Improvement of photometric redshift estimations with Machine Learning

Master thesis

Sara Herminia Navarro Umpiérrez

Supervisor: Aurelio Carnero, Instituto de Astrofísica de Canarias (IAC)
Supervisor: Francisco Kitaura, Instituto de Astrofísica de Canarias (IAC)
July 2023

Facultad de Física

A mis tutores
por estar ahí conmigo al pie del cañón.
A Guetón
por su paciencia y amor incondiconal.

Resumen

Mapear de forma precisa el Universo es crucial para comprender su estructura, evolución y composición a gran escala. Esto nos permite construir modelos cosmológicos más ajustados a la realidad. Por lo tanto, es de vital importancia contar con herramientas confiables que nos permitan obtener una gran cantidad de datos de manera eficiente. Un gran avance para la comunidad científica sería lograr predicciones en redshift fotométricos que compitan en precisión con las mediciones espectroscópicas. Las mediciones espectroscópicas tradicionales requieren un tiempo de observación considerablemente mayor y solo pueden realizarse en un número limitado de objetos a la vez. En contraste, el enfoque del redshift fotométrico, photo-z, ofrece la ventaja de un tiempo de observación reducido y la posibilidad de observar simultáneamente una gran cantidad de objetos.

En este estudio, nuestro objetivo principal ha sido encontrar métodos más precisos para la estimación de photo-z. Para lograrlo, hemos utilizado un código de aprendizaje automático que emplea un modelo de densidad de mezcla y distribuciones Gamma. Este modelo ha sido entrenado exclusivamente utilizando propiedades fotométricas de las galaxias como variables de entrada, y proporciona estimaciones de photo-z en forma de funciones de distribución de probabilidad. El entrenamiento se ha realizado utilizando valores de z espectroscópicos conocidos. Este modelo, denominado γ-MDN, tiene la ventaja de estar restringido a valores reales positivos, lo que lo hace ideal para la predicción de distancias. Además, al ser un modelo de densidad de mezcla, nos permite obtener funciones de densidad de probabilidad con una abundante información estadística, destacando especialmente su capacidad para abordar problemas con cierta multimodalidad, como la incertidumbre a lo largo de la línea de visión en la medición de distancias.

El punto más novedoso de este trabajo consiste en buscar formas de mejorar los resultados de γ-MDN mediante la utilización de información de la estructura a gran escala de la red cósmica, sin depender de asumir una cosmología en particular. Para lograr esto, hemos desarrollado el modelo Cluster-GMDN, que realiza un mapeo del cielo en píxeles y utiliza histogramas de densidad en redshift para generar una función de densidad de probabilidad de grupo del fondo cósmico a lo largo de la línea de visión de la galaxia. Al multiplicar esta función de densidad de probabilidad del fondo con la obtenida a través de γ-MDN, es posible mejorar la precisión en la estimación de photo-z. Aunque en este trabajo se presenta únicamente una primera prueba de este método, ya se observa su gran potencial.

Con el fin de demostrar su potencial, estos dos modelos han sido aplicados a los datos de dos destacados proyectos de investigación, el Dark Energy Survey (DES) y el catálogo 2MASS Photometric Redshift (2MPZ). La elección del primero se debe a su relevancia en el estudio de la energía oscura, mientras que el segundo se seleccionó debido a que sus predicciones fotométricas no incluyen funciones de distribución de probabilidad y es un proyecto colaborativo que se espera que perdure, al menos 10 años más, hasta que ser reemplazado por una iniciativa de mayor.

Con DES, hemos probado y validado el funcionamiento de γ-MDN. Inicialmente, el método se diseñó para tener en cuenta la forma de las galaxias al realizar predicciones, pero decidimos generalizar el código y utilizar solo las magnitudes g, r, i y z de DES, junto con sus colores, como variables de entrada. Los resultados obtenidos se compararon con el modelo de referencia de DES, conocido como *DNF*. Observamos que nuestro modelo es bastante competente, igualando o incluso mejorando la calidad de las predicciones en algunos rangos de redshift. La calidad de las predicciones se evaluó utilizando diversas métricas, como el sesgo en Photo-z, el ancho del percentil 68 del sesgo σ-68, la tasa de valores atípicos y la desviación estándar.

Luego, aplicamos γ-MDN para obtener las funciones de distribución de probabilidad (PDFs) de la submuestra de DES llamada *MagLim*, que no cuenta con mediciones espectroscópicas y se creó con el objetivo de obtener restricciones cosmológicas. A través de este enfoque, obtuvimos una amplia base de datos de aproximadamente 10 millones de galaxias con nuestros propios resultados. Utilizando estos datos, creamos un histograma de densidad que el modelo Cluster-GMDN emplea para generar la PDF del fondo cósmico del grupo. Comparando los resultados de Cluster-GMDN con los obtenidos por γ-MDN, observamos una mejora considerablemente razonable de los resultados de las métricas en ciertos rangos de redshift, principalmente para reshifts por encima de z=0.6. En un estudio individual de ciertas galaxias, se pudo apreciar que existen PDFs multimodales en las que un máximo local, que no es la moda, coincide con el valor de la posición verdadera de la galaxia, y emplendo Cluster-GMDN se ajustaron las probabilidades permitiendo mejorar la precisión de los resultados, dando como moda el pico correcto.

Además, se exploró el potencial de este modelo como una herramienta para la creación de submuestras más limpias, excluyendo valores atípicos. En este sentido, se obtuvieron resultados sólidos y satisfactorios.

Encontramos varios desafíos al evaluar el rendimiento de γ-MDN en la aplicación a 2MPZ, ya que no logramos replicar la muestra de entrenamiento descrita en sus referencias. Sin embargo, podemos concluir que los resultados de nuestro modelo son competitivos en comparación con *ANNz*, el modelo que utilizan. Además, nuestros resultados aportan PDFs confiables,

en lugar de un único valor como estimación de photo-z como devuelve *ANNz*. Sin embargo, la implementación del modelo Cluster-GMDN no mejoró las estimaciones de photo-z. Creemos que esto puede deberse a que el catálogo de 2MPZ tiene limitaciones en cuanto a los redshifts, lo que implica una menor importancia del agrupamiento ya que hay menos densidad de galaxias en la línea de visión.

En conclusión, el modelo Cluster-GMDN presenta resultados muy prometedores y un gran potencial como método para la mejora de estimaciónes de photo-z, a pesar de que en este trabajo solamente se presenta un primer testeo del método y un análisis global de sus resultados. Se ha observado que presenta mejores resultados para muestras de galaxias amplias y extensas en redshift. Sin embargo, es necesario realizar un estudio más exhaustivo del potencial de este modelo, que incluya análisis detallados sobre si existen tipos específicos de galaxias que puedan brindar mejores o peores resultados. Explorar estas posibilidades nos permitirá afinar el modelo y comprender mejor su desempeño en diferentes situaciones.

Contents

1 Introduction

Since time immemorial, human beings have sought to understand the nature of the world around them, from their immediate surroundings to the deepest mysteries of the Universe. In this quest, cosmology emerges as the scientific discipline that explores the laws that govern the Universe as a whole, its origin, and its evolution. From the ancient Greek philosophers who contemplated the stars in search of answers, to the scientific and technological advances of the modern era, cosmology has undergone an incredible evolution. It is in the 20th century that this discipline has experienced a true revolution, thanks to the discovery of the accelerated expansion of the Universe, attributed to the existence of dark energy. However, despite the great advances made, cosmology still presents numerous challenges and unanswered questions. The nature of dark matter and dark energy remains a mystery, as does the pursuit of an increasingly precise determination of cosmological parameters.

In order to answer these questions, it is crucial to map our Universe to better understand its evolution, structure, and large-scale composition. This mapping enables us to construct cosmological models that best fit reality. Therefore, it is essential to obtain accurate measurements of the distances of as many galaxies as possible. Redshift, z, is a fundamental tool in astrophysics, that allows us to measure the distances to galaxies. It emerges as a consequence of the expansion of the Universe, causing the energy spectrum emitted by distant galaxies to be shifted towards longer wavelengths, compared to nearby galaxies. Therefore, the greater the redshift, the greater the distance to the galaxy. For small values of z, the redshift is approximately related to the distance through Hubble's law. One way to calculate z is by comparing the wavelength of an emission or absorption line from the galaxy's spectrum, λ_0, with the position of the same lines in the laboratory, λ, as follows:

$$z = \frac{\lambda_0 - \lambda}{\lambda} \propto d \tag{1}$$

Calculating redshift through spectroscopy is a highly accurate and reliable method. However, this process requires a lot of time and resources since it requires spectroscopic observations with long exposure times per object. Due to this, not all galaxies for which we have images have observed spectra, which limits the amount of available information.

