.029	0.027	0.027	0.031	0.040
σ_{68}	0.018	0.018	0.019	0.018	0.019	0.019	0.021	0.028
NMAD	0.017	0.017	0.016	0.016	0.017	0.016	0.019	0.027
skewness	-0.17	1.39	0.048	-1.26	-1.75	-2.58	-1.56	-3.30
outliers	0.11%	0.04%	0.04%	0.60%	0.40%	0.40%	0.80%	0.60%
$f_{0.05}$	91.7%	93.4%	91.2%	89.9%	90.2%	87.2%	83.8%	76.8%
$f_{0.15}$	99.8%	99.9%	99.9%	99.2%	99.5%	99.5%	99.2%	98.9%
$<\Delta z>$	-0.0006	-0.001	-0.0001	0.0005	-0.0018	0.0025	-0.0015	-0.0015

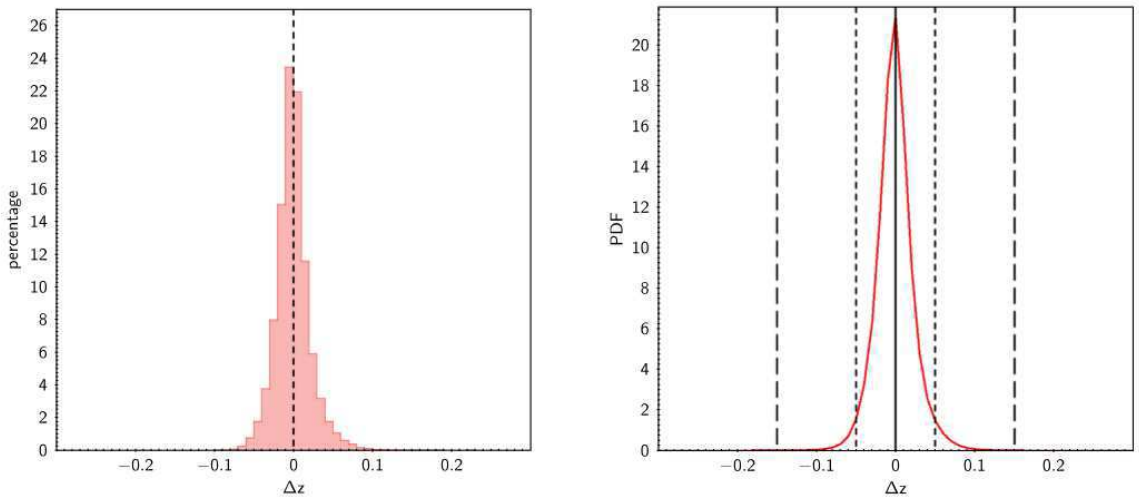

Fig. 4.7 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0, 0.1]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

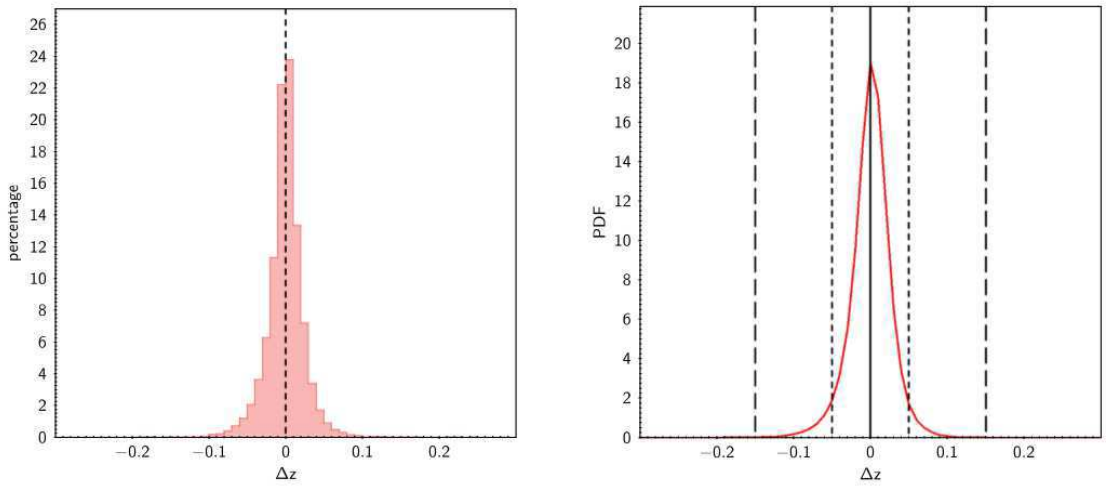

Fig. 4.8 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.1, 0.2]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

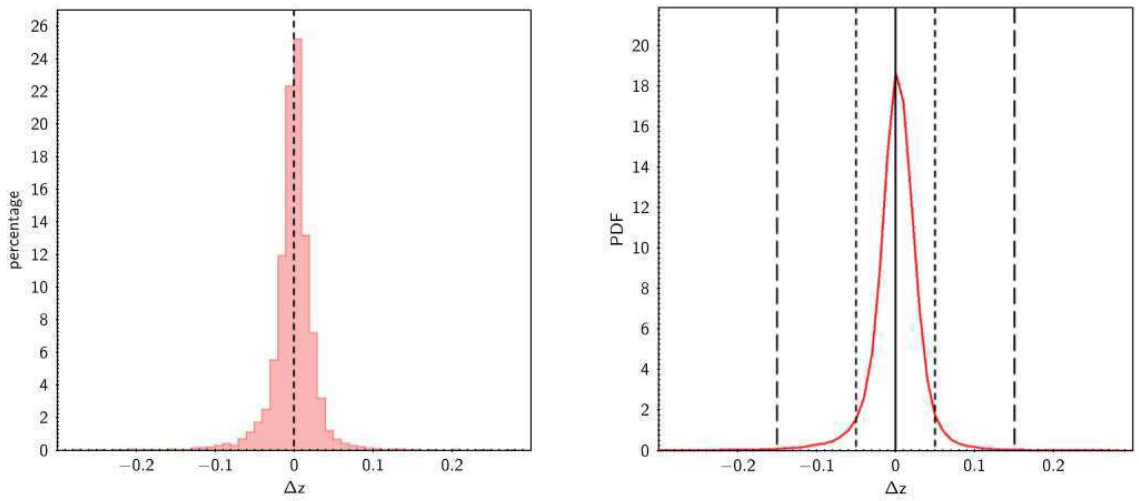

Fig. 4.9 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.2, 0.3]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

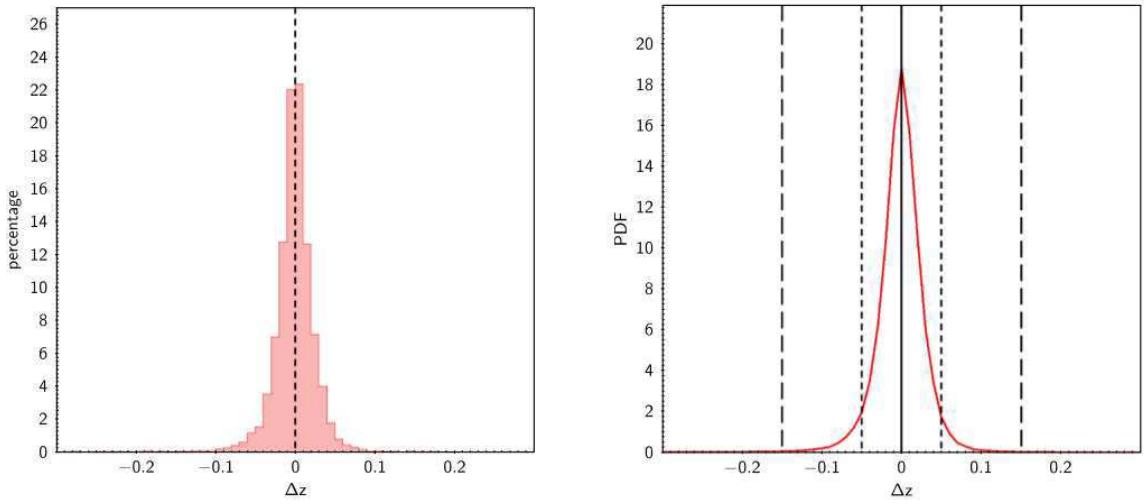

Fig. 4.10 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.3, 0.4]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

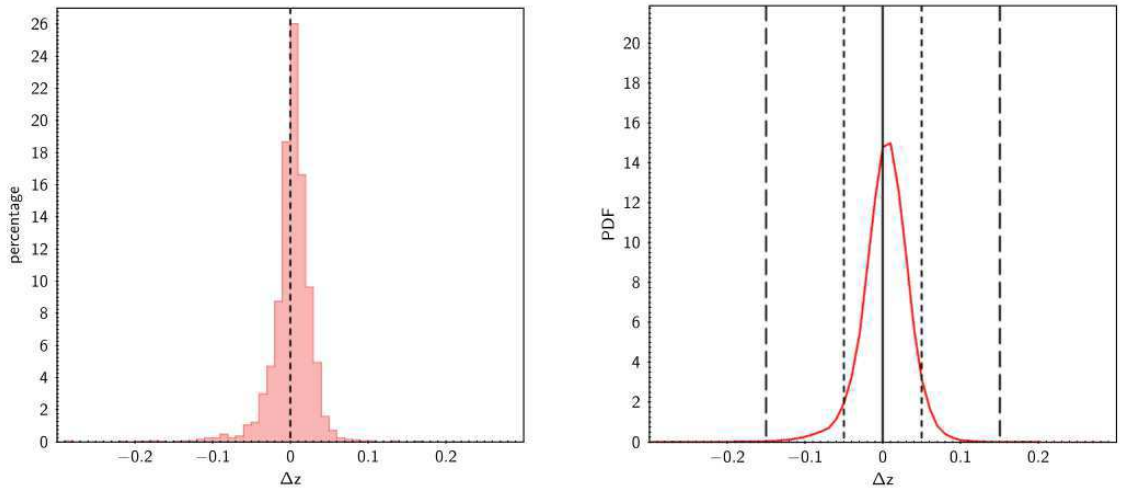

Fig. 4.11 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.4, 0.5]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

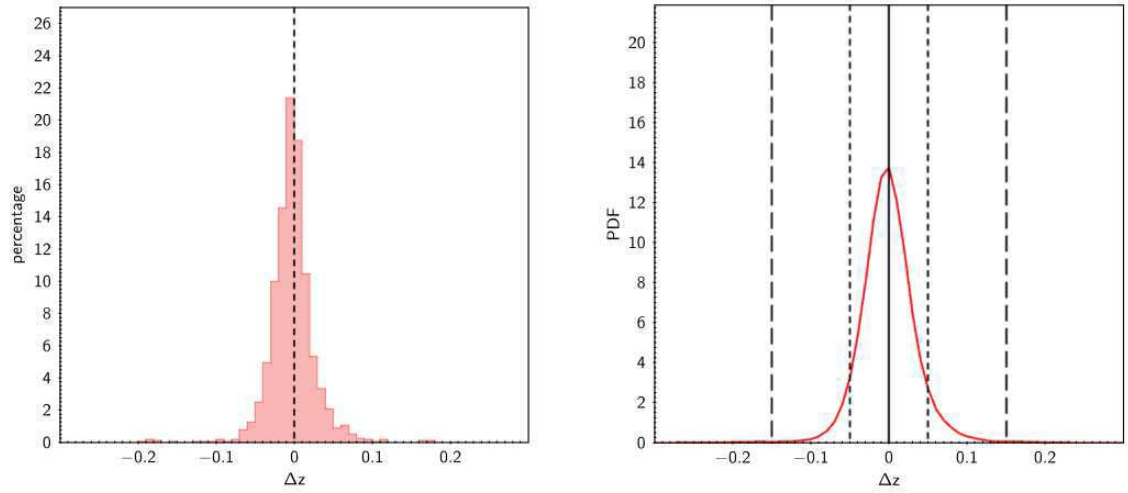

Fig. 4.12 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.5, 0.6]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

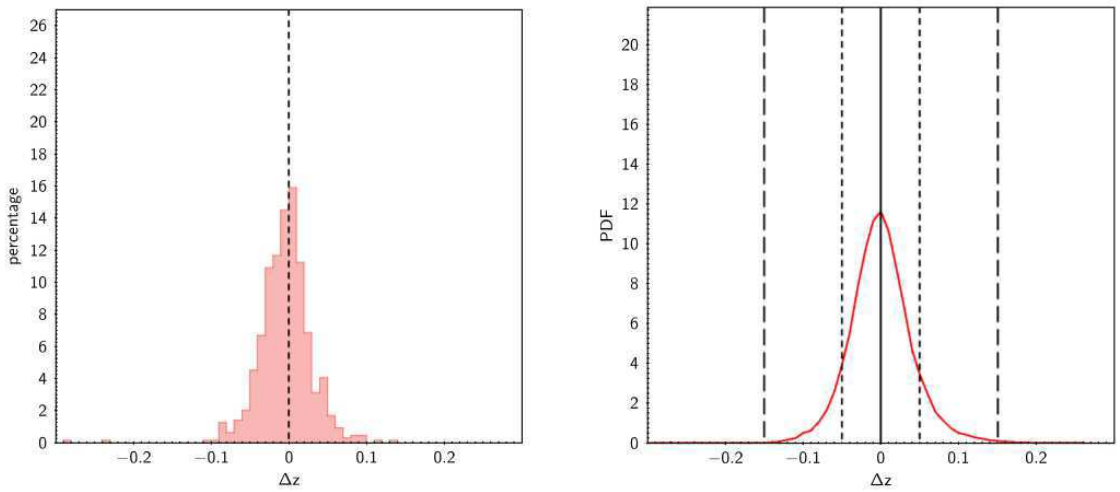

Fig. 4.13 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.6, 0.7]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

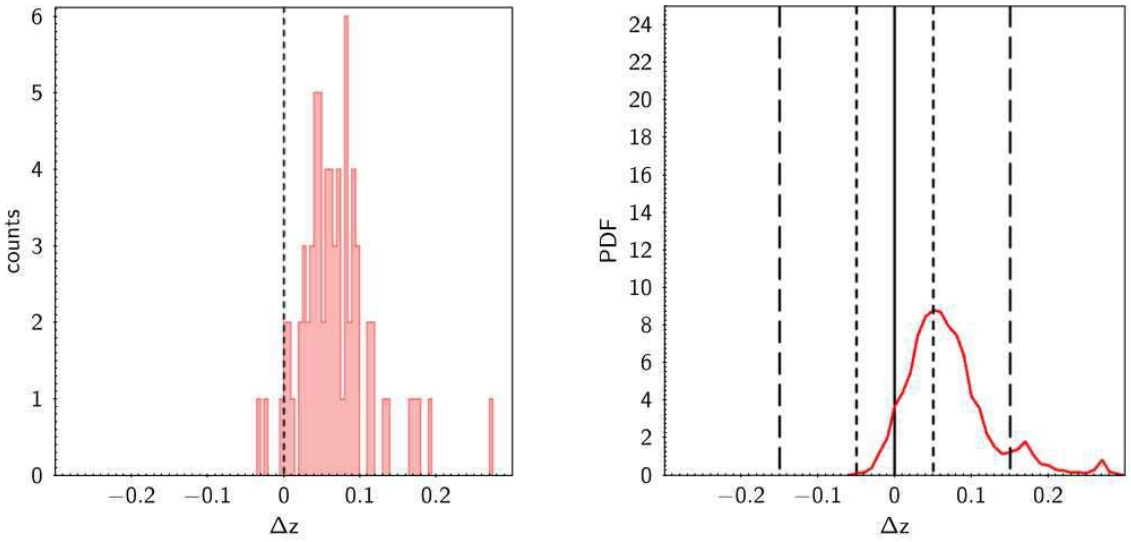

Fig. 4.14 Tomographic analysis of the PDF obtained by MLPQNA in the redshift bin $]0.7, 1]$. Upper panel: histogram of residuals (Δz); lower panel: stacked representation of residuals of the PDFs.

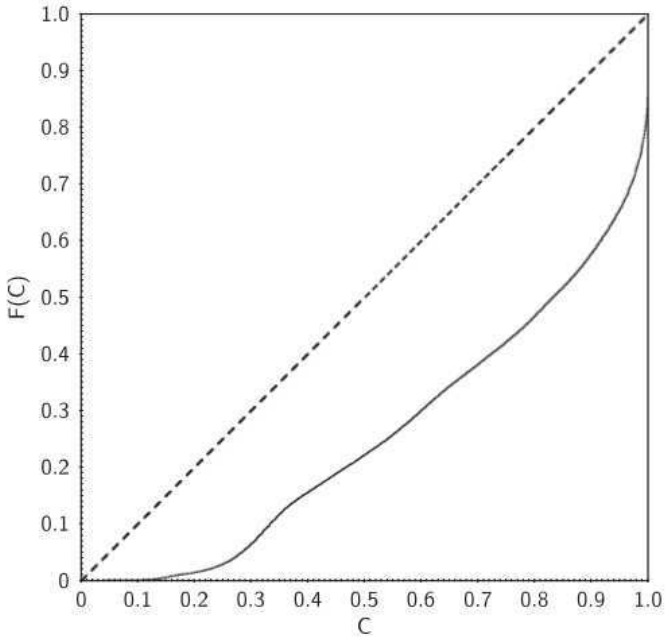

Fig. 4.15 Credibility analysis (Wittman et al. 2016) of the PDFs, as discussed in Sec.2.5.3 and shown in Fig. 2.8. The present figure shows the overconfidence of METAPHOR for the SDSS DR9 data.

4.5 Conclusions

We said in the first Chapter, that it is much harder to obtain a PDF for photo-z predicted by empirical methods, in particular for those based on ML techniques, due to their hidden way to find the flux–redshift correlations in the parameter space. From a theoretical point of view, the characterization of photo-z predicted by empirical methods should be based on the real capability to evaluate the distribution of the photometric errors, to identify the correlation between photometric and spectroscopic error contributions and to disentangle the photometric uncertainty contribution from that one internal to the method itself. This first comparative study of METAPHOR performances exemplifies these difficulties.

One of the most important goals of this analysis was to verify the universality of the procedure with respect to different interpolative models. For this reason, we experimented the METAPHOR processing flow on three alternative empirical methods. Besides the canonical choice of MLPQNA, a powerful neural network, the alternative models selected were RF and KNN. In particular, the choice of KNN has been mainly driven by its extreme simplicity with respect to the wide family of interpolation techniques. For this reason, we tested the METAPHOR strategy and the photo-z estimation models on a sample of the SDSS DR9 public galaxy catalog.

The presented photo-z estimation results and the statistical performance of the cumulative PDFs, achieved by MLPQNA, RF and KNN through the proposed procedure, demonstrate the validity and reliability of the METAPHOR strategy, despite its simplicity, as well as its general applicability to any other empirical method.

Chapter 5

METAPHOR for KiDS DR3

(extracted from Amaro et al., 2017, MNRAS, submitted)

5.1 Introduction

As we have described in Chapter 1, despite the consolidated high accuracy of photometric redshifts reachable by Machine Learning (ML) methods, the derivation of reliable and accurate probability density functions (PDFs) of the residual errors is still a challenging problem. First, because it is important to quantify the different sources of redshift estimate errors, which may arise from the estimation method itself, as well as from the photometric features of the available parameter space. Second, because the problem to define a robust statistical method, always able to quantify and qualify the PDF estimation validity, is still an open issue.

In this chapter, we present a comparison among three different methods: two ML techniques, METAPHOR (see Chapter 2) and ANNz2, plus a spectral energy distribution template fitting method, BPZ (already described these latter in Chapter 1). The data are the galaxies represented by the (Kilo-Degree Survey Data Release 3 (KiDS ESO DR3, hereafter) multi-band photometry and spectroscopy in the Galaxy And Mass Assembly (GAMA) field.

The statistical evaluation of both individual and *stacked* PDFs is done through quantitative and qualitative estimators (described in Sec. 2.5) with also a special *dummy* PDF, derived as benchmark to probe the capability to measure the quality of error estimation and invariance to error sources. We perform also a magnitude tomographic analysis, finding different trends. We conclude that, in order to assess a sufficient trade off between reliability and robustness of any photo-z PDF method, a combined set of statistical estimators is required.

5.2 A deeper analysis of the PDF meaning

The last decade has seen a proliferation of multi-band photometric galaxy surveys, either ongoing (cf. KiDS - Kilo-Degree Survey, de Jong et al. 2015, 2017 ; DES - Dark Energy Survey, Annis 2013) and planned (cf. those of the Large Synoptic Survey Telescope - LSST, Ivezic 2009, LSST Science Book 2009 and Euclid, Laureijs et al. 2014, Euclid 2011).
All these surveys require redshift estimates for hundreds of millions or billions of galaxies which cannot be observed spectroscopically and therefore must be obtained via multi-band photometry (photometric redshifts or photo-z). This is possible due to the existence of a (highly non linear) correlation between photometry and redshift, caused by the fact that the stretching introduced by the redshift induces the main spectrum features to move through the different filters of a photometric system (Baum 1962; Connolly et al. 1995). We have already introduced in the first Chapter that there are two classes of methods commonly used to derive photo-z: the template Spectral Energy Distribution (SED) fitting methods (e.g., Arnouts et al. 1999; Bolzonella et al. 2000; Ilbert et al. 2006; Tanaka 2015) and the empirical (or interpolative) methods (e.g., Brescia et al. 014b; Carrasco and Brunner 2013; Cavuoti et al. 2015; Masters et al. 2015; Sadeh et al. 2016; Tagliaferri et al. 2002; **?**), both characterized by their advantages and shortcomings.
SED methods are capable of deriving photo-z, the spectral type and the Probability Density Function (PDF) of the photo-z error distribution of each source all at once. However, they suffer from several cumbersome shortcomings, such as in particular, the potential mismatch between the templates used for the fitting and the properties of the selected sample of galaxies (Abdalla et al., 2011), color/redshift degeneracies and template incompleteness. Such issues are stronger at high redshift, where galaxies are fainter and photometric errors higher. Furthermore, for what concerns completeness, at high redshifts there are fewer or no empirical spectra available to build the template library.
Among empirical methods, those based on various Machine Learning (ML) algorithms are the most frequently used. We recall furthermore, that ML techniques are endowed with several advantages:

- *(i)* high accuracy of predicted photo-z within the limits imposed by the spectroscopic knowledge base;
- *(ii)* ability to easily incorporate external information in the parameter space (PS), such as surface brightness, angular sizes or galaxy profiles (Bilicki et al., 2017; Brescia et al., 2013; Cavuoti et al., 2012; Soo, 2017).

On the other hand, the weak capability to extrapolate information outside the parameter space defined by training data is one of main shortcomings of ML methods. Hence, for instance,

they cannot be used to estimate redshifts for objects fainter than those in the spectroscopic sample. Furthermore, the methods based on the supervised paradigm are applicable only if accurate photometry and spectroscopy are available for a sufficient (few thousands of objects at least) number of objects. See Hildebrandt et al. (2010), Abdalla et al. (2011) and Sánchez et al. (2014) for reviews about the photo-z estimation techniques.
Due to their intrinsic nature of self-adaptive learning models, the ML based methods do not naturally provide a PDF estimate of the predicted photo-z, unless special procedures are implemented. Over the last several years, particular attention has therefore been paid to develop techniques and procedures able to compute a full photo-z PDF for an astronomical source as well as for an entire galaxy sample (see Chapter 1 for a review of the methods available). The PDF contains more information with respect to the single redshift estimate, as also confirmed by the accuracy improvement of cosmological and weak lensing measurements (Mandelbaum et al., 2008; Viola et al., 2015).

As anticipated in the introduction, in this Chapter we perform a comparative analysis of the performances in terms of photo-z and associated PDFs among different methods. The data used for this analysis were extracted from the KiDS ESO DR3, described in detail in de Jong et al. (2017). In that paper, three different methods for photometric redshifts were used and the related photo-z catalogs made publicly available: two ML methods, respectively, METAPHOR (Machine-learning Estimation Tool for Accurate PHOtometric Redshifts, Cavuoti et al. 2017) and ANNz2 (Bilicki et al. 2017; Sadeh et al. 2016) and one template fitting method: Bayesian Photometric Redshifts (hereafter, BPZ, Benitez 2000).
For the purpose of the present Chapter, we also build a *dummy* PDF, intrinsically invariant to any kind of error source, useful to compare and assess the statistical estimators used to evaluate the reliability of PDFs.
A PDF should provide a robust estimate of the reliability of an individual redshift. In the context of the photo-z estimation the factors affecting such reliability are: photometric errors, intrinsic errors of the methods and statistical biases. In fact, under the hypothesis to reconstruct a perfect photometric redshift, the PDF would consist of a single number.
However, since the photo-z cannot be perfectly mapped to the true redshift, the related PDF represents the intrinsic uncertainties of the estimate. As anticipated, PDFs are useful to characterize photo-z estimates by providing more information with respect to the simple estimation of the error on the individual measurements.
Several works, over the past few years, have shown the capability of PDFs to increase the accuracy of the cosmological parameter measurements. For example, the work discussing the galaxy-galaxy lensing (Mandelbaum et al., 2008), has shown that most common statistics (bias, outliers rate, standard deviation etc.) are not sufficient to evaluate the precise accuracy

of photo-z. In particular the measurement of the critical mass surface density requires a full PDF estimation to remove any calibration bias effect.
In the following Sec. 5.3, we will describe briefly the photo-z and PDF catalogs obtained for the KiDS DR3 survey along with the catalog used to perform the deeper analysis on the PDF meaning, subject of this chapter.
Furthermore in Sec. 5.4, we will report only the useful information about the already described method, compared in this chapter, and the new method referred as *dummy* PDF.

5.3 The data

The sample of galaxies used to estimate photo-z and their individual and stacked PDFs was extracted from the third data release of the ESO Public Kilo-Degree Survey (KiDS-ESO-DR3, de Jong et al. 2017).

When completed, the KiDS survey will cover 1500 deg^2 (de Jong et al., 2017) (de Jong et al. 2017), distributed over two survey fields, in four broad-band filters (*u, g, r, i*). With respect to the previous data releases (de Jong et al., 2015), the DR3 does not only cover a larger area of the sky, but is also based on an improved photometric calibration and provides photometric redshifts along with shear catalogs and lensing-optimized image data. The total DR3 data set consists of 440 tiles for a total area covering approximately 450 deg^2, with respect to the 160 deg^2 of the previous releases (de Jong et al., 2015).

The single-band source lists of products for the DR3 include different aperture magnitudes, star/galaxy separation and mask regions. The reader is referred to the table A.1 in the appendix of de Jong et al. (2017) for the detailed content of this catalog, and to Fig. 2 and Table 3 of the same work for data quality details.

Along with the single-band, DR3 provides also an aperture-matched multi-band catalog for more than 48 million sources, including homogenized photometry based on Gaussian Aperture and PSF (hereafter GAaP) magnitudes (Kuijken, 2008). All the measurements (star/galaxy separation, source position, shape parameters) are based on the *r*-band images, due to their better quality (see Table A.2 of de Jong et al. 2017). The interested reader is referred to de Jong et al. 2017 for a deeper insight into the differences and additional procedures of DR3 with respect to the previous releases.
KiDS was designed primarily for Weak Lensing (WL) studies, in order to reconstruct the Large Scale Structure (LSS) of the Universe. Indeed, the first 148 tiles of the first two data releases produced their first scientific results on weak lensing for galaxies and groups of galaxies in the Gama And Mass Assembly (GAMA, Driver et al. 2011) fields (de Jong et al., 2015) , as the reader can find in Viola et al. (2015).

Now, within the public releases of DR3 products, the interested reader can find also the catalog KiDS-450, providing galaxy shape measurements for more than 14 million galaxies, useful in WL and cosmological parameter constraints studies (Hildebrandt et al., 2017), and used in the next and last Chapter 6 of this thesis to conduct an analysis of the shear measurements obtained with the METAPHOR PDFs.
For what METAPHOR produced data for the KiDS-DR3 survey, we produced a final photo-z catalog of $8,586,152$ objects, by including all data compliant with the magnitude ranges imposed by the KB used to train our model and specified in table 5.1. For convenience, the whole catalog was split into two categories of files, namely a single catalog file with the best predicted redshifts for the KiDS DR3 multi-band catalog, and a set of 440 files, one for each included survey tile, that contain the photo-z PDFs. The file formats are specified in Tables A.3 and A.4, in de Jong et al. (2017).
In order to perform the comparison through a common spectroscopic base, each of the three photo-z catalogs (obtained, respectively, by METAPHOR[1], ANNz2 and BPZ), has been cross-matched with spectroscopic information extracted from the second data release (DR2) of GAMA (Liske et al., 2015), containing spectroscopy for $\sim 70,000$ objects, overlapping with KiDS-North (composed by 77% from GAMA, 18% from SDSS/BOSS DR10, Ahn et al. 2014 and 5% from 2dFGRS, Colless et al. 2001).
For details about the training used for ANNz2, the interested reader is referred to de Jong et al. (2017) and Bilicki et al. (2017). In particular, the ANNz2 catalog released for DR3 does not contain individual PDFs, which have been provided for the analysis presented in this Chapter, and limited to the objects obtained by cross-matching the KiDS DR3 photometry with GAMA DR2 spectroscopy.

5.3.1 Data preparation

In the specific case of DR3, the KB used for METAPHOR is composed of 214 tiles of KiDS data cross-matched with SDSS-III data release 9 (Ahn et al., 2012) and GAMA data release 2 (Liske et al., 2015) spectroscopy.
The photometry is based on the ugri GAaP magnitudes, two aperture magnitudes, measured within circular apertures of 4″ 6″ diameter (referred in Table 5.1 as *MAG_APER*_20_*X* and *MAG_APER*_30_*X*), respectively, corrected for extinction and zeropoint offsets and related colors, for a total of 21 photometric parameters for each object.
The initial combination of the tiles leads to $120,047$ objects, after which the tails of the

[1] in (de Jong et al., 2017) it is referred as MLPQNA

Table 5.1 Brighter and fainter limits imposed to the magnitudes and used to build the parameter space for training and test experiments.

Input magnitudes	brighter cut limit	fainter cut limit
MAG_APER_20_U	16.84	28.55
MAG_APER_30_U	16.81	28.14
MAG_GAAP_U	16.85	28.81
MAG_APER_20_G	16.18	24.45
MAG_APER_30_G	15.86	24.59
MAG_GAAP_G	16.02	24.49
MAG_APER_20_R	15.28	23.24
MAG_APER_30_R	14.98	23.30
MAG_GAAP_R	15.15	23.29
MAG_APER_20_I	14.90	22.84
MAG_APER_30_I	14.56	23.07
MAG_GAAP_I	14.75	22.96

magnitude distributions and sources with missing magnitude measurements were removed. The derived lower and upper limits applied to exclude the tails of the distributions from all the DR3 tiles, are reported in Table 5.1. This was done in order to define the boundaries of the parameter space sampled by the training set.

In de Jong et al. (2017), we performed two experiments with two KBs in two different spectroscopy ranges: *(i)* $0.01 \leq z_{spec} \leq 1$ and *(ii)* $0.01 \leq z_{spec} \leq 3.5$. However, for the present analysis we focused on the training of the experiment *(ii)*. Then we randomly shuffled and split the relative data set, obtaining, respectively, $70,688$ training samples and $17,659$ test objects.

The random shuffle and split ensures the representativeness of the training set with respect to the parameter space covered by the test set as well as the homogeneity of the two data sets: a condition which is crucial to minimize systematics in the calibration of photometric redshifts.

As we have specified in Sec. 2.2.1, at the very base of the PDF estimation in METAPHOR there is the perturbation of the data photometry, based on a proper fitting function of the given flux errors in specifically defined bins of flux. Therefore, in the preparation phase, besides the inspection of the PS feature distributions, it is required also an inspection of the errors on such features provided with the DR3 catalogs.

5.4 The methods

In this Section, we report some useful information, already described in previous chapter for what regards METAPHOR and the SED fitters but useful to be clear.

5.4.1 METAPHOR

As said several times, in the context of ML techniques, the determination of individual PDFs is a challenging task. This is because we would like to determine a PDF by starting from several estimates of photo-z's, actually embedding the information on the photometric error uncertainties on those estimates. Therefore, we derived an analytical law to perturb the photometry by taking into account the magnitude errors provided by the catalogs.

Indeed, the procedure to determine individual source PDFs consists of a single training of the MLPQNA model and by perturbing the photometry of the given test set to obtain an arbitrary number N of test sets, characterized by a variable photometric noise contamination. The decision to perform a single train is mainly due to exclude the contribution of the method intrinsic error from the PDF calculation and due to the fact that the weight of the MLPQNA neural network are randomly initialized at each train.

With this aim, we use the perturbation law, described in Sec. 2.2.1:

$$\tilde{m}_{ij} = m_{ij} + \alpha_i F_{ij} u_{(\mu=0,\sigma=1)} \tag{5.1}$$

where j denotes the j-th object's magnitude and i the reference band; α_i is a multiplicative constant, chosen by the user (generally useful to take into account cases of heterogeneous photometry, i.e. derived from different surveys); the term $u_{(\mu=0,\sigma=1)}$ is a random value from a normal distribution; F_{ij} is the function used to perturb the magnitudes.
For the KiDS DR3 data, the selected perturbation function (F_{ij}) is the bimodal, composed by a constant function (in this case heuristically fixed to 0.03 in all bands while the constant α_i is chosen equal to 0.9 for all the bands) and a polynomial fitting of magnitude average errors on the binned bands. The role of the constant function is like a threshold under which the polynomial term is too low to provide a significant noise contribution to the perturbation (see Cavuoti et al. 2017 for further details). This to take into account the very low error average values for the brighter objects within the catalogs. This perturbation is applied to both GAaP and aperture magnitude types.
For the calculation of the individual PDFs, we submit the N + 1 test sets (i.e. N perturbed sets plus the original one) to the trained model, thus obtaining N + 1 estimates of photo-z. With these estimates we perform a binning in photo-z, thus calculating for each one the probability

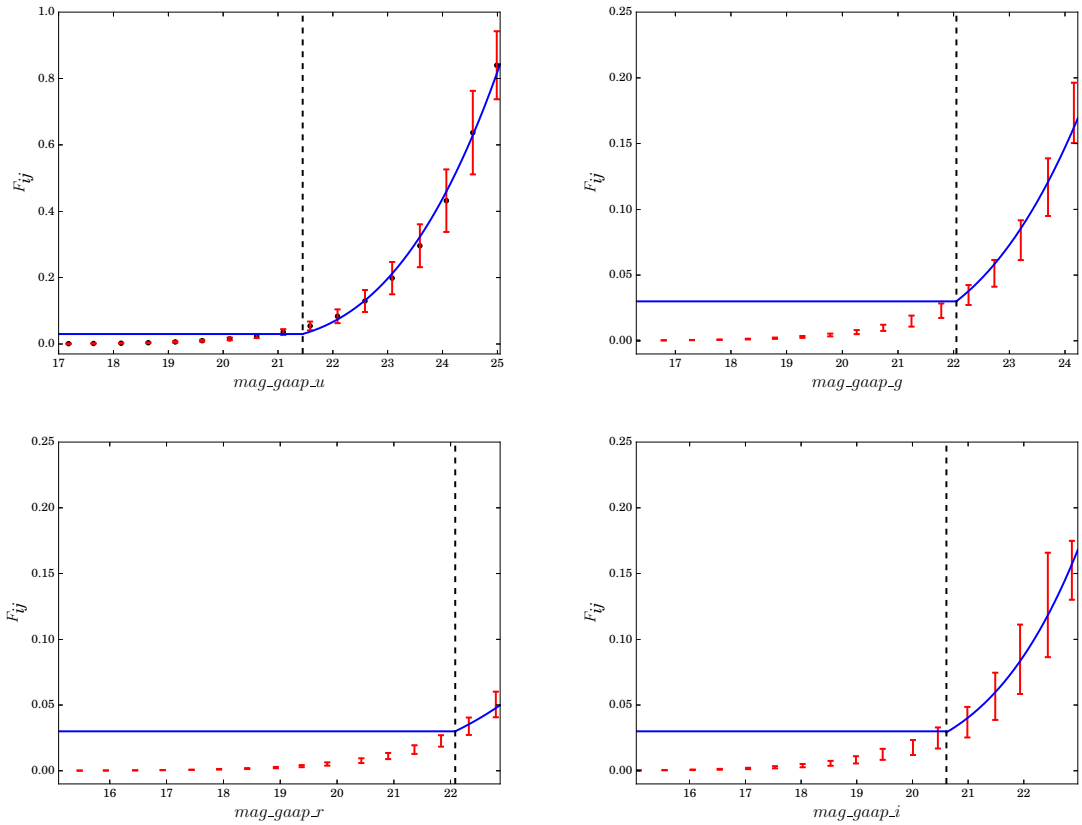

Fig. 5.1 Bimodal function F_{ij} in Eq. 5.1 for the GAaP magnitudes, composed by a flat perturbation for magnitudes lower than a selected threshold (black dashed lines) and a polynomial perturbation $p_i(m_{ij})$ for higher magnitude values (cf. Sec. 2.2.1). The switching thresholds between the two functions are, respectively, 21.45 in u band, 22.05 in g band, 22.08 in r and 20.61 in i band.

that a given photo-z value belongs to each bin. We selected a binning step of 0.01 for the described experiments and a value of N equal to $1,000$. The same binning step has been adopted by all methods presented in this work.