In contrast, by using CCD cameras, we can obtain images of thousands of galaxies simultaneously. Using different broadband filters, we can determine the photometric properties of these galaxies and reconstruct their spectra based on this information. With the resulting broad-band spectra, it is possible to calculate a photometric redshift (photo-z), which, although it is much faster to determine, has a much higher uncertainty than the spectroscopic redshift due its lower spectral resolution. Therefore, the estimation of photometric redshift is a powerful tool that provides a large statistical sample at the cost of individual precision.

There are numerous international projects that have invested significant efforts in scanning the entire sky and thereby increasing the amount of photometric data. Examples of such projects include The Kilo-Degree Survey (KIDS), Hyper Suprime-Cam Subaru (HSC) and the Dark Energy Survey (DES), which aim to shed light on the nature of dark matter and dark energy, among other objectives. These surveys have greatly contributed to our understanding of the Universe, but their capabilities will soon be surpassed. The upcoming Legacy Survey of Space and Time (LSST) is expected to outshine all these previous surveys, generating an unprecedented amount of data within just a few months of operation [Željko Ivezić et al., 2019]. With such a vast amount of data, it is necessary to improve the techniques for obtaining photometric redshifts and thus obtain much more precise results.

Significant progress has been made in addressing the limitations of photometric redshifts (photo-z) through the application of machine learning techniques. One of the most promising approaches is the use of training or regression methods. These methods involve using a set of galaxies with known spectroscopic redshifts, the true target variable, to train a machine learning algorithm. This method takes into account the photometric properties of galaxies, such as their colors and magnitudes, as input features. Subsequently, the trained algorithm is then applied to the galaxy set of interest to estimate their redshifts.

Currently, there are several highly competitive methods that map the redshift using photometric properties. For instance, DES uses the *Directional Neighbourhood Fitting, DNF* method De Vicente et al. [2016], while the 2MASS Photometric Redshift Catalog (2MPZ) employs an Artificial Neural Network called *ANNz* Collister and Lahav [2004]. In this study, we use the γ-MDN method proposed by Cerdán [2020]. This approach is based on a mixture density model that uses Gamma distributions instead of Gaussians. The use of Gamma distributions makes it particularly suitable for problems where the target variable is constrained to the space of positive real numbers, such as in the case of distance measurements.

In particular, the γ-MDN model, being a mixture density model, can be valuable for probability distributions that exhibit multimodality. This is especially relevant in cases where two galaxies with different redshifts display similar photometric properties. Due to the loss of information, the regression model may yield multiple solutions for the same combination of

photometric features. Therefore, the main advantage of the γ-MDN model lies in its ability to provide a reliable probability density function (PDF), offering a comprehensive statistical description of the target variable. This feature distinguishes it from other models that only provide a single output value or employ different types of distributions.

As a result, the γ-MDN model shows great promise in tackling problems characterized by multimodal PDFs. The information provided by the redshift PDFs opens up possibilities for complementing and enhancing its performance through additional data. For instance, using information from the underlying comic web could help to improve the γ-MDN z estimate, a novel approach not previously attempted.

In this work we present for the first time a method to use the information from the underlying cosmic web to improve the redshift determination. We first generate the galaxy probability distribution of the background along the line of sight of each galaxy, using the photo-z estimates from γ-MDN, applied to the entire background sample (ensuring self-consistency between the individual galaxy and the background density field). The resultant background PDF is then multiplied by the PDF of the target galaxy, assigning a different statistical weight to the local maxima of the γ-MDN estimate, or even correcting the photo-z estimate in the case of unimodal distributions. This method has been named "Cluster-GMDN" in this study, and although further refinement is necessary, it holds great promise as it does not impose any pre-existing cosmology. In contrast, other methods that use background information to improve redshift measurements, such as BAO reconstruction, rely on assuming a specific cosmological model.

2 Objectives

Our main interest is to find ways to improve current photo-z measurements in order to fully exploit the information we have from photometric surveys. This includes exploring the potential of the existing γ-MDN model and seeking a new method that leverage the information from the underlying cosmic web to refine photo-z estimates. To achieve this, the main objective has been subdivided into the following specific goals:

- Verify and confirm the results obtained in the previous study completed by Cerdán [2020] on the spectroscopic sample of DES using the γ-MDN method.
- Apply the γ-MDN model to the photometric subsample of DES called *MagLim* [Porredon et al., 2021], in order to generate alternative photo-z to the ones used by the collaboration, including the computation of probability distributions for each galaxy in the sample.
- Apply the γ-MDN model for the first time to the 2MASS Photometric Redshift catalog (2MPZ) data, generating probability density functions (PDFs) for each galaxy, which are not provided by the photo-z method used in the survey.
- Apply the first tests of the method Cluster-GMDN on both the DES and 2MPZ data. As already mentioned, this method uses additional information from the cosmic web to enhance the precision of photo-z estimates, using the statistical information obtained with γ-MDN on a background sample. In addition to improving photo-z measurements, we explore if the method can be used to improve the definition of galaxy samples, i.e.: if the method is able to identify galaxies within a sample for which the photo-z estimate is not reliable enough, decreasing the amount of outliers from a sample.

This work is structured as follows: Sections 3 and 4 present the theoretical framework of the study. In Section 3, we define the main statistical concepts that form the γ-MDN model, as well as the metrics used to evaluate the performance of the neural network. In Section 4, the Cluster-GMDN method is introduced, explaining its main motivation and foundation. Sections 5 and 6 serve as the methodology. Section 5 provides a general description of the surveys used. Section 6 describes the model's implementation, specifying the main parameters used. The results obtained from applying the model are presented and discussed in Section 7. Finally, in Section 8, a concise summary of the main results, conclusions, and future prospects is provided. Additionally, an Appendix is included containing additional figures that were not included in the main body of the work, but provide complementary information.

3 Analytical Framework

In this section, we introduce the γ-MDN model, which will be employed for photo-z predictions throughout this study. This model was originally developed in the Master's Thesis by Cerdán [2020]. We begin by defining the fundamental statistical elements that make up the γ-MDN model, the mixture density network model (MDN), and the γ distribution. Then, I will delve more into the γ-MDN itself and explain the main metrics that we use to calibrate the model.

3.1 Mixture Density Network

It is a feed-forward artificial neural network whose output layer elements correspond to the parameters of a mixture density model that describes the probability density distribution of the target variable. It is an ideal model for problems where the target variable is continuous with certain multimodality or is multivalued, because provide a complete description of the conditional probability distribution of the target variable, given the input features. Other conventional neural network provides the mean as the only statistical variable, which is not enough to fully describe the statistical properties of the target variable and gives wrong values in multimodality problems.

As a formal definition, we have that the probability density function of the target variable conditioned on the input features, $p\left(y \mid \mathbf{x}\right)$, is going to be equal to the weighted sum of simpler distributions:

$$p\left(y \mid \mathbf{x}\right) = p\left(y \mid \pi\left(\mathbf{x}\right), \theta\left(\mathbf{x}\right)\right) = \sum_{i=1}^{K} \pi_i\left(\mathbf{x}\right) \phi_i\left(y \mid \theta_i(\mathbf{x})\right) \tag{2}$$

where K are the number of components of the mixture distribution, π represents the mixing coefficients, i.e., the weights for each component of the mixture. And $\phi_i\left(y \mid \theta_i(\mathbf{x})\right)$ are the conditional probability density functions corresponding to a particular distribution with parameters θ [Cerdán, 2020]. Normally, MDNs make use of Gaussian distributions as probability density function (Figure 1).

This model outputs, for each of the K distributions, the mixture parameter and the θ parameters corresponding to the probability distribution used, which for a normal distribution would be the mean and variance.

However, the mixture model parameters are not obtained directly from the output layer of the neural network, as they are subject to constraints. The mixture parameter, π_i, must satisfy the condition of being positive and the sum must be equal to one, $\sum_{i=1}^{K} \pi_i = 1$, in order to transform real values into a probability distribution. This is achieved by applying a *softmax* transformation to K elements of the output vector z [Bishop, 1994]:

$$\pi_i = \frac{exp(z_i^{\pi})}{\sum_{j=0}^{K} exp(z_j^{\pi})} \tag{3}$$

The constraints on the remaining coefficients will depend on the type of distribution used. In the next section, we will see the constraints for the Gamma distribution.

Conventional MDNs use normal (or Gaussian) distributions for the mixture model due to their formal simplicity and the fact that any given distribution can be approximated by a mixture of normal distributions [Kuruoğlu et al., 1998]. However, this is only true for random variables whose probability density functions are defined over the entire real line (the real numbers, $\mathfrak{R}$). For certain applications where the target variable is strictly positive, such as age or distance measurements, this type of distribution are not the most convenient. The conditional probability distribution should be approximated as a mixture of distributions defined on the positive real numbers [Cerdán, 2020]. This is where the Gamma distribution will be useful.

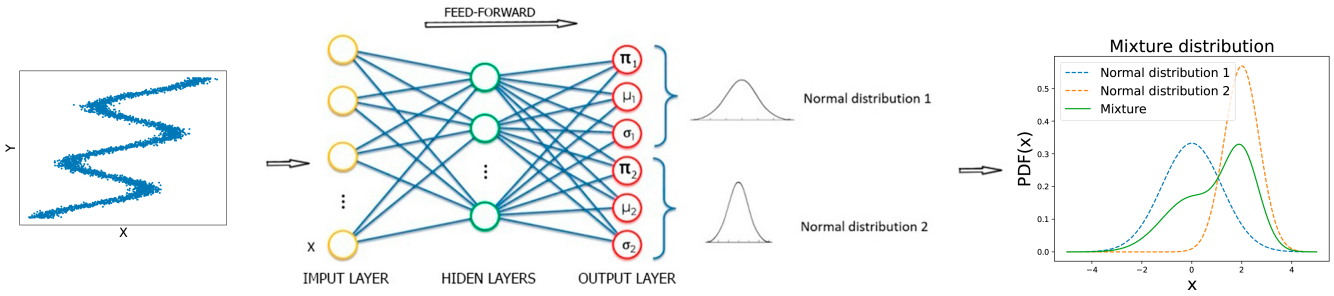

Figure 1: Example implementation of the architecture of a Mixture Density Network (MDN). The features input vector **x**, which belongs to a multimodal or multivalued function, is introduced into a feed-forward ANN. In this example, the output parameters of the ANN correspond to a mixture density model of Gaussian distributions, which are used to construct the probability density function (PDF) of the target variable, conditioned on the input features.

3.2 Gamma distribution

The Gamma distribution is a continuous exponential-type probability distribution with two positive parameters, α and β. It is defined on the positive real line ($\Re^+$), making it ideal for modeling skewed right positive random variables [Cerdán, 2020, Thom, 1958]. The probability density function has the following form, where Γ is the Gamma function:

$$G(x;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{4}$$

The parameters involved are α, also known as the shape parameter, and β, known as the inverse scale or rate parameter. α characterizes the distribution shape, in other words, its skewness, with a higher α resulting in a more symmetric and more concentrated distribution around its mean, and a value less than one indicating a highly skewed and divergent distribution. β characterizes the variance of the distribution, with higher values leading to narrower distributions (Figure 2). Both parameters must be positive. With these parameters, relevant statistics such as the mean (μ) and variance (σ^2) can be calculated:

$$\mu = \frac{\alpha}{\beta} \quad ; \quad \sigma^2 = \frac{\alpha^2}{\beta} \tag{5}$$

Below is the cumulative distribution function (CDF) expression, which we will use later in the metrics, in Section 3.4:

$$F_G(x;\alpha,\beta) = \int_0^x G(x;\alpha,\beta)\, dy = \frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)} \tag{6}$$

where $\gamma(\alpha,\beta x)$ is the lower incomplete Gamma function [Soch, 2020].

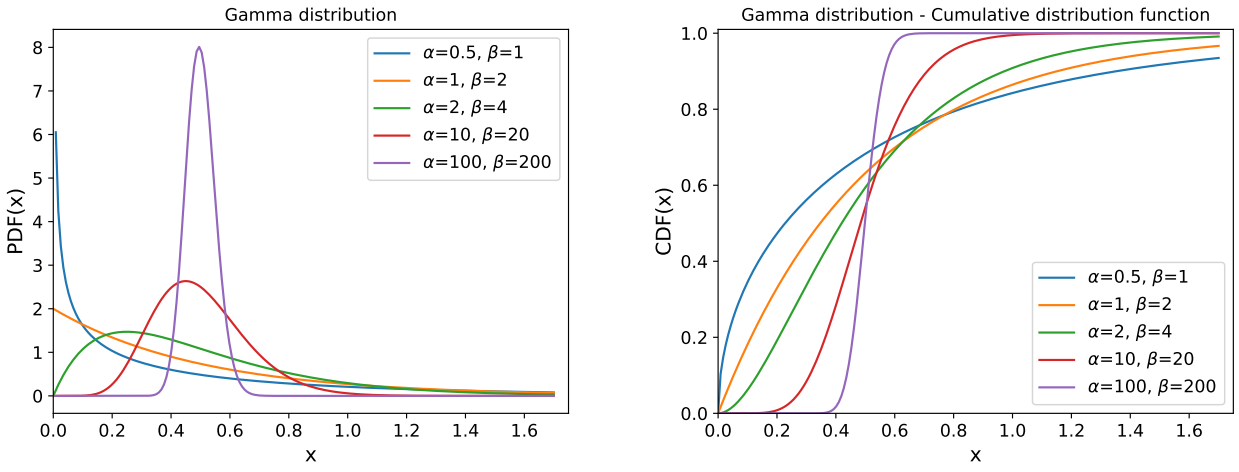

Figure 2: PDF and CDF for Gamma distributions using different values for the shape α and rate β parameters

3.3 γ-MDN

γ-MDN arises from using Gamma distributions in the MDN model, instead of Gaussians, which allows us to model problems with multimodality in positive real numbers (in our case, the target variable, z, is strictly positive real) and providing reliable probability density functions. Therefore, this will be the model we will use to approximate the conditional probability distribution of the galaxy distance.

The output parameter dimensions of the model will be $3xK$ parameters, representing the variables to be predicted by the γ-MDN model: the mixture weights (π_i), as well as the shape and rate parameters (α_i, β_i) for each of the K Gamma distributions.

All of these parameters are subject to constraints. The mixing parameter takes the form for the conventional case of MDN that we saw earlier (in Section 3.1), normalizing real values into a probability distribution, and therefore their corresponding Artificial Neural Network outputs are subjected to a *softmax* equation [Bishop, 1994]. For the α and β outputs, we use a *softplus* transformation, ensuring that they are positive for numerical reasons [Cerdán, 2020]:

$$\alpha_i = \ln\left(exp\left(z_i^{\alpha} + 1\right)\right) + a \tag{7}$$

$$\beta_i = \ln\left(exp\left(z_i^{\beta} + 1\right)\right) + b \tag{8}$$

The *softplus* transformation is used instead of an exponential function because it provides a more numerically stable way to enforce a positive constraint [Wiemann et al., 2021].

3.4 Metrics

In this section, we define the metrics used to assess the calibration and sharpness of the predictions generated by the γ-MDN model. These metrics include the log-likelihood, PIT (Probability Integral Transform), and RI (Reliability Index).

The log-likelihood primarily evaluates the predicted probability density at each observation, without considering the shape of the distribution. Therefore, while the log-likelihood is useful, it may not provide a complete assessment of the model's performance.

To ensure the adequacy and reliability of our conditional mixture distribution model in terms of calibration, we will focus on the PIT and RI. These metrics enable us to verify if the set of observed data is distributed according to the predicted distributions.

Examples of applying these metrics for DES results can be found in Cerdán [2020]. In this work we only present the results for the new estimates applied to the 2MPZ survey. Examples of these metrics are shown in Figure 5 and Figure 6.

3.4.1 The negative log-likelihood

As loss function we use the negative logarithm of the likelihood of the predicted distribution, which will give us an estimate of how likely the observed data is to be produced by the estimated parameters of the model. It is widely used today in many research areas, such as cosmology, to deal with the complexity of large datasets [Dodelson and Schmidt, 2021]. We use the negative logarithm of the likelihood because it is more convenient to minimize a function rather than maximize it. Therefore, to obtain the best possible model, we minimize this function.