In Fig. 5.1 we can see the bimodal functions F_{ij} for the homogenized magnitudes *mag_gaap_x* (with x=*u,g,r,i*). These functions are composed by a constant part under a certain threshold in magnitude, and by polynomial fits of the average error above this threshold.

Concerning the photo-z production, the *best-estimate* photo-z values are not always corresponding to the given unperturbed catalog estimate of photo-z (hereafter photo-z_0), as calculated by MLPQNA. In particular it coincides with photo-z_0 if this measurement

falls into the interval (or *bin*) representing the *peak* (maximum) of the PDF; otherwise, it corresponds to the one closest to photo-z_0 and falling in the *peak* of the PDF.

5.4.2 ANNz2

ANNz2 (Sadeh et al., 2016) is a versatile ML package[2], designed primarily for deriving photo-z's. ANNz2 main method is based on artificial neural networks (ANNs) but it can work also with boosted decision and regression trees. For this Chapter, the co-author who furnished the ANNz2 PDFs, employed only ANNs. The author worked in the randomised regression mode of ANNz2, according to which several networks are randomly designed (the number of networks employed has been 100) and then trained on the spectroscopic data (the knowledge base described in Sec. deriving by the cross-match of GAMA DR2 and KIDS DR3 data). The whole sample of trained network has been used by the author in order to derive both photo-z's and associated PDFs. I report here the description of the author of ANNz2 PDFs about the PDF generation procedure (referring in any case the reader to Sadeh et al. (2016) and to the online documentation of ANNz2 for more details):

- Once the desired number of ANNs have been trained, then in the validation phase (called 'optimization' in ANNz2) each source from the validation set[3] is assigned to a distribution of photo-z solutions from the individual ANNs. These solutions are then ranked by their performance, and the top one is used to derive the individual photo-z estimate, `Z_BEST`, which we use in this work as the point photo-z prediction from ANNz2. In order to derive PDFs, the various ANNs are first folded with their respective single-value uncertainty estimates, derived via the k-Nearest Neighbour method (Oyaizu et al., 2008). A subset of ranked solutions is combined in different random ways to obtain a set of candidate PDFs. In order to select the final PDF, these candidates are compared using their Cumulative Distribution Functions (CDFs), defined as the integrated PDF for redshifts smaller than the reference value of the true redshift, z_{spec}:

$$\mathscr{C}(z_{spec}) = \int_{z_0}^{z_{spec}} p_{reg}(z)dz\,. \tag{5.2}$$

  The function $p_{reg}(z)$ is the differential PDF for a given redshift and z_0 is the lower bound of the PDF ($z_0 = 0$ in our case). The final PDF is chosen as the candidate for which the distribution of $\mathscr{C}$ is the closest to uniform (Bordoloi et al., 2010).
  ANNz2 can generate two types of PDFs, depending on how the $\mathscr{C}$ function is chosen.

[2] Available from https://github.com/IftachSadeh/ANNZ.

[3] We used the ANNz2 option to randomly split the spectroscopic calibration sample into training and validation sets in proportion 1:1.

In the first case, denoted PDF_0, the CDF is based on z_{spec}; in the second option, PDF_1, the results of the best ML solution are used as reference. In this work we use the PDF_1 option as we found it to perform generally better than the other one.

5.4.3 BPZ

The Benitez SED fitter method (BPZ), as any method of this kind, works exactly as the already described in Sec. 4.3.2, in the previous chapter. Here, we have only to add that all the details about the method to derive the BPZ photo-z estimation in KiDS are described in de Jong et al. (2017) along with the reference to the re-calibrated template set selected. While, for the priors used see Hildebrandt et al. (2012).

5.4.4 Dummy PDF

In order to have a benchmark tool useful to analyze and compare the statistical validity of previous methods, we set to zero the multiplicative constant parameter α_i of Eq. 5.1 for all bands to obtain a *dummy* perturbation law.

The relative *dummy* PDF obtained by METAPHOR is made by individual source PDFs, for which the one hundred per cent of the photo-z estimates (coincident with photo-z_0, i.e. the unperturbed estimate of photo-z) fall in the same redshift interval (by fixing the binning step at 0.01, as described in Sec. 2.4).
Main scope of this procedure is to assess the various statistical estimators used to evaluate an ensemble of PDFs. In fact, due to its intrinsic invariance to any kind of error source, it enables the possibility to compare PDF methods independently from the adopted statistical estimator.

5.4.5 Statistical estimators

All the statistical estimators used in the context of this PDF performance analysis have been explained in Sec. 2.5.

5.5 Comparison among methods

A preliminary comparison among the three methods METAPHOR, ANNz2 and BPZ, only in terms of photo-z prediction performance, has been already given in de Jong et al. (2017). That comparison was based on statistics applied to the residuals defined by the Eq. 2.7,

Table 5.2 Statistics of photo-z estimation performed by MLPQNA (photo-z estimation engine of METAPHOR), ANNz2, BPZ, on the GAMA field: respectively, the bias, the sigma, the Normalized Median Absolute Deviation, the fraction of outliers outside the 0.15 range, kurtosis and skewness.

Estimator	MLPQNA	ANNz2	BPZ
bias	-0.004	-0.008	-0.020
σ	0.065	0.078	0.048
NMAD	0.022	0.018	0.028
outliers	0.97%	1.60%	1.13%
Kurtosis	774.1	335.9	52.5
Skewness	-21.8	-15.8	-2.91

reported in Table 8 and Fig.11 of de Jong et al. (2017). In the upper panel of that figure, the plots of photo-z vs GAMA-DR2 spectroscopy and residuals vs *r*-magnitude were shown for the three methods.

More recently, in Bilicki et al. (2017) a comparison among the three methods has also been presented on KiDS DR3 data, more in terms of photo-z estimation quality at the full spectroscopic depth available, confirming the better behavior of ML methods at bright end of KiDS ($z < 0.5$) as well as comparable quality of ML methods and BPZ at higher redshift ($z \sim 1$).

The content of the present chapter is mainly focused on the photo-z PDF comparison among the three methods using the data set obtained by cross-matching KiDS DR3 photometry with GAMA DR2 spectroscopy, which amounts to $63,749$ samples. The reason why the number of objects does not coincide with that reported in Sec. 5.3.1 is that the spectroscopy strictly used for the comparison is based only on GAMA DR2 data.

The statistical comparison among the three methods on the dataset obtained by cross-matching KiDS-DR3 and GAMA data, is summarized in Table 5.2. It shows a better performance in terms of bias and fraction of outliers for METAPHOR, while BPZ and ANNz2 obtain, respectively, a lower σ and *NMAD* of the residual errors.

In figures 5.2 and 5.3 we show the comparison on the GAMA field between METAPHOR and respectively, BPZ and ANNz2, also in terms of graphical distributions of predicted photo-z and stacked PDF residuals.

From Fig. 5.2 it is apparent that the correlation between z_{phot} and z_{spec} is tighter for METAPHOR than for BPZ and that the residuals have a more symmetric distribution, as well as a more peaked shape. Fig. 5.3 shows slightly higher symmetry and peak for the ANNz2 residuals distribution.

In table 5.3 we report the fraction of residuals in the two ranges $[-0.05, 0.05]$ and

Table 5.3 Statistics of the photo-z error stacked PDFs for METAPHOR, ANNz2, BPZ and *dummy* obtained by METAPHOR, for the sources cross-matched between KiDS DR3 photometry and GAMA spectroscopy.

Estimator	METAPHOR	ANNz2	BPZ	dummy
$f_{0.05}$	65.6%	76.9%	46.9%	93.1%
$f_{0.15}$	91.0%	97.7%	92.6%	99.0%
$< \Delta z >$	-0.057	0.009	-0.038	-0.006

Table 5.4 *zspecClass* fractions for METAPHOR, ANNz2 and BPZ on the GAMA field.

zspecClass	METAPHOR		ANNz2		BPZ	
0	9042	(14.2%)	12426	(19.5%)	4889	(7.7%)
1	16758	(26.3%)	19040	(29.9%)	9650	(15.1%)
2	37233	(58.4%)	31927	(50.1%)	49170	(77.1%)
3	200	(0.3%)	8	(0.01%)	0	(0%)
4	516	(0.8%)	324	(0.5%)	31	(0.05%)

$[-0.15, 0.15]$ and the average of residuals for all the methods probed. Last column shows such statistics also for the *dummy* PDF. Table 5.4 summarizes the distribution of fractions of samples among the five categories of individual PDFs, referred to their spectroscopic redshift position with respect to the PDF.

From table 5.3 it appears evident that in terms of stacked PDF, ANNz2 performs quantitatively better than other two methods. This behavior is also supported by looking at table 5.4, where ANNz2 has a percentage of 49.4% of samples falling within one bin from the PDF peak (the sum of fractions for *zspecClass* 0 and 1) against, respectively, the 40.5% and 22.8%, of the other two methods. However, by introducing the statistics for the *dummy* PDF in the analysis, there is a clear improvement of all stacked PDF estimators. By construction of the *dummy* PDF and due to the fact that such PDFs are *peaked* in a single value, it is not worth to report the statistics regarding the *zspecClass* estimator (cf. Sec. 2.5.2), since it is expected that the most part of the spec-z of the GAMA sources fall outside the PDF.

Therefore, from the statistical results of table 5.3, the *dummy* PDF, derived from METAPHOR, obtains the best quality estimation. This demonstrates that the statistical estimators adopted for the stacked PDF show low robustness in terms of quality assessment of photo-z error evaluation and there is need for deeper understanding of the real meaning of a PDF in the context of photo-z quality estimation and a careful investigation about the statistical evaluation criteria.

In Fig. 5.4, we superimpose the stacked distribution of PDFs, derived by the three methods plus the *dummy* PDF, on the photometric and spectroscopic redshift distributions. By observ-

ing the stacked trend of the *dummy* PDF method, as expected, it results able to reproduce the spectroscopic distribution, since by construction it does not take into account any redshift error contribution (in particular that one arising from the photometric uncertainties through the perturbation law in Eq. 5.1).
Very close to the spectroscopic redshift distribution is also the stacked PDF of ANNz2, while BPZ and METAPHOR, although still able to follow the spectroscopic distribution, differ from the first two methods. Nevertheless, METAPHOR and ANNz2 show a better similarity between their spectroscopic and photometric redshift distributions.
We proceed further by introducing two graphical estimators, respectively, the credibility analysis on the cumulative PDFs and the PIT, described in Sec. 2.5.3. The Fig. 5.5 and Fig. 5.6 show these two respective diagrams for the three methods and the *dummy* PDF. The credibility analysis trend of METAPHOR (top right panel of Fig. 5.5) reveals an evident better degree of credibility than ANNz2 and BPZ (respectively, bottom and top left panels of Fig. 5.5), which are characterized by a higher degree of *underconfidence*. However, the credibility diagram of the *dummy* PDF (bottom right panel of Fig. 5.5) puts in evidence the inability to evaluate the credibility of a photo-z error PDF in an objective way, since in the case of the *dummy* PDF the Eq. 2.8 is identically 1 for each galaxy of the data set. In other words, according to the construction of the HPDCI for the credibility analysis (cf. Sec. 2.5.3), the *dummy* PDFs method shows that the 100% of photo-z_0's fall in the 100% of the HPDCI, thus resulting totally *overconfident*.

The statistical evaluation of the three methods and the *dummy* PDF based on the PIT diagram is shown in Fig. 5.6. In the case of METAPHOR (top right panel) we can observe a good degree of uniformity (flatness of the distribution F_t around the threshold at 0.05, obtained as the inverse of number of classes, fixed to 20 in this case, i.e. *c=1/20*), in the range between 0.05 and 0.6 of the random variable and for 11 up to 20 classes in total. Among the three methods the worst performance occurs for BPZ, with only two uniform classes up to 20, while ANNz2 appears slightly better than METAPHOR, with 13 uniform PIT classes up to 20. However, the PIT histogram for *dummy* PDFs shows a totally degraded *underdispersive* behavior of the photo-z's distribution (bottom right panel of Fig. 5.6), and this result is in clear contrast with the previous statistics.

Another interesting comparison concerns the analysis of a specific quality parameter provided by BPZ in KiDS DR3 photo-z catalogs. For each source there is the so-called ODDS parameter (Coe et al., 2006), ranging in the interval]0,1[. It measures the increasing uni-modality of the redshift PDF: the closer the ODDS parameter is to 1, the higher is the reliability of the photo-z estimate. Therefore, we decided to evaluate the possible improvements of the cumulative statistics given in table 5.3, by selecting objects filtered through a

Table 5.5 Statistics of the stacked PDF for METAPHOR and BPZ for the objects characterized by values of the ODDS parameter higher (h) than two chosen thresholds, respectively, 0.8 and 0.9.

Estimator	METAPHOR		BPZ	
	h 0.8	h 0.9	h 0.8	h 0.9
$f_{0.05}$	65.7%	67.3%	47.0%	47.5%
$f_{0.15}$	91.1%	91.7%	92.8%	93.3%
$< \Delta z >$	-0.056	-0.053	-0.037	-0.037

fixed threshold value imposed by the BPZ ODDS parameter. The results of the cumulative PDF performance are reported in table 5.5 for BPZ and METAPHOR. The number of objects on the entire sample are: 263 for $ODDS < 0.8$ (0.4%) and 63,486 for $ODDS > 0.8$ (99.6%); 5,337 for $ODDS < 0.9$ (8.4%) and finally 58,372 for $ODDS > 0.9$ (91.6%).

By comparing tables 5.5 and 5.3, the improvement of the PDF quality for both METAPHOR and BPZ appears not particularly relevant, also by considering that the thresholding imposed by the ODDS parameter implies the loss of a considerable amount of objects from the dataset.

Finally, in order to analyze the stacked PDFs obtained by the three estimation methods in different ranges of magnitude, we performed a binning of the magnitude *mag_gaap_r* in the range [16.0, 21.0] with a step Δmag=0.5, resulting in a tomography of 10 bins in total. The range has been chosen in order to ensure a minimum quantity of objects per bin to calculate the statistics.

The results in terms of the fraction of residuals and the total average for the stacked PDFs (cf. Sec. 2.5.3), are reported in Table 5.6, while the fraction of residuals $f_{0.05}$ is shown as a function of magnitude in Fig. 5.7.