Below is the expression for the negative log-likelihood for the specific case of the γ-MDN model:

$$\ell = -\log L = -\sum_{n=1}^{N} \log\left\{p\left(y_n \mid x^n\right)\right\} = -\sum_{n=1}^{N} \log\left\{\sum_{k=1}^{K} \pi_k\left(x^n\right) G\left(y_n; \alpha_k\left(x^n\right), \beta_k\left(x^n\right),\right)\right\} \tag{9}$$

3.4.2 Probability Integral Transform

The PIT is calculated as the value of the cumulative distribution function (CDF) reached at the actual target value yn [Cerdán, 2020]. For a mixture of Gamma distributions, it has the following form:

$$p_n = \sum_{i=1}^{K} \pi_i F_G\left(y_n; \alpha_i \beta_i\right) \tag{10}$$

PIT transforms the predictions into the range [0,1] and the shape of the histogram give us information about if the predicted distributions are well calibrated. If they are, then the PIT values should follow a uniform distribution and the predicted distributions closely follow the real distribution of the data. In the case where the distribution is not uniform, we may encounter various patterns in the histograms: a hump-shaped histogram indicates an overdispersed predictive distribution, while a U-shaped histogram suggests that the predictive distributions is too narrow. A triangle-shaped histograms typically means the predictive distribution is biased [Gneiting et al., 2007].

3.4.3 Reliability Index

Using the PIT histograms to evaluate the calibration of predictions during the training phase of a neural network becomes impractical due to the large number of histograms that need to be assessed. For this, the Reliability Index is useful, which provides a cumulative measure of how much the PIT histograms deviate from uniformity, and it is calculated as follows [Cerdán, 2020]:

$$RI = \sum_{b}^{B} \left\| k_b - \frac{1}{B} \right\| \tag{11}$$

where k_b defines the relative number of observations in bin b of the PIT histogram, and capital B is the number of bins used. A good prediction is characterized by values that are as close to 0 as possible, with a perfect prediction having an RI of 0.

4 Clustering of the Background in the galaxy line of sight

In this section we present the Cluster-GMDN model. This model aims at improving the photo-z estimate by using information about the cosmic web distribution in the background of the target galaxy, considering that the probability in redshift increases where there is a density peak (in a galaxy cluster, for example) and decreases where the density is lower (in voids and filaments).

The PDF generated by the γ-MDN for the background galaxies is used to generate a general probability distributions for each galaxy along the line of sight. In general, the technique we have developed involves a discretization of the celestial sphere into pixels, and the generation of histograms of galaxy density distributions for each pixel using the photo-z values obtained with γ-MDN itself, for the survey under analysis. With this histograms, a PDF is constructed for each pixel, which is then modified to 'increase the contrast' of the galaxy density, highlighting the most important density peaks. To account for the fact that a galaxy might sit at the border of a galaxy cluster centered at a neighbouring pixel, we create a weighted sum of the PDFs coming from contiguous pixels, giving more weight to the PDFs of the closest pixels.

Once the background PDF is calculated, we multiply with the target PDF of the galaxy to create a new PDF (Cluster-GMDN PDF), corrected by the distribution of the background cosmic web. This method intends to reduce the number of outliers in multimodal PDF distributions and adjust the maximum probability in the unimodal cases. More details about its implementation are given in Section 6.3.

The novelty of the approach, which makes it innovative, is the fact that it does not rely on any prior cosmology assumptions. Although it is not the first time that a model uses the information of overdensities in the large-scale structure of the Universe to make corrections in redshifts, these models assume a specific cosmology that can lead to incorrect values or biases. The most important of these models is the technique called "BAO reconstruction" [Padmanabhan et al., 2012]. This method aims at improving the measurement of the baryonic acoustic oscillations (BAO) scale used as a standard ruler. To achieve this, it mitigates distortions produced by nonlinear gravitational interactions in the distribution of galaxies by estimating the motion of tracers under the background cosmic web and reversing it. The drawback of this model is that must assume a fiducial cosmology to convert positions and redshifts into distances [Sherwin and White, 2019]. Incorrect assumptions about the underlying cosmology result in variations in the shape of the BAO, which might have a significant impact on the precision of the BAO scale, specially for future surveys like DESI and LSST [Sherwin and White, 2019].

In this work, in Section 7, we unveil for the first time the application of the Cluster-GMDN to a real set of galaxies from DES and 2MPZ. As the model is still in the testing phase, further evaluations and refinements are required to identify the optimal operating conditions. For example, some aspects not yet explored are how the model performs for different galaxy types or how to group along the line of sight considering projection and evolutionary effects. However, we anticipate that this model will enhance distance estimation accuracy and mitigate systematic errors in the future.

5 Sampled Galaxy Surveys

In this section we describe the two surveys used in this study, DES and 2MPZ. DES is by far the largest and more homogeneous photometric survey to date, although will soon be surpassed by the much larger LSST survey. On the contrary, 2MPZ is expected to remain as the leading survey at very low redshifts for at least another 10 years, before being replaced by other surveys. This makes it an interesting survey to focus on for enhancing its photo-z measurements.

5.1 DES

The Dark Energy Survey (DES) is a large-scale international project with the main objective of studying and characterizing the properties of dark energy and dark matter, while also seeking to obtain precise measurements of cosmological parameters. To achieve this goal, several methods have been employed, including the study of the number of galaxy clusters as a function of redshift , weak gravitational lensing, BAO and the Hubble diagram of Type Ia supernovae [Ponce Aguilar, 2015]. Additionally, the project has led to advances in other fields apart from cosmology. Some examples are the discovery of "hot" transneptunian objects, a survey of dwarf satellite galaxies around the Milky Way and the detection of a large number of quasars at very high redshifts, among other discoveries [Abbott et al., 2016].

To carry its study, the DES project requires photometric measurements of large quantities of objects. DES observations began in 2013 and concluded in 2019 after a 6-year campaign, during which galaxies with redshifts up to $z \approx 1.4$ were observed. DES covered approximately 5000 square degrees of the south Galactic cap using the Dark Energy Camera (DECam) installed on the 4-meter Blanco Telescope at the Cerro Tololo Inter-American Observatory (CTIO) in Chile [Rosell et al., 2022]. It uses five broad-band filters from the visible to near-infrared, (namely g, r, i, z, Y) achieving a depth of approximately 24 magnitudes at a signal-to-noise ratio of 10 [Abbott et al., 2021].

DES has made two important data releases: data release 1 (DES DR1), which was based on the first three seasons of observations, and DR2, which increased the total number of cataloged objects to approximately 700 million, making it the largest photometric data set to date at the achieved depth and photometric precision [Abbott et al., 2021]. Our work is based in the DR1 data.

The main photo-z estimator used in DES is the Directional Neighborhood Fitting (*DNF*) algorithm. *DNF* is a machine-learning algorithm based on the nearest neighbor approach, called neighborhood fitting (NF), but it takes into account a 'directional neighborhood' instead of Euclidean. Where the authors of the model define the directional neighborhood as the product of the Euclidean and angular neighborhoods: $DN = D^2 \sin \alpha^2$, with D representing the Euclidean distance and α denoting the angle between two multi-magnitude vectors in the feature space. In short, *DNF* uses a training sample of galaxies with photometric measurements to calculate a photo-z hyperplane that best fits the directional environment of a particular photometric galaxy. This hyperplane is used as a prediction function to estimate the redshift (photo-z) of that galaxy. The main idea is that by considering only nearby neighboring galaxies in the training sample, a simple linear fit can be made to predict the redshift more accurately [De Vicente et al., 2016].

5.1.1 DES Spectroscopic sample

The dataset we have used to train and validate the model consists of a sample of 387,889 galaxies observed by DES in their Y3 Gold release, for which we have their photometric properties and spectroscopic measurements. This dataset is a good reproduction of the one used by DNF to train and validate their results. As shown in Figure 3, the dataset extends up to a redshift of 1.4, with a median spectroscopic redshift of $z = 0.51$. To apply the model, the dataset has been randomly split into 80% for training and 20% for testing. The chosen input features for the neural network are, firstly, the apparent magnitudes in the different filters used by DES, except for the Y filter, i.e. the g, r, i and z filters. Secondly, the color indices, which represent the differences between the magnitudes of the different filters (g-i, i-r, etc.), and provide us information about certain physical characteristics such as age or metallicity.

We decided to remove the Y filter from the input characteristics because it overlaps with the z filter and does not contribute significantly to new information. It was observed in Cerdán [2020] that training the neural network with the Y filter did not result in a noticeable difference. The spectroscopic redshift values of these galaxies are considered as the true target values, against which we will compare our predictions of photo-z and evaluate our model.

To improve the quality of the training and test samples, the samples were cleaned by applying various filters. Galaxies with extremely high redshift values were removed, discarding those that met the condition of $z > 2$. Additionally, all galaxies

with negative errors in their magnitudes were excluded. These restrictions resulted in the removal of only a small number of galaxies, leaving a training set with 300,833 galaxies and a test set with 79,484 galaxies.

5.1.2 MagLim sample for Cluster-GMDN

MagLim is a subset of $\approx$ 10 million galaxies from the DES Y3 Gold data, sharing the same specified photometric characteristics as the spectroscopic sample. It was created by Porredon et al. [2021] with the aim of finding the most optimal sample of lens galaxies to obtain cosmological constraints from galaxy clustering and galaxy-galaxy lensing measurements [Chiu et al., 2023].

The sample goes from z=0.2 to z=1.2 (photo-z's calculated with *DNF* [Porredon et al., 2021]). It has been constructed with a magnitude cut that varies linearly across redshift, following the form i< $a \times z_{\rm phot} + b$, where a and b are arbitrary numbers and $z_{\rm phot}$ is the *DNF* estimate. This sample does not have a fixed magnitude limit, as is commonly done, in order to ensure the inclusion of brighter galaxies at lower redshifts while progressively incorporating fainter galaxies as redshift values rise. This avoids an excessive inclusion of less luminous galaxies at lower redshifts, improving the precision of photo-z estimates [Porredon et al., 2021]. Additionally, we apply the same filters as for the spectroscopic sample, selecting galaxies with positive magnitude errors.

Our aim with Maglim is to conduct our own alternative photo-z estimations for the entire dataset, generating as well redshift probability distributions. We use this sample to construct the background density distributions in the Cluster-GMDN model, applied over the test spectroscopic sample.

5.2 2MPZ

The Two Micron All Sky Photometric Redshift catalog, 2MPZ, was born out of the need for a large 3D catalog that extends beyond the local universe, covering the entire sky and providing complete redshift information. 2MPZ is a combination of several photometric catalogs, including 2MASS, WISE and SuperCOSMOS data. Each catalog contributes to different wavelengths: J, H, and Ks in the near-infrared from 2MASS; W1 and W2 in the infrared from WISE; and B, R, and I in the optical from SuperCOSMOS. To estimate the photometric redshifts, 2MPZ employed an Artificial Neural Network (ANN) known as *ANNz*. This network learns the relationship between photometry and redshifts by using an appropriate training set of galaxies with known spectroscopic redshifts [Bilicki et al., 2013].

For the study of the 2MPZ survey, we have used the sample referred as the "final sample" in Bilicki et al. [2013]. This sample consists of 934,175 galaxies with redshifts up to 0.3 and a median redshift of z = 0.070 (Figure 3). It includes the deredened B,R,I,J,H,K,W1,W2 magnitudes, as well as the photo-z predictions from *ANNz*, and spectroscopic values for approximately 311,229 galaxies.

We attempted to reproduce the training and testing sample of 2MPZ described in Bilicki et al. [2013] for galaxies with spectroscopic measurements. However, the dataset they provide is not the training and the testing sample they used, but rather the final sample to which they applied their *ANNz* method. As a result, we encountered several challenges. The main issues were that the provided data is truncated at magnitudes K<13.9, leading them to use galaxies with fainter magnitudes for training than us. Another difference is that they used the magnitudes W1 and W3 from WISE for training, while they provide only W1 and W2. These discrepancies may have influenced the comparison of results between γ-MDN and *ANNz* in Section 7.2. Despite this, we decided to proceed with the comparison, and the same ratio between training and testing was selected, with a ratio of 1:2 respectively. The redshift distribution for the spectroscopic 2MPZ sample can be seen in Figure 3.

The input features used by the neural network are all magnitudes provided by 2MPZ: J, H, K, W1, W2, B, I, and R, as well as their color indices. Although in Bilicki et al. [2013] they only use the color W1-W3 for training, we make use of all available color combinations.

The data has been cleaned using the restrictions employed in Bilicki et al. [2013], which include the removal of galaxies with a spectroscopic redshift $z_{\rm spec}$0<0.003, the removal of galaxies with high magnitude errors δmag>0.2, and galaxies with reddening E(B-V) > 0.25. These filters resulted in training and testing samples of 100,916 and 202,619 galaxies, respectively.

We use the complete sample, comprising galaxies with both spectroscopic and non-spectroscopic measurements, to construct the background density distribution in the Cluster-GMDN.

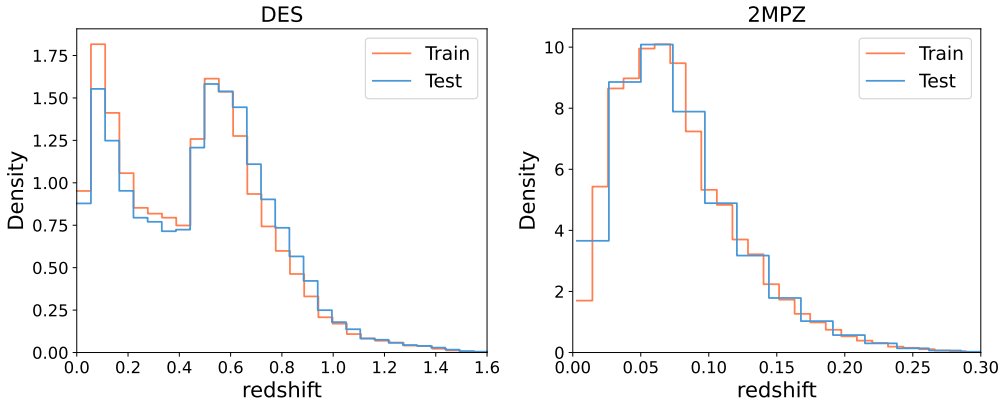

Figure 3: Spectroscopic redshift distributions of the train (orange) and test (blue) samples for the DES and 2MPZ datasets used in this work.

6 Model implementation

In this section, we describe the implementation of γ-MDN to DES and to 2MPZ data. In the case of DES, it is based on the previous work from [Cerdán, 2020], but with modifications that allow for the generalization to any other survey relying solely on magnitudes and colors. Likewise, the training validation is also shown for the 2MPZ case, as it has not been previously applied to this survey, making it a novel aspect of this work. Furthermore, it is worth noting the importance of applying γ-MDN to 2MPZ to obtain photo-z PDFs, not generated before for this survey, highlighting the importance of this work as it provides valuable information for the scientific community.

Finally, we implement the Cluster-GMDN method both in the DES and 2MPZ data. For the first case, we will use the MagLim sample described above to create the background cosmic web distribution, while for the 2MPZ case, we will use all the photometric sample available to characterize the background.

6.1 Data preprocessing

Some data processing has been performed in advance, following the guidelines used by Cerdán [2020] in his master's thesis.

It was observed that the input features exhibit multicollinearity, meaning that two or more features are highly linearly related to each other. This can slow down the convergence of back-propagation learning methods due to factors such as the presence of degeneracies. To solve this, certain data preprocessing is required. In this case, principal component analysis (PCA) was used to calculate a new set of variables, the principal components (PCs), that contain the same information as the original features but expressed in a more practical way. PCs are, by construction, uncorrelated and represent a change of basis that is constructed to point in orthogonal directions where the data varies most (Figure 4, right panel).

From now on, when we refer to the input features of the neural network, instead of corresponding to the magnitudes and colors, we are actually referring to the result of applying PCAs. As shown in the left panel of Figure 4, for the case of 2MPZ, we have practically 100% of the information contained in the 36 magnitudes and colors in the first 8 principal components, significantly reducing the input variables. However, we will use the complete set of principal components as input variables simply to avoid multicollinearity.

Other methods for optimizing the input variables, such as Convolutional Autoencoder, are possible, but in our case it was concluded that the PCA method suit our needs. Our goal with the data preprocessing was to avoid the multicollinearity among the input features, rather than compressing the information and PCA is a simple an realiable method for this purpose.