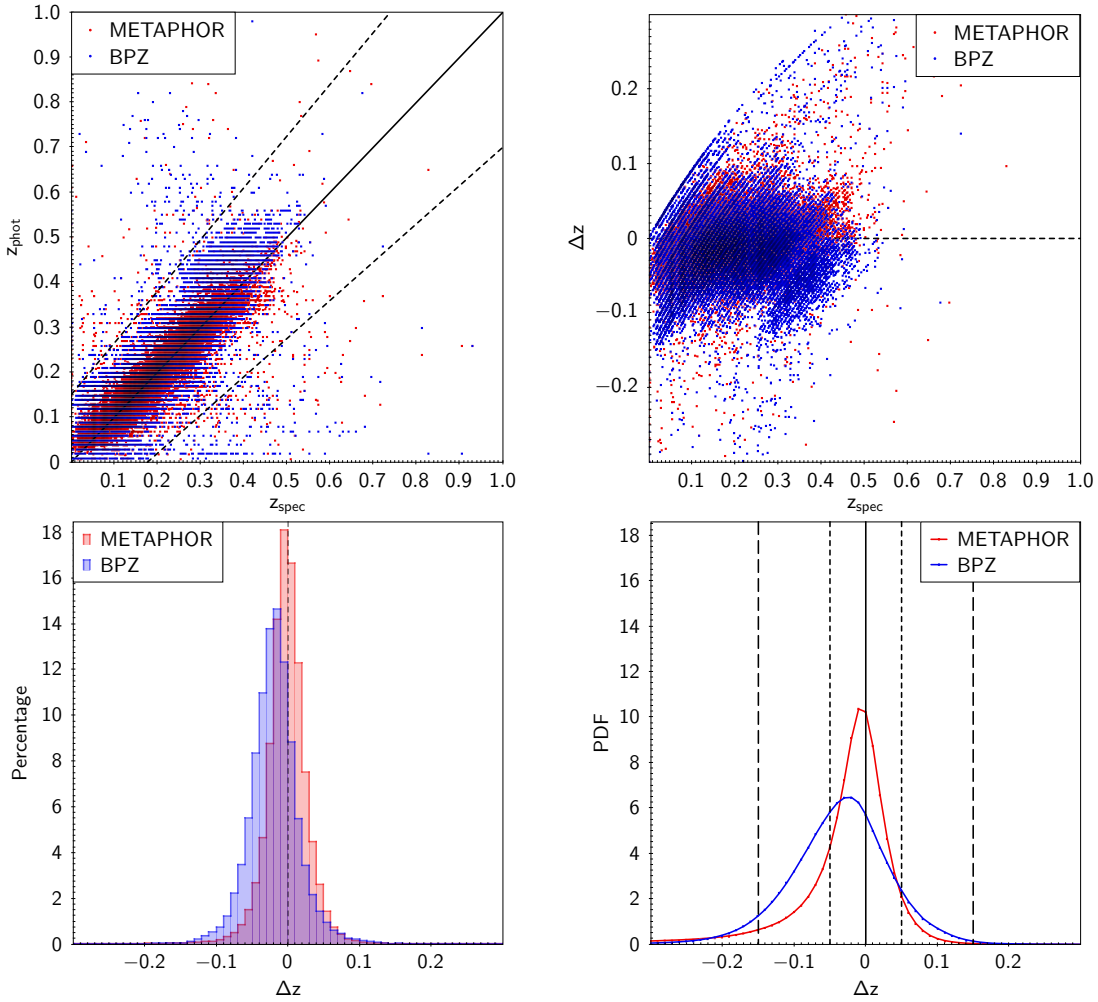

Fig. 5.2 Comparison between METAPHOR (red) and BPZ (blue). Left-hand panel of upper row: scatter plot of photometric redshifts as function of spectroscopic redshifts (z_{spec} vs z_{phot}); right-hand panel of upper row: scatter plot of the residuals as function of the spectroscopic redshifts (z_{spec} vs Δz); left-hand panel of the lower row: histograms of residuals (Δz); right-hand panel of lower row: *stacked* representation of the residuals of PDFs (with redshift bin equal to 0.01).

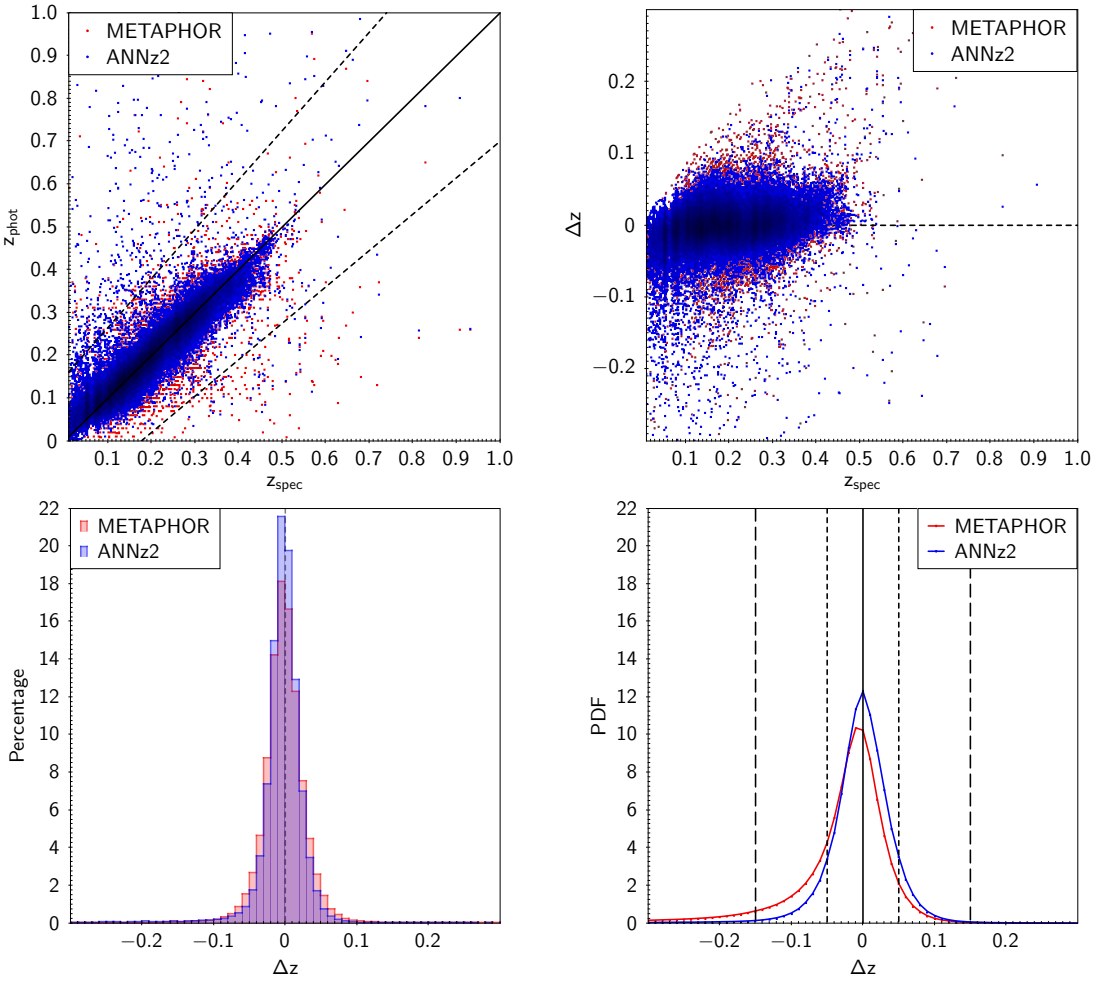

Fig. 5.3 Comparison between METAPHOR (red) and ANNz2 (blue). Left-hand panel of upper row: scatter plot of photometric redhifts as function of spectroscopic redshifts (z_{spec} vs z_{phot}); right-hand panel of upper row: scatter plot of the residuals as function of the spectroscopic redshifts (z_{spec} vs Δz); left-hand panel of the lower row: histograms of residuals (Δz); right-hand panel of lower row: *stacked* representation of the residuals of PDFs (with redshifts bin equal to 0.01).

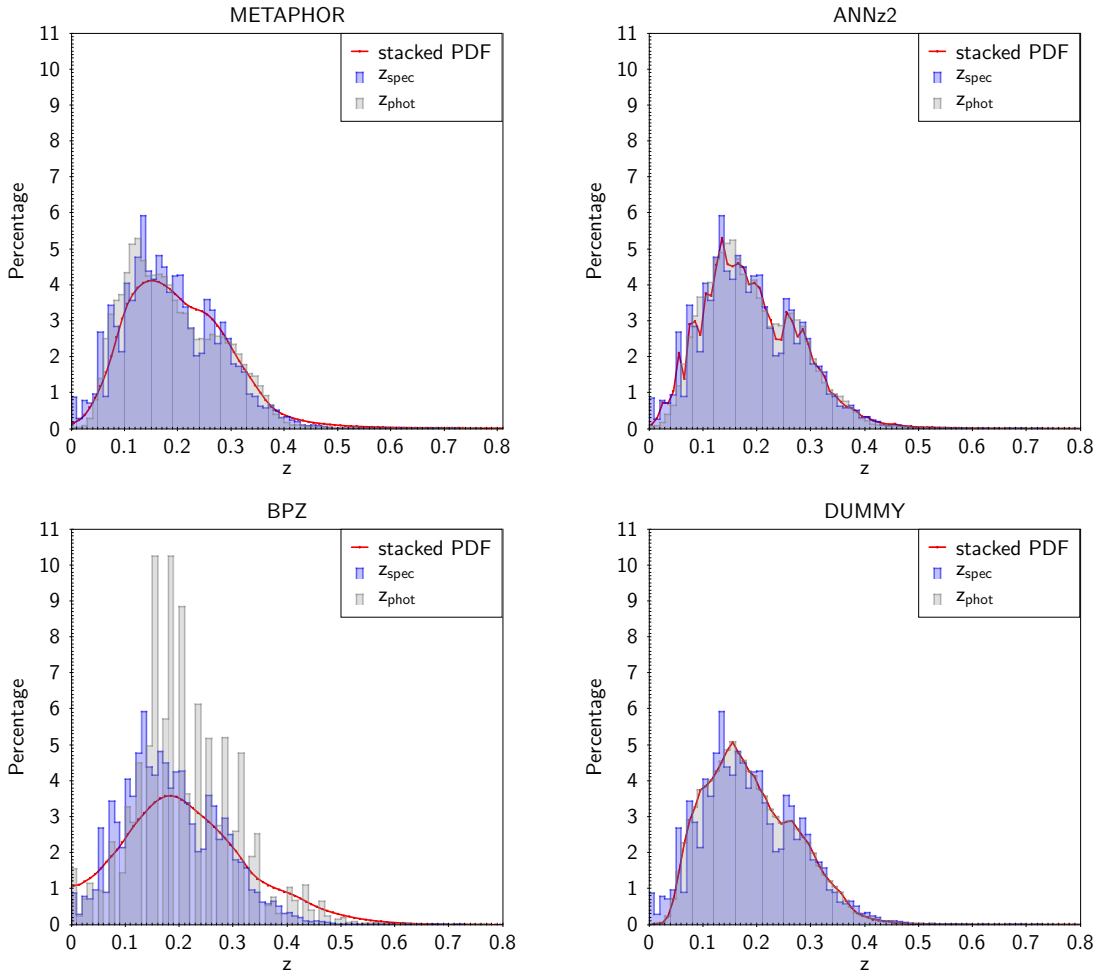

Fig. 5.4 Superposition of the stacked PDF (red) and estimated photo-z (gray) distributions obtained by METAPHOR, ANNz2, BPZ and for the *dummy* (in this last case the photo-z distribution corresponds to that of the photo-z_0 estimates, Sec. 5.4.4) to the z-spec distribution (in blue) of the GAMA field.

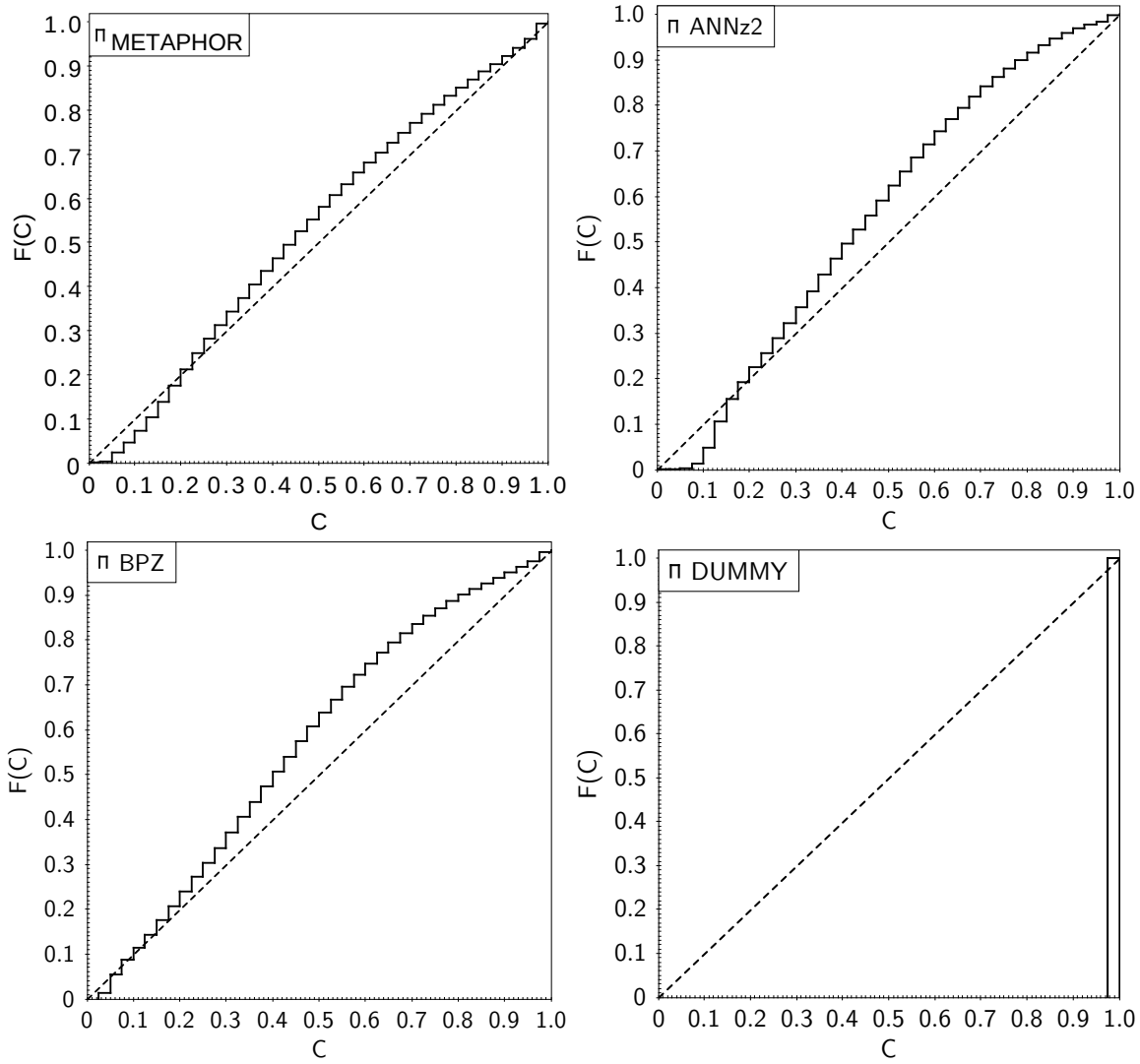

Fig. 5.5 Credibility analysis (cf. Sec. 2.6) obtained for METAPHOR, ANNz2, BPZ and the *dummy* PDF, calculated by METAPHOR.

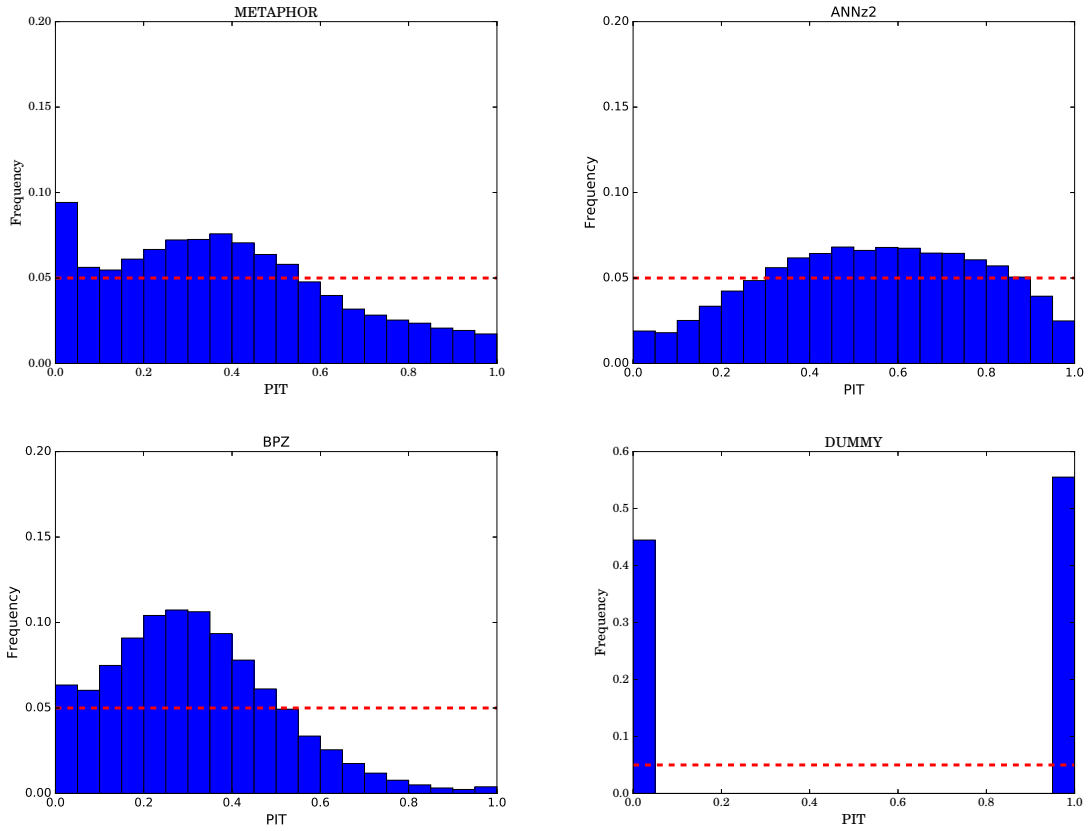

Fig. 5.6 Probability Integral Transform (PIT) obtained for METAPHOR (top left panel), ANNz2 (top right panel), BPZ (bottom left panel), and for the *dummy* PDF, calculated by METAPHOR (bottom right panel).

Table 5.6 Tomographic analysis of the stacked PDFs for METAPHOR, ANNz2, BPZ and *dummy* PDF calculated by METAPHOR, respectively, in ten bins of the homogenized magnitude *mag_gaap_r*.