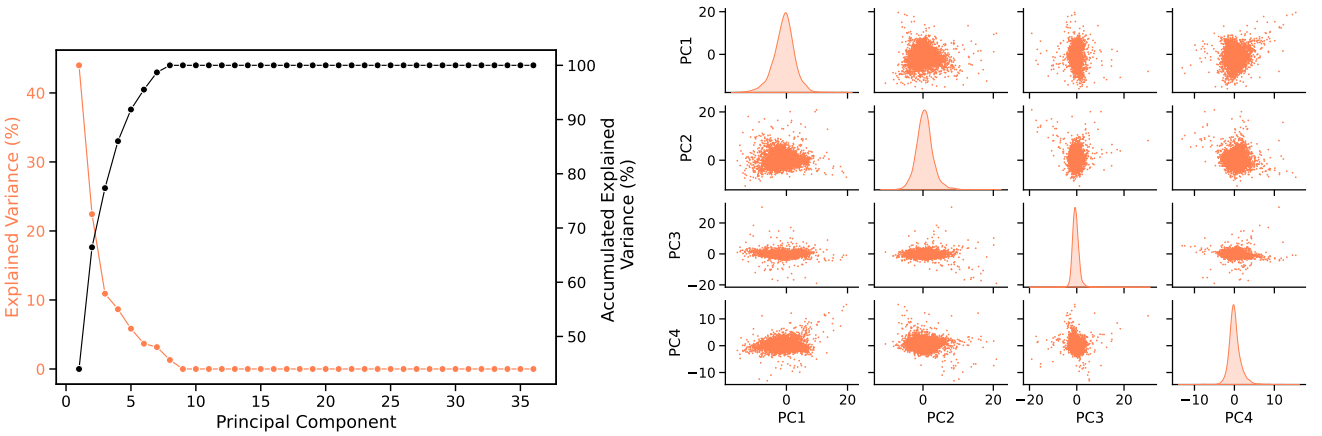

Figure 4: Decomposition of the input features using Principal Component Analysis (PCA) for the 2MPZ dataset. Left plot: Variability or variance explained ratio (left axis) and cumulative variability ratio (right axis) plotted against the number of PCs. Right plot: Correlation analysis of the first four PCs with a pair-wise relationships between them. Only 5000 points are included.

6.2 Model Architecture And Training for γ-MDN

In this section we provide an overview of the model architecture and present only the training analysis results for 2MPZ, as this catalog is one of the novel aspects of this work. For a more comprehensive analysis of the method or specific training results for DES, please refer to the study conducted by Cerdán [2020].

6.2.1 Summary of the Model Architecture

The γ-MDN model has been implemented using the Python programming language. For the development of this project, several Python packages and libraries have been used, such as *TensorFlow*, *Keras* and *NumPy*, among others. These packages provide specific tools and functions for neural network training and data processing.

The architecture of the γ-MDN model comprises five layers: the input layer, whose dimension is given by the number of features used for prediction, in our case it will be the number of PCs (10 for DES and 36 for 2MPZ).

Three hidden layers of 128, 64, and 32 units, respectively. The He-Uniform technique has been used for the network weights initialization, as there is evidence that it is an optimal initializer for ReLU non-linear activation functions. It also has a Batch Normalization (BN) layer before the activation function, which makes it unnecessary to use Dropout, accelerating training, as BN acts as a kind of regularization.

Finally, the dimension of the output layer will consist of the $3xK$ parameters parameters needed to define the mixture of gamma distributions. As the last activation function, it has a softplus function to avoid numerical instabilities. Additionally, the outputs of the neural network pass through an additional activation layer, where the mixing coefficients π_i are passed through a softmax function to ensure they are positive and sum to one (and thus normalize the output of a network to a distribution of probability). The shape parameters α_i and rate parameters β_i of the model's output elements are modified by adding constants $a = 2$ and $b = 1$, respectively (Equations 7 and 8). This adjustment is made to the existing softplus distribution provided by the model. The purpose of adding these constants is to ensure that the distributions approach zero as x approaches zero and to prevent any numerical instabilities.

The loss function employed is the average of the logarithm of the likelihood given by Equation 9 over all observations, $LF = \frac{1}{N}\sum_{n=1}^{N}\ell_n$, where N is the number of training samples. To minimize this error function in the γ-MDN, the Adam optimizer was used, which is a highly efficient first-order stochastic gradient descent optimization algorithm. The default parameters are used for the Adam optimizer, using the RI metric.

A summary of the layers already described that the model uses would be the following:

- Input layer: Number of features (36 PC for 2MPZ)
- Hidden layer (x3): 128, 64, 32
  - Weights Initialization: He-Uniforme
  - Batch Normalization
  - Rectified Linear Unit (ReLU)
- Output layer: dimension 3 x K (softplus + softmax)
  - Loss Function: Negative log-likelihood
  - Optimizer: Adam
  - Metric: RI

6.2.2 Training of γ-MDN for 2MPZ.

To obtain the optimal parameters for the 2MPZ training, we first studied how statistical metrics varied with the random seed and the batch size. In Figure 26 of the Appendix, some results of this analysis are presented. It was determined that the most optimal parameters for this model were a seed of 350 and a batch size of 456. These parameters were used in the remaining training processes of the project.

The performance of predictions using only magnitudes and both magnitudes and colors was firstly analyzed. Figure 5 shows the PIT histograms along with their corresponding RI for the training and testing samples of the 2MPZ catalog, using magnitudes (left figures) or magnitudes and colors (right figure). In both cases, the RI values are quite close to 0, which is the perfect value. However, the RI values are slightly smaller when using colors and magnitudes. Certain asymmetries are observed in the histograms, indicating the presence of bias. The statistics metrics were also calculated for both trainings, with only colors and with both magnitudes and colors, as shown in Figure 25 of the Appendix. It was observed that the variations between the two were not significant. Based on these findings, we decided to use colors and magnitudes as input variables since the distributions are slightly more homogeneous and with a smaller RI value.

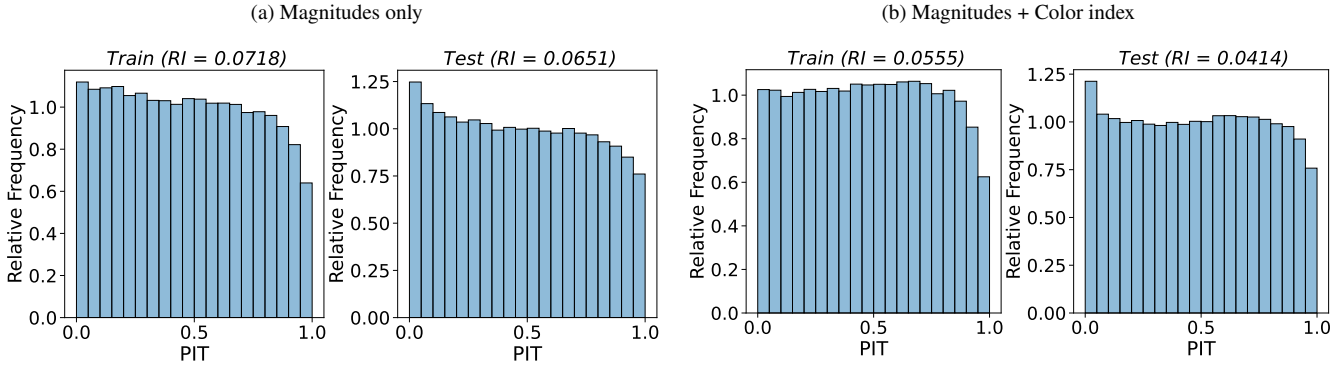

Figure 5: Probability Integral Transform (PIT) histograms and Reliability Index (RI) for the train and test partitions of the 2MPZ dataset. These results were obtained using γ-MDN. In plot (a), only the magnitudes were considered as input variables, while in plot (b), both the magnitudes and color indices were taken into account. The use of magnitudes and colors as input features is a more stable training process and yields a histogram with a slightly more uniform distribution, with a smaller RI value. Therefore, this combination of input features has been selected to train the 2MPZ data.

Figure 6 shows the progression of the training process. In each plot, the vertical line marks the point at which the corresponding metric reaches its minimum in the validation partition. This value has been selected as the optimal number of epochs, as beyond it there is a risk of overfitting. A study was also conducted to evaluate the effect of increasing the number of epochs beyond this reference value. The results are shown in Figure 27 in the Appendix. It was observed that the results

worsened as the number of epochs increased, suggesting that excessive epochs once the model has converged can lead to overfitting.

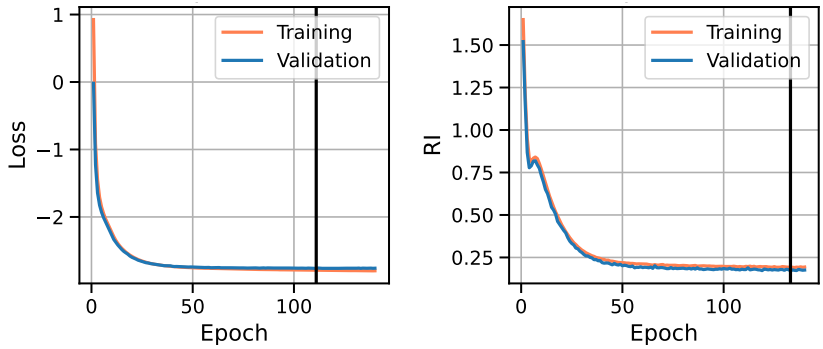

Figure 6: Training progress of γ-MDN for predicting true redshift, using magnitudes and colors as input features from the 2MPZ dataset. On the left, the evolution of the loss function with the number of epochs. On the right, the progression of the RI index as the number of epochs increases. The vertical line represents the point at which the minimum value for each metric was achieved in the validation partition.

6.3 Cluster-GMDN Methodology

Here we present the detailed application of the Cluster-GMDN method.

Firstly, the celestial sphere has been discretized into pixels of equal area using *HEALPix* (Hierarchical Equal Area isoLatitude Pixelization) algorithm with its NESTED scheme [Górski et al., 2005]. Then, the samples defined in Section 5 for the Cluster-GMDN application are associated to each pixel given its pixel index. The pixel index is an integer that indicates the location of the galaxy within the nested structure, and it was derived using the *healpixNestIndex* function in TopCat. The input variables used for this calculation are the astronomical coordinates right ascension, RA, and declination, Dec, along with the desired resolution order expressed as an exponent in base 2. In this study, we computed the *HEALPix* index for resolutions 64, 128, and 256.

Next, in each *HEALPix* pixel, we generate histograms of the galaxy density distribution as a function of photo-z using the γ-MDN estimates, for all specified resolutions. This enables the creation of probability density functions (PDFs) defined for the background along the line of sight in every pixel.

For each test galaxy (from the spectroscopic sample), we first identify the pixel index to which it belongs. Once selected, we search for neighboring pixels using a nested hierarchical structure. The indices of these neighboring pixels are then transformed into positions (RA, Dec), given the coordinate of the center of the pixel.

Next, we calculate the angular separation between the test galaxy (determined by the galaxy's RA and Dec coordinates) and the positions of the pixel and all neighboring pixels. We next select the four closest pixels and arrange them based on their distances, with the closest pixel being obviously the one where the galaxy is located.

Next, each pixel is assigned a weight based on its distance to the galaxy. This weight is calculated as the difference from 1 to the distance from the center of the pixel to the galaxy, normalized with respect to the distance to the fifth farthest pixel. This way, pixels closer to the galaxy are given greater weight and the fifth pixel is given a weight=0.

We select the calculated density histograms for the 4 nearest pixels to the galaxy. From these histograms, we determine the 'Spine' of the distributions, representing a simplification of the pixel's PDF, where we retain only the main structure identifying the local maxima of the distribution. Using the spine, we reconstruct the PDF of each pixel by connecting the maxima through catenary curves to highlight the overdensity peaks of the distribution.

Finally, the final background PDF, in the line of sight of the test galaxy, is calculated as the weighted sum of the distributions constructed for the 4 nearest pixels. This final function contains information about the clustering of galaxies along the line of sight at the right ascension and declination position of the target galaxy. It is this distribution that we use to multiply it with the PDF generated by γ-MDN, in order to obtain a photo-z probability density distribution modified by the background galaxy clustering.

6.4 Photo-z metrics for Result Evaluation

In order to evaluate the accuracy of our photo-z results, we have employed the following four statistical metrics for the test samples:

- The mean of the bias of the distribution, $\triangle z$, where bias is defined as the difference between the photometric redshift z_{phot} and the spectroscopic redshift z_{true}.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(z_{\text{phot}} - z_{\text{true}})_i = \frac{1}{N}\sum_{i=1}^{N}\triangle z_i \tag{12}$$

- The precision in 68-quantile, σ_{68}, which represent the 68% spread of data about the median value of the bias $\triangle z$ [Cerdán, 2020]. For a completely symmetric Gaussian distribution this value would be equal to the standard deviation, however we use Gamma distributions and also the PDFs we generate are not usually symmetric, so this value measures the width of the core of the $\triangle z$ distribution [Sanchez et al., 2014]. This metrics is defined as the difference between the 84th and the 16th percentiles of the cumulative distribution,P_{84} and P_{16}, of the bias.

$$\sigma_{68} = \frac{1}{2}\left(P_{84} - P_{16}\right) \tag{13}$$

- The outlier rate is defined as those values with a bias greater than 0.15 for DES and 0.09 for 2MPZ, considering the difference in the redshift range of each survey.

$$\text{OutRate}_{\text{DES}} = \frac{1}{N}\sum(|\triangle z| > 0.15) \quad ; \quad \text{OutRate}_{\text{2MPZ}} = \frac{1}{N}\sum(|\triangle z| > 0.09) \tag{14}$$

- The standard deviation, which is sensitive to the tails of the $\triangle z$ distribution [Sanchez et al., 2014], is defined as follows:

$$Stdv = \sqrt{\frac{1}{N}\sum \triangle z^2} \tag{15}$$

The results of the metrics applied to our predictions have been grouped into bins along the redshift range. For DES, these metrics are calculated for sets of observations divided into separate z_{spec} bins of width 0.1, starting from z=0.2 and extending up to z=1. For 2MPZ, the redshift interval goes from z=0 to z=0.25, with a bin width of 0.05. Since the γ-MDN model predicts probability density functions (PDFs), which represent continuous distributions, we decided to use the mode of the distribution as the γ-MDN photo-z best estimate.

Additionally, we have computed the metrics specified in Bilicki et al. [2013] for 2MPZ, to compare our results with theirs, despite the challenges encountered in the sample selection. These metrics include: the mean $\langle z\rangle$ and median $\bar{z}$; 1σ scatter between the spectroscopic and photometric redshifts; the scaled median absolute deviation; the net bias; the percentage of outliers; and the median of the relative error. We have calculated these metrics for both the mode and the mean of the γ-MDN predictions. The comparison is presented in Table 1. It is important to note that these metrics have been calculated for the entire dataset, following the approach used in the 2MPZ paper, not divided in redshift bins.

In the case of the Cluster-GMDN, when we explore the possibility of using the method as a way to cleaning the test sample, by removing outliers, we use the following the definition:

$$\text{ClusterOutRate}_{\text{DES}} = \left|z_\gamma - z_{Cluster}\right| > 0.15 \quad ; \quad \text{ClusterOutRate}_{\text{2MPZ}} = \left|z_\gamma - z_{Cluster}\right| > 0.03) \tag{16}$$

where z_γ is the redshift obtained as the mode of the γ-MDN distribution and $z_{Cluster}$ is the redshift obtained as the mode of the Cluster-GMDN distribution. The potential of using Cluster-GMDN result to clean a galaxy sample is shown in Section 7.1.5, demonstrating a promising application of the method, identifying outliers and generating a cleaner dataset.

7 Results and Discussion

This section presents the findings and detailed analysis obtained in this study. The results of applying the γ-MDN and Cluster-GMDN models to the DES and 2MPZ samples are discussed. The performance of the models is evaluated and compared in terms of accuracy in estimating photo-z's.

7.1 Application to DES

7.1.1 The Effect of Removing Galaxy Shape Variable

The original code developed by Cerdán [2020] included the parameter SOF_CM_T as one of the input features of the neural network for DES. This parameter provides information about the size of the galaxy and it is defined as the size squared of the object, T=+[arcsec^2]. However, in this project, we decided to modify the code and exclude this input feature to develop a more versatile model that can be applied to any survey.