			METAPHOR			ANNz2			BPZ			dummy		
Bin	Range	Number	$f_{0.05}$	$f_{0.15}$	$<\Delta z>$	$f_{0.05}$	$f_{0.15}$	$<\Delta z>$	$f_{0.05}$	$f_{0.15}$	$<\Delta z>$	$f_{0.05}$	$f_{0.15}$	$<\Delta z>$
1	$]16.0,16.5]$	122	16.3%	37.2%	-0.330	80.9%	99.5%	-0.016	26.5%	87.0%	-0.080	97.5%	100%	-0.015
2	$]16.5,17.0]$	290	23.9%	49.0%	-0.249	81.7%	99.2%	-0.015	28.5%	86.7%	-0.080	97.9%	99.3%	-0.009
3	$]17.0,17.5]$	858	34.2%	62.4%	-0.185	82.0%	98.4%	-0.016	36.4%	89.7%	-0.068	95.1%	98.7%	-0.006
4	$]17.5,18.0]$	1,873	48.0%	75.7%	-0.132	81.6%	97.4%	-0.017	41.0%	90.7%	-0.060	94.2%	97.8%	-0.010
5	$]18.0,18.5]$	4,427	59.0%	84.6%	-0.086	82.2%	98.2%	-0.011	45.4%	92.5%	-0.050	95.3%	98.7%	-0.006
6	$]18.5,19.0]$	8,230	64.9%	89.4%	-0.067	81.1%	98.0%	-0.008	47.6%	93.1%	-0.043	94.3%	98.8%	-0.008
7	$]19.0,19.5]$	15,388	68.9%	92.6%	-0.051	79.2%	97.9%	-0.008	48.5%	93.2%	-0.037	93.7%	98.9%	-0.007
8	$]19.5,20.0]$	22,952	68.5%	93.8%	-0.043	75.9%	98.0%	-0.006	47.8%	92.9%	-0.033	93.4%	99.2%	-0.003
9	$]20.0,20.5]$	9,178	65.8%	94.2%	-0.040	61.4%	97.0%	-0.010	45.4%	91.6%	-0.033	89.9%	98.9%	-0.007
10	$]20.5,21.0]$	367	55.5%	88.4%	-0.061	44.5%	80.4%	-0.104	43.1%	88.9%	-0.033	74.6%	94.0%	-0.025

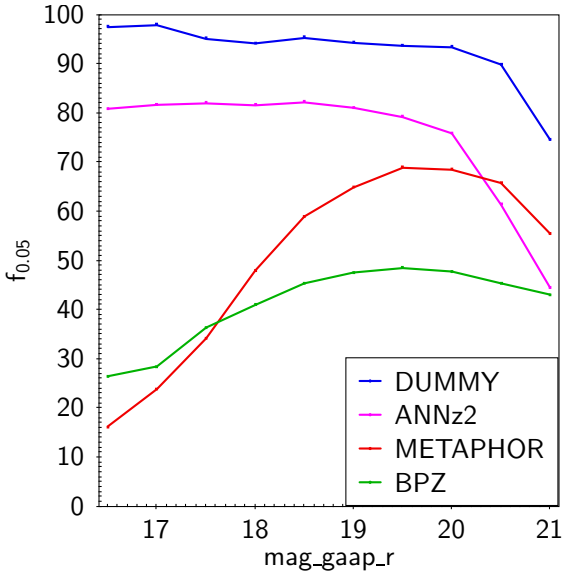

Fig. 5.7 Residuals fraction in the range [-0.05, 0.05] of the PDFs versus magnitude *mag_gaap_r* in the range [16.0, 21.0], used for the tomographic analysis shown in Tab. 5.6. From top to bottom, *dummy* (blue), ANNz2 (violet), METAPHOR (red) and BPZ (green).

By considering the statistics from Table 5.6 and the Fig. 5.7, METAPHOR and BPZ show a similar trend of residual fractions by increasing source faintness, with several trend fluctuations between contiguous bins. Such fluctuations could be due to *edge* effects in the photometric parameter space of the probed magnitudes ranges, and can shed light on different trends from different methods.
In the case of ANNz2 and *dummy* PDFs, we can observe a similar trend (Fig. 5.7) , both showing a decrease of the fraction $f_{0.05}$ from bin 8 to 9, and from 9 to 10, with respect to the other two methods. These trends are an indication of the different feedback to the *edge* effects in the fluctuations between contiguous bins of the parameter space among the compared methods. As for the whole sample, also in the case of the magnitude tomographic analysis ANNz2 obtains better results, except for the last two magnitude bins. Here, in fact, by looking at the Table 5.6, the value of $f_{0.05}$ is better for METAPHOR in the bin 9 and all the statistical estimators are better in the bin 10.
Finally, the tomographic analysis confirms that the *dummy* PDFs obtains the best performance in all bins.

5.5.1 A qualitative discussion

As we have anticipated in Sec. 1.2, a PDF is an intrinsic property of a certain phenomenon regardless the way in which you calculate it. However, in the context of astroinformatics

methods to derive PDFs, the method itself, if not controlled in a proper manner, introduces the contribution of its error within PDF estimate. The comparison just developed in this section is actually among very different methods, whose characteristics can be summarized as follow:

- METAPHOR, is able to remove the method error contribution and to take into account only the uncertainties of the photometric parameter space, provided within the DR3 multi-catalogs, and used to apply the perturbation law, in order to find the PDF and relative point photo-z best- estimates;
- ANNz2 is mainly concerned with a PDF arising from a multi-training of the network, during which the error of the method is propagated. This method do not use at all the uncertainties on the photometric features used for training;
- BPZ makes use of the $flux\_gaap$ and relative errors (see Eq. 4.2) to run and calculate photo-z point estimates and relative PDFs, and it embeds the contribution of the error due to the method itself. The dummy PDF obtained by METAPHOR, does not contain neither the error contribution arising from the method nor that arising from the propagation of the photometric uncertainties

The overall picture which arises from the comparison among methods, shown in this section, would seem as follows: the dummy PDF shows the best quantitative estimator results both for the whole sample and in the tomographic analysis but this PDF is completely useless to fix the nature of the PDF. ANNz2 obtains quantitative results that are just a little bit worse than those of the dummy: this is because we are introducing the method error in the PDF. However again, the information about the uncertainties on the photometric features is totally missing. BPZ shows the worst results in terms of quantitative estimators and METAPHOR shows a quantitative performance that is in the middle between BPZ and ANNz2: this assesses that the removal of the method error improves the quantitative results of METAPHOR with respect to BPZ. If we look only at the performance of the quantitative estimators, we could confirm ANNz2 as the best choice for future studies requiring the use of PDFs; however the point is that the absence of the photometric information is a huge shortcoming of this approach. In any case, a comparison aimed at fixing the scores for the methods it is not possible at this stage, because we should fix the same work conditions for all the methods (except for the benchmark method dummy). For example, if we wanted to fix the best performance between METAPHOR and ANNz2, we know that for this latter the photometric errors can be used somehow, in a particular set-up of the algorithm itself and that the multi-threading implementation of METAPHOR can generate, with N training, the method error.

5.6 Conclusions

Due to the increasing demand for reliable photo-z's and the intrinsic difficulty to provide reliable error PDF estimation for machine learning methods, a plethora of solutions have been proposed. PDFs for machine learning models are conditioned by their intrinsic mechanism to find the hidden flux-redshift relationship. In fact, this mechanism imposes the necessity to disentangle the contributions to the photo-z estimation error budget, by distinguishing the intrinsic method error from the photometric uncertainties. Furthermore, due to the large variety of methods proposed, there is also the problem of finding objective and robust statistical estimators of the PDFs.

In fact, in the absence of systematics, there are some factors affecting the photo-z reliability, such as photometric and intrinsic errors of the methods as well as statistical biases. We believe that it is extremely useful to measure the photo-z error estimation through the intrinsic photometric uncertainties, by considering that the observable photometry cannot be perfectly mapped to the true redshift and that a reliable PDF should exclude the contribution of the intrinsic error of the method.

In Cavuoti et al. (2017), we presented METAPHOR, a method designed to provide a PDF of photometric redshifts calculated by machine learning methods. METAPHOR has already been successfully tested on SDSS (Cavuoti et al., 2017) and KiDS-DR3 (de Jong et al., 2017) data, and makes use of the neural network MLPQNA (Brescia et al., 2013, 014a; Cavuoti et al., 015c) as the internal photo-z estimation engine.

Main subject of the present work is a deeper analysis of photo-z PDFs obtained by different methods, for instance two machine learning models (METAPHOR and ANNz2) and one based on SED fitting techniques (BPZ), through a direct comparison among such methods. The investigation was focused on both cumulative (*stacked*) and individual PDFs reliability, moreover subject to a comparative analysis among different kinds of statistical estimators to evaluate their degree of coherence. Exactly for this reason, by modifying the METAPHOR internal mechanism, we have also derived a *dummy* PDF method (see Sec. 5.4.4), helpful to obtain a benchmark tool to evaluate the objectivity of the various statistical estimators applied on the presented methods.

The credibility analysis, through the Wittman diagrams shown in Fig. 5.5, appears in contrast with the results indicated by the Probability Integral Transforms on the PDFs obtained by the three compared methods (Fig. 5.6). Also in terms of quantitative statistics, i.e. the evaluation of various fractions of error residuals in some ranges and the weighted average of residuals for the *stacked* PDF (Table 5.3), as well as the fractions of occurrences of individual PDFs in different categories based on their relation with the spectroscopic redshift (Table 5.4), there are discordant assessments among the estimators.

The statistical behavior of the *dummy* or *ideal* PDFs confirms only partially the quoted analysis. In fact, all the quantitative estimators improve (Table 5.3). This means that, the more the PDF is representative of an almost perfect mapping of the parameter space on the true redshifts, the better are the performances in terms of quantitative estimators. On the contrary, the PIT histogram and the credibility analysis provide a negative concordant result: the first showing the total *underdispersive* character of the reconstructed photometric redshifts distribution; the second assessing an *overconfidence* of all photo-z estimates.

All these considerations lead us to affirm that only a deep analysis of the performances through a wide combination of statistical estimators may help to understand the whole nature of the measured photo-z error PDFs and to assess the objective validity of the method employed to derive them.

Chapter 6

Weak Lensing measurements with METAPHOR photo-z and relative PDFs

As said many times before, one of the main drivers behind the implementation of METAPHOR, was the need to produce PDFs as requested by the analysis of Weak Lensing (WL) data. A requirement which has been enforced also for the Euclid space mission. In this final Chapter after a short introduction to "weak lensing", I will describe a first attempt to use ML based photo-z's to the analysis of weak lensing data. What presented here is just a preliminary one. It was in fact obtained toward the end of my PhD and there was not enough time to put it on firmer ground.

6.1 Introduction to Weak Lensing

Weak gravitational Lensing, is the weak distortion of the galaxies images induced by the inhomogeneities in the Universe Large Scale Structure (LSS, made up of voids, filaments, halos) along the line of sight. The larger the amplitude of the inhomogeneity of the cosmic web is, the larger these deformations are.
The typical weak distortions of high-redshift galaxies are of the order of a few per cent, much smaller than the width of the intrinsic galaxy shape and size distribution: therefore, WL is not detectable for a single galaxy but only from the average distortion of many different sources. By measuring galaxy shape correlations between different redshifts, the evolution of the LSS can be traced, thus unabling us to detect the effect of dark energy on the growth of structure and its amount along with that of dark matter.
The measurements of the coherent distorsions of millions of galaxy images as a function of

angular separation on the sky and redshift is therefore a probe for the standard ΛCDM model, along with the observation of the large scale geometry, the Universe expansion rate, and the structures formation.
The ΛCDM model in which Λ is responsible for the late acceleration of the Universe and the Cold Dark Matter (CDM) drives structure formation, has been already confirmed by some important observational evidences such as: statistics of anisotropies in the cosmic microwave background, Hubble diagram of supernovae of type Ia, big bang nucleosynthesis and galaxy clustering.

Several future planned surveys (like the Large Synoptic Survey Telescope (Ivezic 2009; LSST Science Book 2009) and Euclid (Laureijs et al. 2014, Euclid 2011) have been conceived with the aim of measuring the effect of WL, since both the volume covered by the ongoing surveys (cf. KiDS - Kilo-Degree Survey, de Jong et al. 2015, 2017; DES - Dark Energy Survey, Annis 2013) and their measurement precision are still not sufficient.
In figure 6.1, the effect of the light distortion due to the presence of Universe Large Scale Structure, is represented. In the next section, the cosmological background along with the basis of the WL theory will be described in some detail. This in order to understand the quantities that will be calculated using the METAPHOR photometric redshift PDFs (as well as the punctual photo-z estimates) and the shear measurements. The latter being provided in the KiDS-450 catalog which is described in Section 6.4.

6.2 Weak Lensing: cosmological background

To give a detailed exposition of the formal theory of weak lensing goes beyond the scope of this Section. However, the basic WL equations will be derived in order to underline which are the actual physical observables of the problem, along with the quantities that can be inferred. For a complete review of the WL formalism, the interested reader is referred to Kilbinger (2015), while for a complete review of the the systematics involved we recommend Hildebrandt et al. (2017).
First, it is essential to start from tracing the cosmological background in order to fix the framework in which weak lensing is explainable.
It is known that the relationship between space-time geometry and matter-energy content of the Universe is fixed by the field equations of General Relativity (GR): a solution of such equations exists and represents a homogeneous and isotropic universe. However, to quantify gravitational lensing, we have to consider the light propagation in an inhomogeneous universe. In such a Universe, and in the so-called regime of weak field, the Friedmann

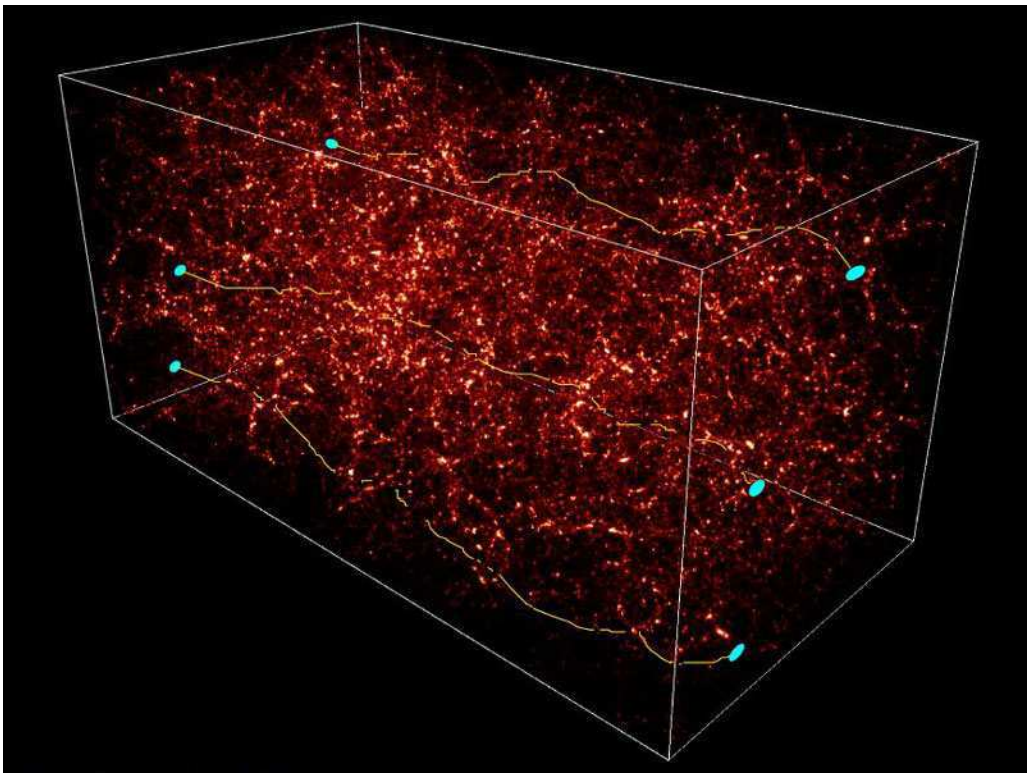

Fig. 6.1 A schematic representation of the deflection of light rays from distant objects due to the presence of matter along the line of sight.

Lamaitre Robertson Walker (FLRW) metric contains the Bardeen potentials ϕ and ψ which are $\ll c^2$. Such potentials represent first-order approximation of the metric in an expanding Universe. The FLRW metric is expressed in the equation 6.1.

$$ds^2 = \left(1 + \frac{2\psi}{c^2}\right) c^2 dt^2 - a^2(t) \left(1 - \frac{2\phi}{c^2}\right) dl^2 \tag{6.1}$$

where the scale factor a depends on time, c is the speed of light and l is the comoving coordinate which remains constant when the universe expands. The spatial line element dl^2 can be split in a radial and angular part, as in equation 6.2:

$$dl^2 = d\chi^2 + f_K^2 d\omega \tag{6.2}$$

where χ is the comoving coordinate, and f_K the comoving angular distance.
According to the value of the universe curvature K, three different forms for f_K are allowed, in a three dimensional space:

$$f_K = \begin{cases} K^{-1/2} \sin\left(K^{1/2}\chi\right) & for\ K > 0\ (spherical) \\ \chi & for\ K = 0\ (flat) \\ -K^{-1/2} \sinh\left(-K^{1/2}\chi\right) & for\ K < 0\ (hyperbolic) \end{cases}$$

each of them characterized by an equation of state, linking the universe pressure p to its density ρ, via the parameter ω:

$$p = \omega c^2 \rho \tag{6.3}$$

There are three main density components in the Universe, each of them usually scaled by the present-day value of the universe critical density:

$$\rho_{c,0} = \frac{3H_0^2}{8\pi G} \tag{6.4}$$

for which the Universe has a flat geometry. In equation 6.4, $H_0 = H(a=1) = (\dot{a}/a)_{t=t_0} = 100 h km s^{-1} Mpc^{-1}$) is the present value of the Hubble constant, and h parametrizes the uncertainty on the knowledge of such constant: nowadays the best estimates give $h = 0.7$ (Planck Collaboration XVI (2014)).
These components are:

- the density parameter of non-relativistic matter (baryonic plus Cold Dark Matter, CDM, and possibly heavy neutrinos):

$$\Omega_m = \Omega_m + \Omega_b + \Omega_\nu$$
$$\Omega_m = \rho_{m,0}/\rho_{crit,0}$$

- the relativistic matter (Ω_r) consisting of photons (with the main contribution given by the cosmic micro-wave background) and light neutrinos;
- the dark energy density component (Ω_{DE}) which regulates the expansion of the universe.

Finally, the curvature density parameter (Ω_K) is defined by the sum of all the mentioned density parameters, in such a way that $\Omega_m + \Omega_{DE} + \Omega_r = 1 - \Omega_K$, with $\Omega_K = -(c/H_0)^2 K$ which has opposite sign with respect to the curvature K.

In an expanding universe, density fluctuations evolve with time: from the tiny quantum fluctuations set during the inflationary phase, afterward driven substantially by ρ_{CDM} which interacts only gravitationally inducing small-amplitude density fluctuations which then evolve in the large structures visible today (clusters, filaments, halos).
There are cases (e.g., in the early epochs of the Universe and on large enough scales) in which the linear perturbation theory is sufficient to treat the growth of the density fluctuations. In such cases, fluctuations are small, thus leading to small values of the density contrast (see below equation 6.5). On the contrary, when the perturbations grow, a non-linear perturbation theory is needed or other approaches (e.g., N-body simulations or analytical models) in order to treat them.
The density contrast δ parametrizes the fluctuations of the density ρ around the mean density $\bar{\rho}$:

$$\delta = \frac{\rho - \bar{\rho}}{\bar{\rho}} \tag{6.5}$$

The gravitational instability of the original fluctuations are described by the solutions, for the dimensionless density contrast δ in equation 6.5, of a system of three hydrodynamic equations: the Euler and Poisson ones, and the continuity equation.
This quoted system is linearized for small fluctuations ($\delta \ll 0$) and presents two different solutions: a homogeneous one for the Hubble expansion and an inhomogeneous solution for the density contrast. The solutions for the density contrast δ are expressed by a second order differential equation, which has the form of a damped wave equation and governs the gravitational amplification of the fractional density contrast itself.
The form of the differential equation, all the mathematical steps and the considerations on

the physical quantities involved, will not be reported here since they are beyond the scope of this Section. However we say only that from the differential equation for the density contrast, it is inferable the growing solution, for the *growing factor* D_+, which relates the density contrast at time a to an earlier, initial epoch a_i (Kilbinger, 2015):

$$\delta(a) = D_+(a)\delta(a_i) \tag{6.6}$$

In different cosmological models the perturbation amplitudes evolve differently.
In an Einstein-de Sitter Universe they grow proportionally to the scale factor a, while in a low density Universe there is a slower growth rate at low redshifts. In a ΛCDM model, i.e. the accepted model of structure formation, there is an intermediate degree of structure evolution with respect to the two quoted models, since the Hubble expansion must have been slower in the past (with respect to an open Universe). Accordingly to the ΛCDM model, the structure formation is hierarchical and galaxy groups and clusters result built from subsequent merging of smaller systems.

6.3 The weak lensing formalism

To describe WL, one of the approaches is to derive the equations describing the deflection of light rays in the presence of massive bodies using the Fermat's principle of minimal light travel. The principle:

$$\delta t = 0 \tag{6.7}$$

is applied to the light ray travel time inferable from the FLRW metric in equation 6.2 since photons propagate on null geodesics ($ds_{FLRW} = 0$):

$$t = \frac{1}{c}\int \left(1 - \frac{2\psi}{c^2}\right) dr \tag{6.8}$$

The integral in equation 6.8 is along the light path in physical coordinates dr.
The gravitational lensing takes its name from the fact that, as in geometrical optics, the potential ψ in equation 6.8 is the analogous of a medium with a refractive index $n = 1 - 2\psi/c^2$.
From the Fermat principle, we can derive the Euler-Lagrange equations along the light path which, integrated, leads to the determination of the deflection angle $\hat{\alpha}$ (i.e., the difference between the directions of emitted and received light rays):

$$\hat{\alpha} = -\frac{2}{c^2}\int \nabla_\perp^p \phi \, dr \tag{6.9}$$

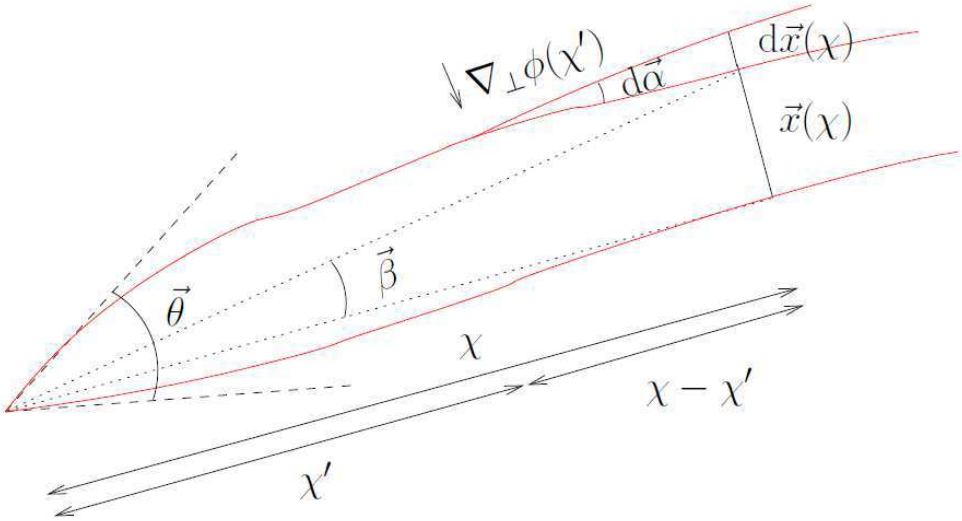

Fig. 6.2 Propagation of two light rays (red solid lines), converging on the observer on the left. The light rays are separated by the transverse comoving distance **x**, which varies with distance χ from the observer. An exemplary deflector at distance χ' perturbes the geodescics proportional to the transverse gradient of the potential. The dashed lines indicate the apparent direction of the light rays, converging on the observer under the angle θ. The dotted lines show the unperturbed geodesics, defining the angle β under which the unperturbed transverse comoving separation **x** is seen. Caption and figure are taken from Kilbinger (2015).

The gradient of the potential in equation 6.9 is perpendicular to the light travel with respect to physical coordinates p.
With reference to figure 6.2, in which the difference between two nearby geodesics is represented, the transverse comoving separation $\mathbf{x}_0(\chi)$ between two light rays in function of the comoving distance χ from the observer, is proportional to the comoving angular distance:

$$\mathbf{x}_0(\chi) = f_K(\chi)\theta \tag{6.10}$$

In figure 6.2:

- the separation vector $\mathbf{x}_0$ is seen by the observer under a small angle θ;
- the comoving distance between the observer and the lens is χ';
- the comoving distance between the observer and the lensed object is χ;
- the comoving distance between the lens and the lensend source is $\chi - \chi'$.

The amount of deflection of a light ray, in presence of the potential ϕ at distance χ' is:

$$d\hat{\alpha} = -\frac{2}{c^2} \bigtriangledown_\perp \phi\left(\mathbf{x}, \chi'\right) d\chi' \tag{6.11}$$

The induced change in the separation vector, in the vantage point of the deflector is:

$$\mathbf{x} = f_K\left(\chi - \chi'\right) d\hat{\alpha} \tag{6.12}$$

At the end, we have to integrate over the line of sight χ' in order to obtain the total separation:

$$\mathbf{x}(\chi) = f_K(\chi)\theta - \frac{2}{c^2}\int_0^{\chi} d\chi' f_K\left(\chi - \chi'\right)\left[\bigtriangledown_\perp \phi(\mathbf{x},\chi') - \bigtriangledown_\perp \phi^{(0)}(\chi')\right] \tag{6.13}$$

In absence of lensing, the vector $\mathbf{x}$ would be seen by the observer under an angle:

$$\beta = \mathbf{x}(\chi)/f_K(\chi) \tag{6.14}$$

One can write the *lens equation* for the scaled deflection α:

$$\beta = \alpha - \theta \tag{6.15}$$

with:

$$\alpha = \frac{2}{c^2}\int_0^{\chi} d\chi' \frac{f_K\left(\chi - \chi'\right)}{f_K(\chi)}\left[\bigtriangledown_\perp \phi(\mathbf{x},\chi') - \bigtriangledown_\perp \phi^{(0)}(\chi')\right] \tag{6.16}$$

By using the fact that the vector $\mathbf{x}$ can be approximated by the 0th-order solution $\mathbf{x}_0(\chi) = f_K(\chi)\theta$ (equivalent to integrate the potential gradient along the unperturbed ray, Born approximation), the equation 6.13 can be approximated.

6.3.1 The Actual observables in Weak Lensing

We can linearize the lens equation and define the inverse amplification matrix as the Jacobian $\partial\mathbf{A} = \partial\beta/\partial\theta$, which describes the linear mapping from the lensed (image) coordinates θ to the unlensed (source) coordinates β (Kilbinger, 2015),

$$A_{ij} = \frac{\partial\beta_i}{\partial\theta_i} = \delta_{ij} - \frac{\partial\alpha_i}{\partial\theta_i} = \delta_{ij} - \frac{2}{c^2}\int_0^{\chi} d\chi' \frac{f_K\left(\chi - \chi'\right) f_K(\chi'))}{f_K(\chi)} \frac{\partial^2}{\partial x_i \partial x_j}\phi(f_K(\chi')\theta, \chi') \tag{6.17}$$

In the integral of equation 6.17 we have two factors, the first involving the quotient of functions f which represents the *cosmological component* and the other, containing the second derivatives which is referred to as the *structure component*.
In the approximation in which the second term in equation 6.16 vanishes, the deflection angle α is the gradient of a bi-dimensional potential ψ, given by:

$$\psi(\theta,\chi) = \frac{2}{c^2}\int_0^{\chi} d\chi' \frac{f_K(\chi-\chi')}{f_K(\chi)f_K(\chi')}\phi(f_K(\chi')\theta,\chi') \tag{6.18}$$

With equation 6.18, we are simply projecting the information of the 3D potential ϕ into the two-dimensional surface, on the sky plane via the 2D potential ψ: this allows to re-write the Jacobi matrix **A** as:

$$A_{ij} = \delta_{ij} - \partial_i\partial_j\psi \tag{6.19}$$

The matrix **A** can be further parametrized in terms of the scalar *convergence* κ, and of the two component *shear*, $\gamma = (\gamma_i, \gamma_j)$, as:

$$\mathbf{A} = \begin{bmatrix} 1-\kappa-\gamma_1 & -\gamma_2 \\ -\gamma_2 & 1-\kappa+\gamma_1 \end{bmatrix} \tag{6.20}$$

In this way *convergence* and *shear* can be written as second derivatives of the potential, as:

$$\kappa = \frac{1}{2}(\partial_1\partial_1 + \partial_2\partial_2)\psi;\ \gamma_1 = (\partial_1\partial_1 - \partial_2\partial_2)\psi;\ \gamma_2 = \partial_1\partial_2\psi \tag{6.21}$$

In the weak lensing regime, as anticipated, the values of κ and γ are of a few percent, thus the Jacobian matrix A is invertible. Its inverse A^{-1} describes the local mapping of the source light distribution to image coordinates. The convergence, i.e. the trace of the matrix, is an isotropic increase or decrease of the observed size of a source image. Shear, the part outside the trace, quantifies an anisotropic stretching, turning a circular into an elliptical light distribution as it is schematized in figure 6.3.
We are finally able to define, after this discussion, the actual observable of the weak lensing effect since the factor κ only affects the size but not the shape of the source. Therefore we can take out it from the matrix **A** in equation 6.20 and define the actual observable, i.e. the reduced shear, as:

$$g = \frac{\gamma}{1-\kappa} \tag{6.22}$$

Furthermore, if we associate to a galaxy a complex source ellipticity ε^s, cosmic shear will modify it as a function of the complex reduced shear. If we define this ellipticity for an image with elliptical isophotes, minor-to-major axis ratio b/a, and position angle ϕ, as:

$$\varepsilon = (a-b)/(a+b)\exp^{-2i\phi} \tag{6.23}$$

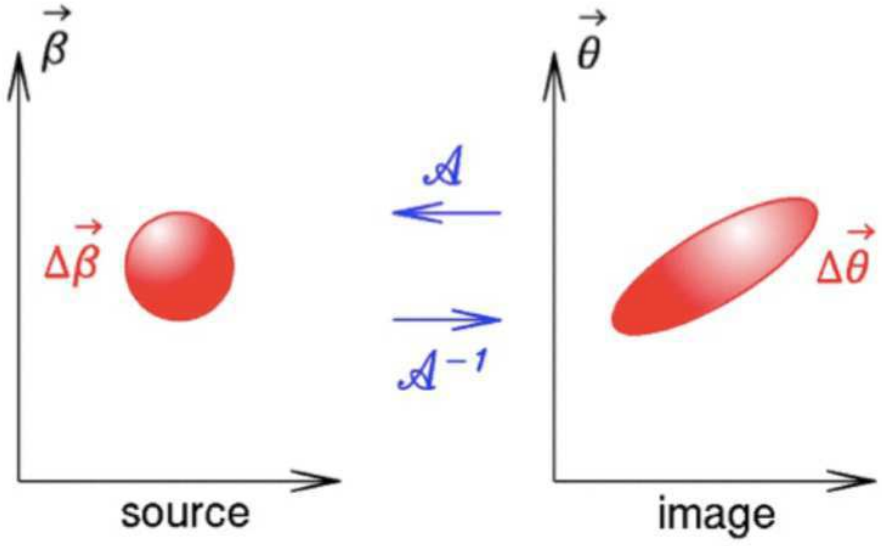

Fig. 6.3 Representation of the size and shear effects.

The ellipticity becomes:

$$\varepsilon = \frac{\varepsilon^s + g}{1 + g^* \varepsilon^s}; \;\; \varepsilon \approx \varepsilon^s + \gamma; \;\; \langle \varepsilon \rangle = g \tag{6.24}$$

The observed ellipticity is then an unbiased estimator of the reduced cosmic shear, at least in absence of intrinsic alignment, which is one of the sources of systematics in weak lensing (anyway, the intrinsic alignment discussion is beyond the scope of this Section).

6.3.2 Cosmology from the convergence factor and photometric redshifts

In this section we describe the connection between the measurable convergence factor and cosmology. From this we obtain the relation of the observable with the astronomical sources distances (obtained via photometric redshifts).
As we said, the convergence factor is related to the lensing potential ψ via a 2D Poisson equation, then it can be interpreted as a (projected) surface density. In order to take into account the effect of the matter in the universe on this parameter (through the introduction of the 3D density contrast δ), we have:

- To apply a 2D Laplacian to the 3D potential ϕ;
- To add a third second derivative in comoving coordinate to such Laplacian;

- To replace the 3D Laplacian of ϕ with the over-density δ using the Poisson equation for it;
- To consider the proportionality relation between the average density and the scale factor:

$$\bar{\rho} \propto a^{-3}$$

We obtain in this way a convergence factor weighted by geometrical factors involving the distances between source, deflector, and observer:

$$\kappa = \frac{3H_0^2\Omega_m}{2c^2}\int_0^{\chi}\frac{d\chi'}{a(\chi')}\frac{f_K(\chi-\chi')}{f_K(\chi)}f_K(\chi')\delta(f_K(\chi')\theta,\chi') \tag{6.25}$$

The mean convergence from a population of galaxies is obtained by weighting the above expression with the galaxy probability distribution in comoving distance out to the comoving distance χ_{lim} of the galaxy sample. The distance is usually obtained using photometric redshifts $n(z)dz$:

$$\kappa(\theta) = \int_0^{\chi_{lim}} d\chi n(\chi)\kappa(\theta,\chi) \tag{6.26}$$

The convergence is then a linear measure of the total matter density, projected along the line of sight, with dependences on the geometry of the universe via the distance ratios and the source galaxy distribution. In any case, the expectation values of convergence and shear are zero since $\langle\delta\rangle = 0$. Therefore the first non-trivial statistical measure of the distribution of κ and γ are second moments.
We shall not discuss all the theory about the two-point correlation function (2PCF) which represents the way to take into account the second moment of the distribution, since it is beyond the scope of this thesis. For a complete review, the interested reader is again referred to Kilbinger (2015).
Finally, we conclude that the important quantities are:

- the shear factor that is the actual observable, inferable from the measurements of galaxy ellipticities;
- the convergence factor κ that is calculated from the observable shear and the photometric redshifts distributions, via the equations shown above. This factor, due to its link to cosmological parameters, allows also to constraint such quantities.

6.3.3 Galaxy-galaxy lensing

With respect to weak lensing, galaxy-galaxy lensing or GGL usually correlates the shapes of high-redshift galaxies with the position of galaxies at lower redshifts (Kilbinger, 2015).
The GGL is therefore used to measure the density distribution around foreground galaxies (lenses) using the shear (see section 6.3.1) of a suitable amount of background galaxies (sources).
We said that the shear γ is not measurable on a single galaxy since the distortion is only a few percent of the shape of the source. Therefore, GGL can only be calculated statistically by azimuthally averaging the shear of many sources around lenses and then by stacking the lens signals for many lenses grouped together according to some of their observable properties.
The measured quantity is the tangential shear along the line of sight joining the lens to the source galaxy: such shear component is then averaged for all the couples lens-source for all available lenses, by obtaining in this way the average tangential shear $\langle \gamma_t \rangle (R)$.
The average tangential shear can be connected to the Excess Surface Density (ESD) profile $\Delta\Sigma(R)$. This is the surface mass density $\Sigma(R)$ at the projected radial distance R from the lens centre subtracted to the average density $\bar{\Sigma}(< R)$ within that radius, i.e:

$$\langle \gamma_t \rangle (R)\Sigma_{crit} = \Delta\Sigma(R) = \bar{\Sigma}(< R) - \Sigma(R) \tag{6.27}$$

where the Σ_{crit} is the critical surface mass density which contains information on the angular diameter distances (D) of the lens (z_l) and of source (z_s) from the observer, and between the lens and the source (z_l, z_s):

$$\Sigma_{crit} = \frac{c^2}{4\pi G} \frac{D(z_s)}{D(z_l)D(z_l, z_s)} \tag{6.28}$$

The effective critical surface density for each pair of spectroscopic redshift of the lens (z_l) and the full posterior redshift distribution of the source $p(z_s)$ is:

$$\tilde{\Sigma}_{crit}^{-1} = \frac{4\pi G}{c^2} \int_{z_l}^{\infty} \frac{D_l(z_l) D_{ls}(z_l, z_s)}{D_s(z_s)} p(z_s)\, dz_s \tag{6.29}$$

The approach explained in equation 6.29 is the same followed by Viola et al. (2015) to which the interested reader is referred for more details. The GGL pipeline (Dvornik et al., 2017) used by the authors of the quoted paper calculates the quantities in equation 6.29 in order to find the ESD. The weights, which contain information on possible shear measurements uncertainties, are calculated by the code *lensfit* used for the measurement of the object shapes (ellipticities). Weights and shapes are combined to find, through the two-point correlation function, the tangential shear in equation 6.27.

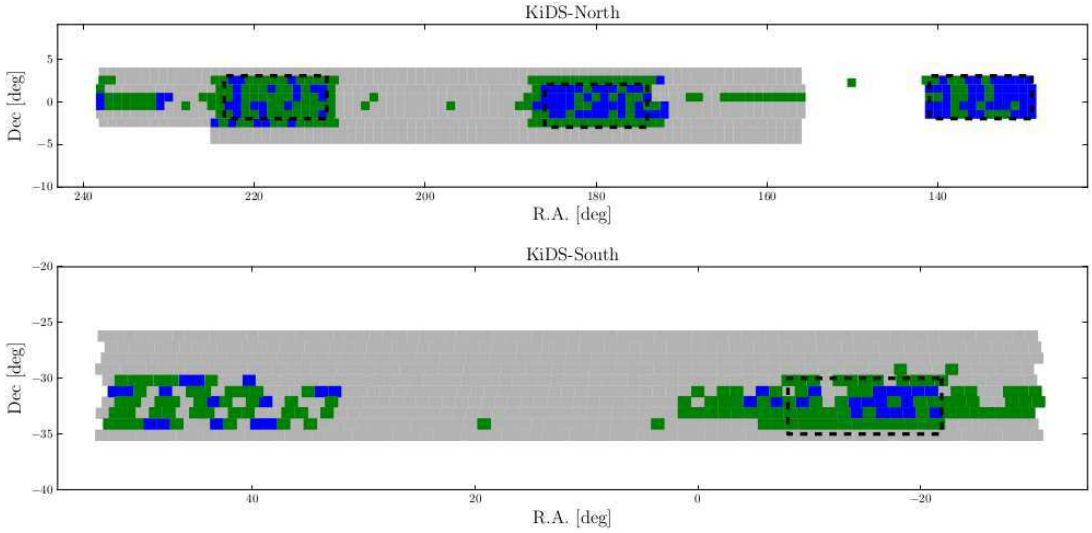

Fig. 6.4 Sky distribution of survey tiles released in KiDS-ESO-DR3 (green) and in the previous releases KiDS-ESO-DR1 and -DR2 (blue). The multi-band source catalog covers the combined area (blue + green) and the full KiDS area is shown in grey. Top: KiDS-North. Bottom: KiDS- South. Black dashed lines delineate the locations of the GAMA fields. Caption and figure are taken from de Jong et al. (2017).

6.4 The data

The Leiden University WL group designed and developed a GGL pipeline able to perform several measurements, among which that of the ESD as a function of the projected physical distance R from the lens centre. For details on such pipeline we refer the interested reader to the paper by Dvornik et al. (2017).

In order to conduct a study on the new KiDS Data Release 3 galaxies, by using METAPHOR derived probability density functions (see eq. 6.29), we used the GGL pipeline, which was also employed in the work of Viola et al. 2015, who produced the shear measurements for the galaxies in KiDS DR1 and 2 (de Jong et al., 2015). The first two KiDS releases covered 100 KiDS tiles, in all four optical bands (u,g,r,i). The effective area after removing masks and overlaps between tiles was $68.5 deg^2$.

For the present work instead, the galaxy shear measurements contained in the catalog KiDS-450, publicly available, have been used. This catalog contains 15 million galaxies over a total effective area of $360.3 deg^2$ (masking and overlap of the tiles considered).

In figure 6.4 it is shown the different tile coverage of KiDS DR3 (de Jong et al., 2017) with respect to the release 1 and 2 (de Jong et al., 2015) and the overlapping Galaxy And Mass Assembly (GAMA, Liske et al. 2015) spectroscopy.

For what the lenses used concerns, the GAMA (Driver et al., 2011) galaxy groups (with

at least 5 galaxies) have been utilized. The three GAMA equatorial regions used are the $G9$, $G12$, $G15$ ($180 deg^2$ is the sky area covered from these patches) of the GAMA Release 2 (Liske et al., 2015).
In the equatorial region, the KiDS footprint overlaps with the footprint of the GAMA spectroscopic survey (see figure 6.4). This catalog contains $180,960$ galaxies and $23,838$ galaxy groups (found by a *fof* algorithm, see the paper of Robotham et al. (2011) for details on the identification of the galaxy groups). In the paper of Viola et al. (2015), the $\sim 1,400$ galaxy groups with at least five members, were selected, and are used also in the present work.
The equatorial patches GAMA galaxies overlap the KiDS DR3 METAPHOR galaxies for which we provide, as it was described in Chapter 5, more than 8 million galaxy PDFs and photo-z punctual estimates.
The coordinate cross-match between the METAPHOR KiDS DR3 galaxy sources and those of the KiDS-450 catalog (GAMA patches $G9$, $G12$, $G15$) leads to $2,622,700$ objects for which we have shear along with METAPHOR punctual photo-z's and relative PDFs.
In the paper of Viola et al. (2015) the distances to the individual source (background galaxies) and the PDFs were derived using the SED fitter BPZ.
In the present work, we want to test the performance of Machine Learning PDFs as well as of punctual photo-z estimates in WL ESD measurements, these latter not probed in the paper of Viola et al. (2015).
In figure 6.5 we show the redshift distribution of the GAMA groups used in this work (as well as that used in the work Viola et al. (2015)) with superimposed the stacked representation of the whole KiDS DR3 METAPHOR distributions for either the PDFs and for the punctual estimates (calculated through the dummy PDF version, see the reason why in the next section) in order to obtain punctual photo-z estimates (see next Sec. 6.4.1) .

6.4.1 Data preparation

The accuracy of METAPHOR PDFs for KiDS DR3 was fixed at $\Delta z = 0.01$, instead for measurements of shear an accuracy of 0.05 is sufficient. The first action was then to rebin the original data to the new required accuracy.
A cross-match has been performed between the shear catalog KiDS-450 and the METAPHOR sources.
Successively, in order to be compliant to the required input catalog format, the rebinned METAPHOR PDFs, for objects in the range of redshift $]0, 3.5]$ (70 redshift bins) had to be compressed in a vector.

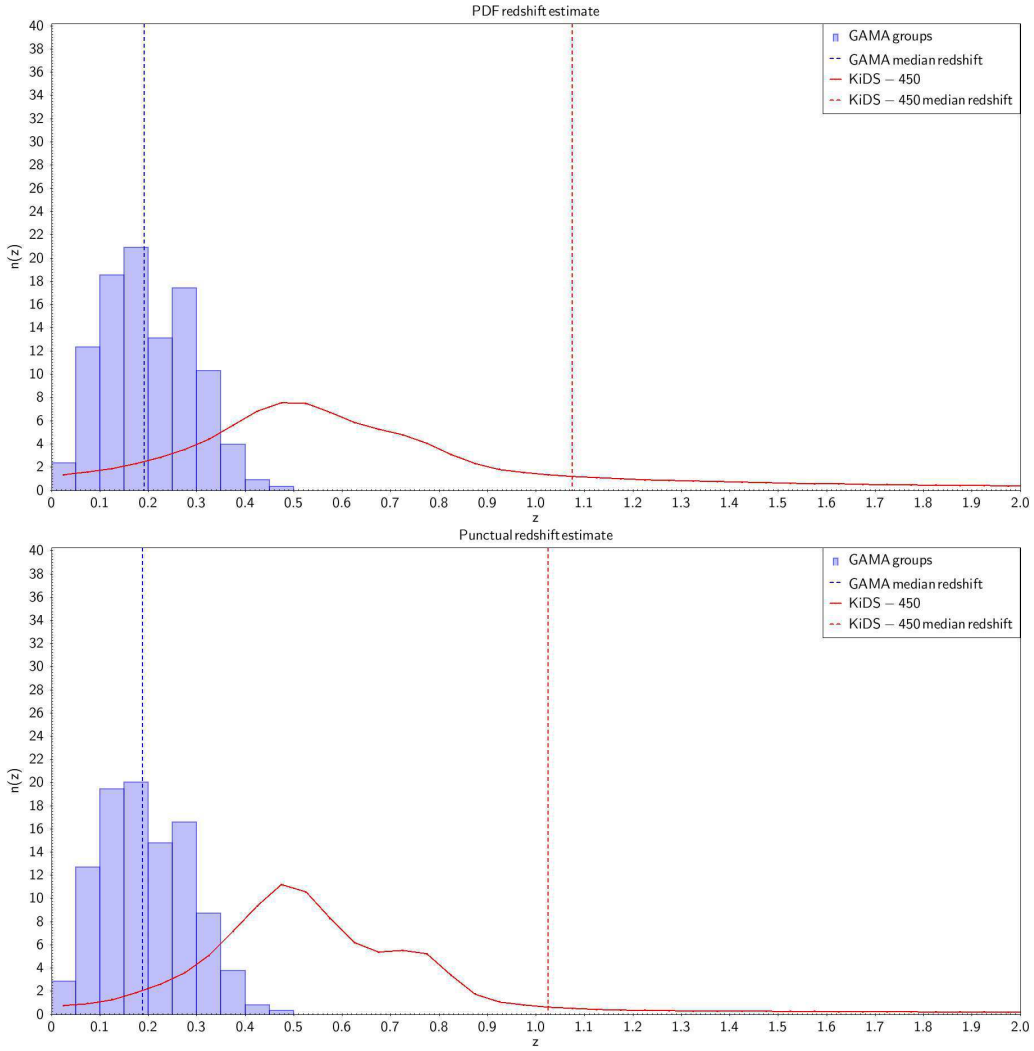

Fig. 6.5 Redshift distribution of the GAMA groups used in this analysis (blue histogram) and the KiDS DR3 galaxies (red lines). For the KiDS DR3 galaxies the redshift METAPHOR distribution is computed as a staked PDF (top panel) and a stacked dummy PDF (bottom panel). The dummy PDF allows to use the GGL pipeline with the punctual photo-z estimates (cf. Sec. 6.4.1).

Some other procedures had been in order to process the data. In particular, since the GGL pipeline does not foresee the possibility to use punctual photo-z estimates but only individual full posterior probabilities, an artifact was used: the *dummy* PDFs (see Sec. 5.4.4) were employed in order to derive the punctual photo-z estimates peaked in the corresponding redshift bin (of accuracy 0.05). In other words, the punctual estimates have been produced as dummy PDFs. The aim was to see a possible difference in the ESD results in the comparison between photo-z and PDFs ESD measurements.
Finally, the produced catalogs containing the cross-match of METAPHOR KiDS DR3 products and the KiDS-450 shear were further cross-matched with the KiDS tiles in order not to create memory issues in running the GGL pipeline. This led to a processing on lighter catalogs (divided in tiles) with respect to those corresponding to the GAMA patches.

6.5 Results and Conclusions

In this section we report the preliminary ESD measurements obtained using the procedure described above. First of all, we wish to point out, as it can be seen in figure 6.7, that the trend of the ESD obtained using ML based PDFs is almost identical to that found by Viola et al. (2015) who used the BPZ SED based approach. This is even more evident if the smallest radial projected distance R is set to $30kpc$ (see figure 6.8) rather than $20kpc$.
For convenience we report in figure 6.6 the ESD profile obtained by Viola et al. (2015) with GAMA lenses and KiDS DR1-2 source galaxy shear measurements. We remember that in such work the full individual posterior redshift distribution $p(z_s)$ for each object, calculated by the SED fitter BPZ, has been used in order to find the ESD, according to the equations 6.27 and 6.29. We also find that both using photo-z and PDFs, with a lowest radial distance of $20kpc$ negative values of ESD appear, while in the work of Viola et al. (2015) such negative values are present for radial distances lower than $20kpc$ (see figure 6.6). This is due to the fact that the signal to noise is very poor at scales smaller than $20kpc$ since many objects close to the group centres are blended. The discrepancy on the ESD behavior in the first radial bin (at $20kpc$) between the present results and those found by Viola et al. (2015), is most likely due to the different available statistics, i.e. to the different number of source galaxies available in our dataset and in that of Viola et al. (2015).
We remember, in fact, that with respect to the KiDS-450 shear measurements for the three GAMA patches, containing more than 10 million galaxies with measured shear, we have found a cross-match with our KiDS DR3 objects with estimated PDFs only $2,622,700$ objects. This difference being due to the need for ML methods to remain within the boundaries imposed by the training set. A limit which is not present in SED fitting methods.

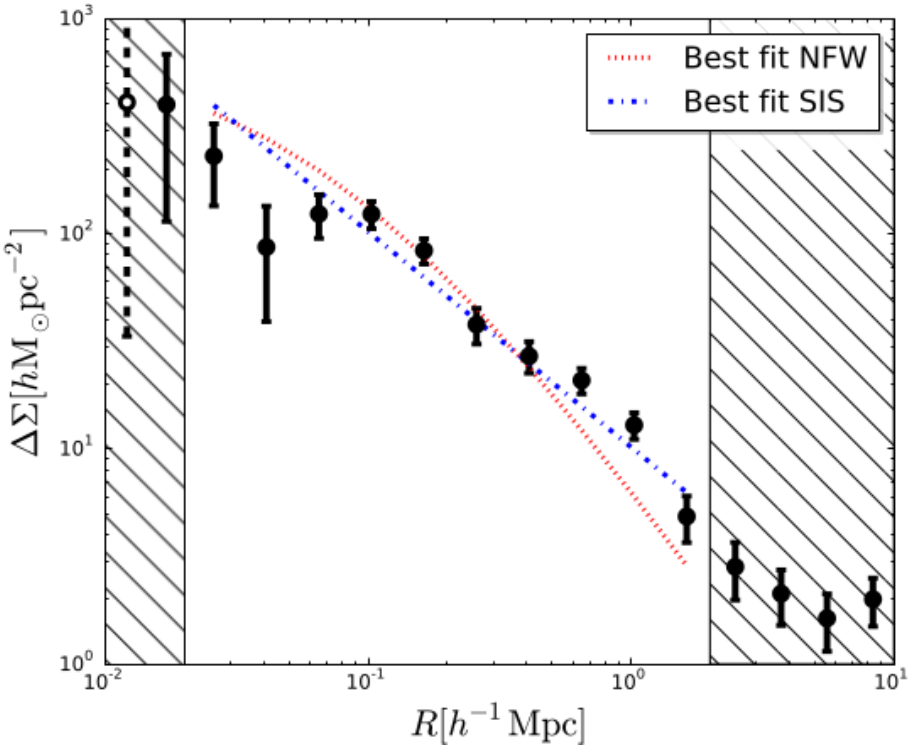

Fig. 6.6 ESD profile measured from a stack of all GAMA groups with at least five members (black points). Here, we choose the BCG as the group centre. The open white circle with dashed error bars indicates a negative . The dotted red line and the dash–dotted blue line show the best fits to the data of NFW (Navarro et al. 1995) and the best-fitting singular isothermal sphere (SIS) profiles, respectively. Caption and figure are from Viola et al. (2015)).

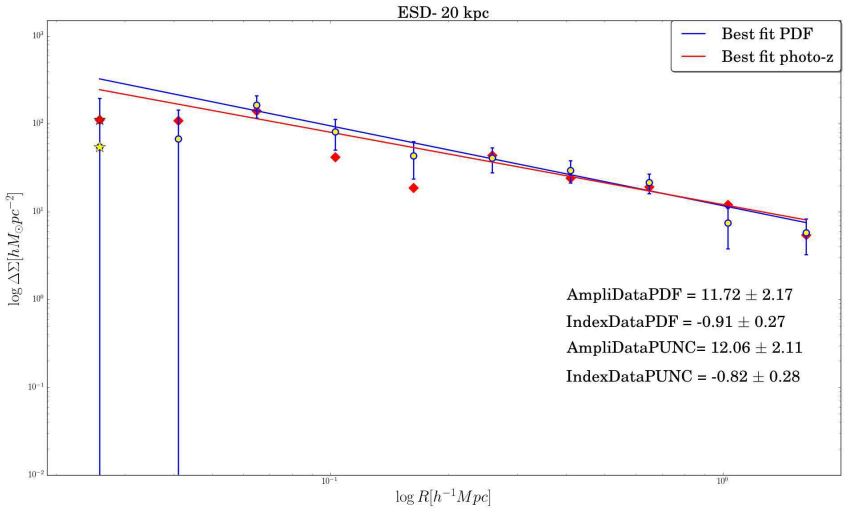

Fig. 6.7 ESD profile measured from a stack of GAMA groups with at least five members by using the individual source PDFs (blue yellow-filled circles) and the punctual redshift photo-z (red diamonds) and relative best fits (in blue and red respectively). The radial bin distance from the lens centre ranges from $20kpc$ up to $2Mpc$. The star dots correspond to a negative $\Delta\Sigma$. The quantities shown are the amplitude ("Ampli") and the exponent ("Index") of the relative best fit power laws of the type $y = Ampli \times x^{Index}$. Therefore Index is the actual angular coefficient of the plotted best fit straight lines.

In figure 6.8, one can see, that by increasing the lowest radial distance from the lens centre from $20kpc$ to $30kpc$, the effect of the presence of negative ESDs disappears. However a deeper comparison between the dataset used by Viola et al. (2015) and the one probed in the present work should be performed in order to grasp the distributions of the two datasets. The most interesting implication of the present work is however the fact that the ESD obtained by using PDFs and punctual redshift estimates are indistinguishable within the errors (figures 6.7 and 6.8).
The quantities indicated by "Index" in figures 6.7 and 6.8, are the exponents of relative best fit power laws and they confirm that they are almost indistinguishable. This is an important clue at least within the accuracy required in galaxy-galaxy lensing studies ($\Delta z = 0.05$), the use of PDFs does not improve with respect to the punctual photo-z estimates.

Finally, in figure 6.9, we superimpose the points in the Viola ESD profile and the relative best fit, to our results for a radial distance range, from the lens centre, of $[20kpc, 2Mpc]$. As we can see the values of the "Index" of the corresponding power laws best fits are the same for both Viola and METAPHOR PDFs. The relative straight lines are indeed parallel since

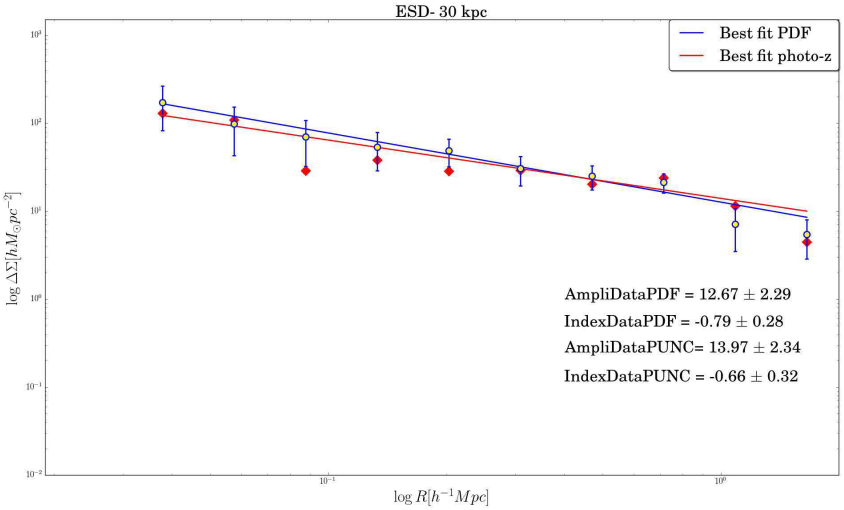

Fig. 6.8 ESD profile measured from a stack of GAMA groups with at least five members by using the individual source PDFs (blue yellow-filled circles) and the punctual redshift photo-z (red diamonds) and relative best fits (in blue and red respectively). The radial bin distance from the lens centre ranges from $30kpc$ up to $2Mpc$. The quantities shown are the amplitude ("Ampli") and the exponent ("Index") of the relative best fit power laws of the type $y = Ampli \times x^{Index}$. Therefore Index is the actual angular coefficient of the plotted best fit straight lines.

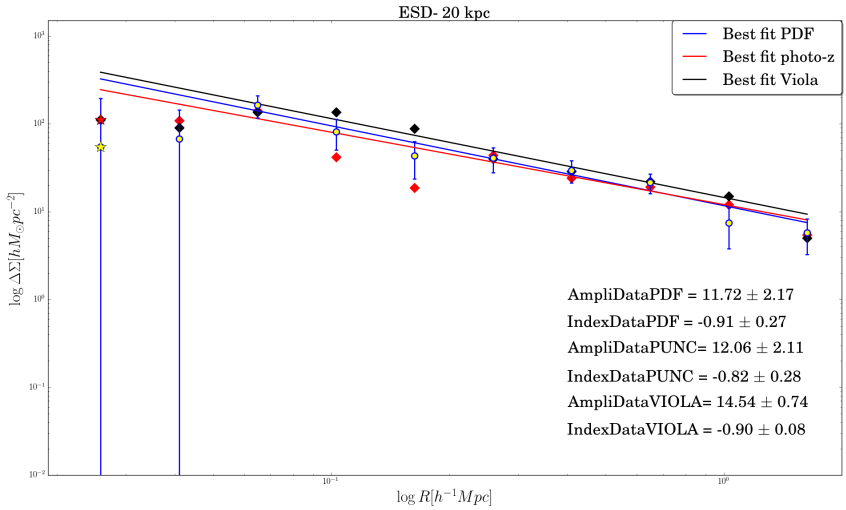

Fig. 6.9 ESD profile measured from a stack of GAMA groups with at least five members by using the individual source PDFs (blue yellow-filled circles) and the punctual redshift photo-z (red diamonds) and relative best fits (in blue and red respectively). The radial bin distance from the lens centre ranges from $30kpc$ up to $2Mpc$. Superimposed to them, the ESD profile (black diamonds points) of Viola et al. (2015) visible in figure 6.6 with relative best fit in black. The quantities shown are the amplitude ("Ampli") and the exponent ("Index") of the relative best fit power laws of the type $y = Ampli \times x^{Index}$. Therefore Index is the actual angular coefficient of the plotted best fit straight lines.

the "Index" represents the angular coefficient of the fitted lines. Differences in the "Ampli" factor, i.e. the amplitude of the power law and the intercept of the straight lines, can be understood statistically as an effect of the different number of objects at disposal in the two datasets.
In conclusion: for the first time Machine Learning techniques have been used to perform measurements of Weak Lensing quantities (e.g. the ESD) by using both photo-z in their punctual estimates and their PDF representations. We found that the trends are almost indistinguishable. Furthermore, despite the difference in the object number of the datasets, sampled by METAPHOR and by a SED fitter method (e.g. BPZ), we find very similar trends. ML techniques seem therefore capable to reproduce results obtained using SED fitting methods, using a small fraction of the objects (and hence of the information).
Due to the relevance of this result which goes against the widespread belief that PDFs are crucial to derive ESD's, much work is still needed before reaching a final conclusion.

References

Abdalla, F., Banerji, M., Lahav, O., and Rashkov, V. e. a. (2011). *MNRAS*, 417:1891.

Ahn, C. P., Alexandroff, R., and Allende Prieto, C. e. a. (2012). *ApJS*, 203:21.

Ahn, C. P., Alexandroff, R., and Allende Prieto, C. e. a. (2014). *ApJS*, 211:17.

Annis, J. T. (2013). *American Astronomical Society, AAS Meeting 221*.

Aragon-Calvo, M. e. a. (2015). 463. *MNRAS*, 454.

Arnouts, S., Cristiani, S., and Moscardini, L. e. a. (1999). *MNRAS*, 310:540.

Baran, S. and Lerch, S. (2016). Mixture emos model for calibrating ensemble forecasts of wind speed. *Environmetrics*, 27:116–130.

Baum, W. (1962). Problems of extra-galactic research. *in IAU Symposium*, Vol. 15:390.

Benitez, N. (2000). *ApJ*, 536:571.

Bilicki, M., Hoekstra, H., and Amaro, V. e. a. (2017). *A&A, arXiv:1709.04205*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bolzonella, M., Miralles, J. M., and Pello, R. (2000). *A&A*, 363:476–492.

Bonnet, C. (2013). *MNRAS*, 449:1043.

Bordoloi, R., Lilly, S. J., and Amara, A. (2010). *MNRAS*, 406(2):881.

Brammer, G. B., van Dokkum, P. G., and Coppi, P. e. a. (2008). *ApJ*, 686:1503.

Brescia, M., Cavuoti, S., D'Abrusco, R., Mercurio, A., and Longo, G. (2013). *ApJ*, 772:140.

Brescia, M., Cavuoti, S., Longo, G., and De Stefano, V. (2014b). *A&A*, 568:A126.

Brescia, M., Cavuoti, S., and Longo, G. e. a. (2014a). *PASP*, 126:783–797.

Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., and T., P. (2012). *MNRAS*, 421:1155.

Bruzual, G. and Charlot, S. (2003). *MNRAS*, 344:1000.

Byrd, R. H., Nocedal, J., and Schnabel, R. B. (1994). *Math.Program.*, 63:129.

Capozzi, D., de Filippis, E., Paolillo, M., D'Abrusco, R., and Longo, G. (2009). *MNRAS*, 396, 2:900–917.

Carrasco, K. and Brunner, R. (2013a). *MNRAS*, 432:1483–1501.

Carrasco, K. and Brunner, R. J. (2013). San francisco: Astronomical society of the pacific. *Astronomical Data Analysis Software and Systems XXII*, page 69.

Carrasco, K. and Brunner, R. J. (2014b). *MNRAS*, 442:3380–3399.

Carrasco Kind, M. and Brunner, R. J. (2014a). *MNRAS*, 438(4):3409–3421.

Cavuoti, S., Amaro, V., and Brescia, M. e. a. (2017). *MNRAS*, 465:1959.

Cavuoti, S., Brescia, M., De Stefano, V., and Longo, G. (2015c). Volume 39, Issue 1, 1 March 2015:45–71.

Cavuoti, S., Brescia, M., Longo, G., and Mercurio, A. (2012). *A&A*, 546:13.

Cavuoti, S., Brescia, M., and Tortora, C. e. a. (2015). *MNRAS*, 452, 3(a):3100–3105.

Cavuoti, S., Tortora, C., Brescia, M., and Longo, G. e. a. (2017a). *MNRAS*, 466:2039.

Coe, D., Benitez, N., Sanchez, S., Jeee, M., Bouwens, R., and Ford, H. (2006). *ApJ*, 132:926–959.

Coleman, G. D.and Wu, C.-C. and Weedman, D. W. (1980). *ApJS*, 43:393.

Colless, M., Dalton, G., and Maddox, S. e. a. (2001). *MNRAS*, 328:1039.

Collister, A. A. and Lahav, O. (2004). *PASP*, 116:345.

Connolly, A. J., Csabai, I., and Szalay, A. S. e. a. (1995). *AJ*, 110:2655.

Cover, T. M. and Hart, P. E. (1967). *IEEE Trans. Inf. Theory*, 13:21.

Davidon, W. (1991). *SIAM J. Optim.*, 1:1.

de Jong, J. T. A., Verdoes Kleijn, G. A., and Boxhoorn, D. R. e. a. (2015). *A&A*, 582:A62.

de Jong, J. T. A., Verdoes Kleijn, G. A., Erben, T., Hildebrandt, H., Kuijken, K., Sikkema, G., Brescia, M., B., and M., Napolitano, N. e. a. (2017). *A&A*, 604, A134:26.

Driver, S. P., Hill, D. T., and Kelvin, L. S. e. a. (2011). *MNRAS*, 413:971.

Dvornik, A., Cacciato, M., and Kuijken, K. e. a. (2017). *MNRAS*, 468:3251–3265.

Euclid, R. B. (2011). *ESA Technical Document, ESA/SRE(2011)12, arXiv:1110.3193*, Issue 1.1.

Feldmann, R., Carollo, C. M., and Porciani, C. e. a. (2006). *MNRAS*, 372:565–577.

Flewelling, H. A., Magnier, E. A., and Chambers, K. C. e. a. (2016). *arXiv:1612.05243v2*.

Gerdes, D. e. a. (2010). *AJ*, 715:823.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243–268.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics.

Hildebrandt, H., Arnouts, S., and Capak, P. e. a. (2010). *A&A*, 523:A31.

Hildebrandt, H., Erben, T., and Kuijken, K. e. a. (2012). *MNRAS*, 421:2355.

Hildebrandt, H., Viola, M., and Heymans, C., e. a. (2017). *MNRAS*, 465:145.

Hoyle, B., Rau, M. M., Zitlau, R., Seitz, S., and J., W. (2015). *MNRAS*, 449:1275.

Ilbert, O., Arnouts, S., and McCracken, H. J. e. a. (2006). *A&A*, 457:841.

Ilbert, O. e. a. (2009). *ApJ*, 690:1236.

Ivezic, Z. (2009). *American Physical Society, APS April Meeting*, May 2-5, W4.003.

Kilbinger, M. (2015). *Review article, arXiv:1411.0115v2*.

Kuijken, K. (2008). *A&A*, 482:1053.

Laureijs, R., Racca, G., and Stagnaro, L. e. a. (2014). *Proceedings of the SPIE*, Vol. 9143:id. 91430H.

Liske, J., Baldry, I. K., and Driver, S. P. e. a. (2015). *MNRAS*, 452:2087.

LSST Science Book, V. . (2009). *Science Collaborations and LSST Project, arXiv:0912.0201*.

Mandelbaum, R., Seljak, U., and Hirata, C. M. e. a. (2008). *MNRAS*, 386, 2:781–806.

Masters, D., Capak, P., and Stern, D. e. a. (2015). *ApJ*, 813,1:53.

Oyaizu, H., Lima, M., Cunha, C. E., Lin, H., and Frieman, J. (2008). *ApJ*, 689:709.

Pedregosa, e. a. (2011). *J. Mach. Learn. Res.*, 12:2825.

Planck Collaboration XVI, e. a. (2014). *A&A*, 571:A16.

Polletta, M. e. a. (2007). *ApJ*, 663:81.

Rau, M. M., Seitz, S., Brimioulle, F., Frank, E., Friedrich, O., Gruen, D., and B., H. (2015). *MNRAS*, 452:3710.

Robotham, A. S. G., Norberg, P., and Driver, P. e. a. (2011). *MNRAS*, 416:2640–2668.

Sadeh, I., Abdalla, F., and Lahav, O. (2015).

Sadeh, I., Abdalla, F. B., and Lahav, O. (2016). *PASP*, 128:104502.

Sánchez, C., Carrasco Kind, M., Lin, H., and Miquel, R., e. a. (2014).

Schlafly, E. F. and Finkbeiner, D. P. (2011). *ApJ*, 737:103.

Serjeant, S. (2014). *AJ*, 793,1:l10.

Silva, L., Granato, G. L., Bressan, A., and L., D. (1998). *ApJ*, 509:103.

Soo, J. e. a. (2017). *Morpho-z: improving photometric redshifts with galaxy morphology. submitted to MNRAS*.

Speagle, J. and Eisenstein, D. (2015). (4).

Tagliaferri, R., Longo, G., and Andreon, S. e. a. (2002). *ArXiv*.

Tanaka, M. (2015). *AJ*, 801:1,20.

Viola, M., Cacciato, M., and Brouwer, M. e. a. (2015). *MNRAS*, 452:3529.

Wittman, D., Bhaskar, R., and Tobin, R. (2016). *MNRAS*, 457:4005.

York, D. G. e. a. (2000). *AJ*, 120:1579.

List of publications

- Amaro, V., et al., “Statistical analysis of astroinformatics based photometric probability density functions through the KiDS-ESO-DR3 galaxies”, submitted to MNRAS.
- Amaro, V., et al., Astroinformatics, Proceedings of the International Astronomical Union, IAU Symposium, Volume 325, pp. 197-200
- Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
- S. Cavuoti, C. Tortora, M. Brescia, G. Longo, M. Radovich, N. R. Napolitano, V. Amaro and C. Vellucci Astroinformatics, Proceedings of the International Astronomical Union, IAU Symposium, Volume 325, pp. 166-172
- S. Cavuoti, C. Tortora, M. Brescia, G. Longo, M. Radovich, N. R. Napolitano, V. Amaro, C. Vellucci, F. La Barbera, F. Getman and A. Grado, MNRAS, 466,2039
- Cavuoti, S., Brescia, M., Amaro, V., Vellucci, C., Longo, G., Tortora, C., et al., 2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016 7849953
- de Jong, J.T.A., Verdoes Kleijn, G.A., Erben, T., Hildebrandt, H., Kuijken, K., Sikkema, G., Brescia, M., Bilicki, M., Napolitano, N.R., Amaro, V., et al., et al., 2017, A&A, accepted, DOI: https://doi.org/10.1051/0004-6361/201730747
- M. Bilicki, H. Hoekstra, V. Amaro, C. Blake, M. J. I. Brown, S. Cavuoti, J. de Jong, H. Hildebrandt, C. Wolf, M. Brescia, S. Brough, M. V. Costa-Duarte, T. Erben, K. Glazebrook, T. Jarrett, S. Joudaki, G. Longo, C. Vellucci, and L. Wang, 2017, “Photometric redshifts for the Kilo-Degree Survey Machine-learning analysis with artificial neural networks”, submitted to A&A
- N. Roy, N.R. Napolitano, F. La Barbera, C. Tortora, F. Getman, M. Radovich, M. Capaccioli, M.Brescia, S. Cavuoti, G. Longo, E. Puddu, G. Covone, V. Amaro, C. Vellucci, A. Grado, K. Kuijken, 2017,”Evolution of galaxy sizestellar mass relation from the Kilo Degree Survey”, submitted to MNRAS