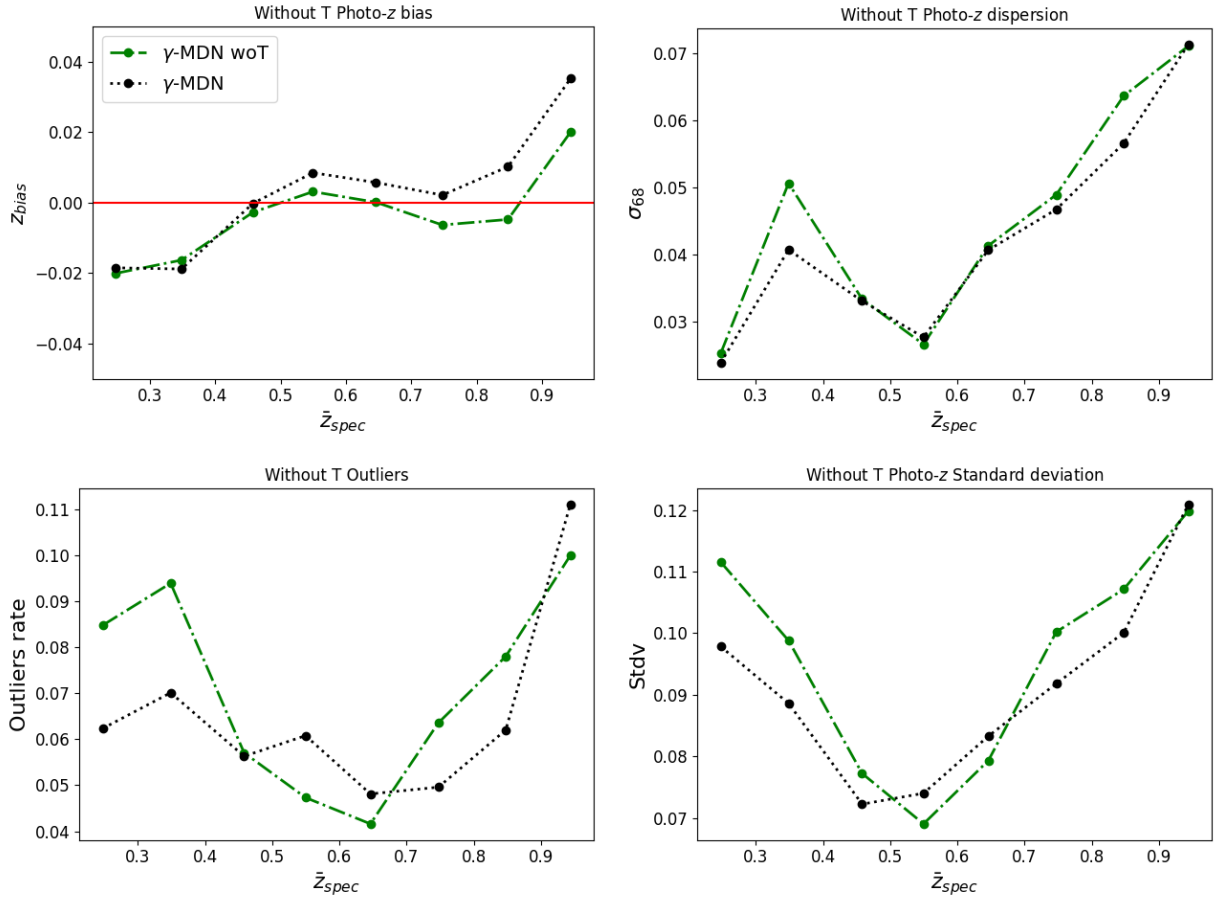

Figure 7: Photo-z bias (top left), 68-percentile width of the bias σ_{68} (top right), outliers rate (bottom left) and standard deviation (bottom right) as a function of true redshift for the original γ-MDN, in black, and γ-MDN without the galaxy shape parameter T, in green.

Figure 7 shows the effect of removing SOF_CM_T. The metrics are shown for two scenarios: one where the parameter is included as an input variable (green line) and another where the parameter is excluded (black line). It was expected that removing this parameter would lead to a more significant variation in the results, especially at lower and higher redshifts, due to the change in the apparent size of objects when projected onto the celestial sphere. It is observed that including the parameter improves the results for almost the entire redshift range in terms of bias. However, upon examining the results as a whole, no evident improvement is generally observed when not using the SOF_CM_T parameter. Therefore, the decision

was not to include this input through the analysis, and allow the code to be based solely on magnitudes alone. This makes the model more general, allowing it to be used to more diverse data, without relying on specific survey reduction.

From this point onwards, all the results were obtained by training the neural network without including the SOF_CM_T parameter, although it is not explicitly stated.

7.1.2 Assessing Model Performance: Comparative Analysis of γ-MDN and the DES Reference Method

In this section, we analyze the predictive performance of γ-MDN and compare it with the photo-z prediction used by DES, *DNF*. First, we provide a general comparison of the results. Later, a more detailed analysis will be conducted using the statistical metrics defined in Section 6.4.

Figure 8, in the rigth side, illustrates the mode of the predicted PDFs by γ-MDN plotted against the spectroscopic values. On the left, the same plot is shown for *DNF*. It is evident that both methods perform quite well, as the majority of the data density is concentrated around the bisector line, highlighted in red. In an ideal scenario, where the model performs perfectly, all the results would lie on the bisector line. Although both models appear to perform similarly, it can be observed that γ-MDN yields fewer scattered galaxies in certain redshift regions. Hence, it is already apparent that γ-MDN matches or even outperforms *DNF* in some areas. On the other hand, *DNF* seems to give a lower number of outliers in the middle region.

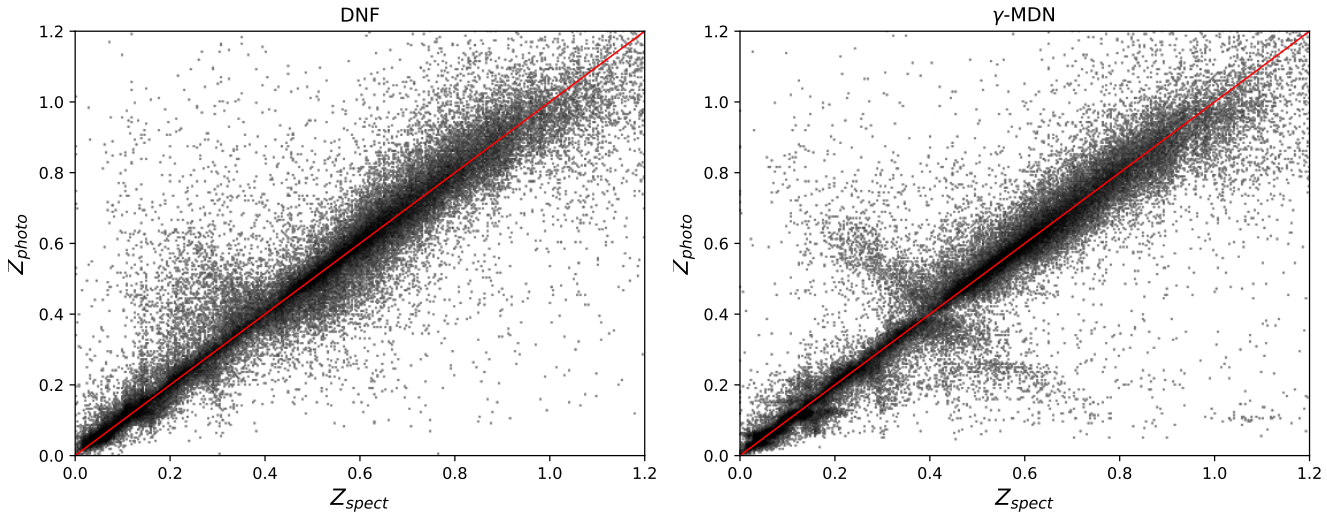

Figure 8: Scatter plot of the spectroscopic values, z_{spect}, versus the photometric predictions, z_{photo}, obtained by the *DNF* (left) and γ-MDN models (right). For clarity, the bisector is shown overlaid in red.

Although in Figure 8 we represent the mode as the photo-z prediction, the model actually outputs PDFs as results. Figure 9 illustrates the reconstruction of the PDFs based on the resulting parameters from the model, namely π_i, α_i, and β_i. The PDF generated by γ-MDN is depicted in orange, with the mode represented in blue as the reference value to compare, while the spectroscopic position of the galaxy is shown in black. The first row show four randomly selected examples of galaxies with unimodal PDFs. The next row features another random selection of four galaxies with multimodal PDFs. Finally, the third row presents examples where the true position of the galaxy is near another local maximum that is not the mode. It is a very promising model because, as observed, it performs quite competently for unimodal galaxies. However, its greatest potential lies in galaxies with multimodal probability density functions (PDFs). The fact that the model provides PDFs as output variables instead of a single value gives us a wealth of statistical information about the possible position of the galaxy. For instance, knowing that we have a multimodal distribution allows us to consider all local maxima instead of just the mode, and leverage additional statistical information, such as the background galaxy clustering, to identify the correct position. The potential of this approach is clearly evident in galaxies that have a second peak near the true value. The benefits of applying this information are analyzed in Section 7.1.4.

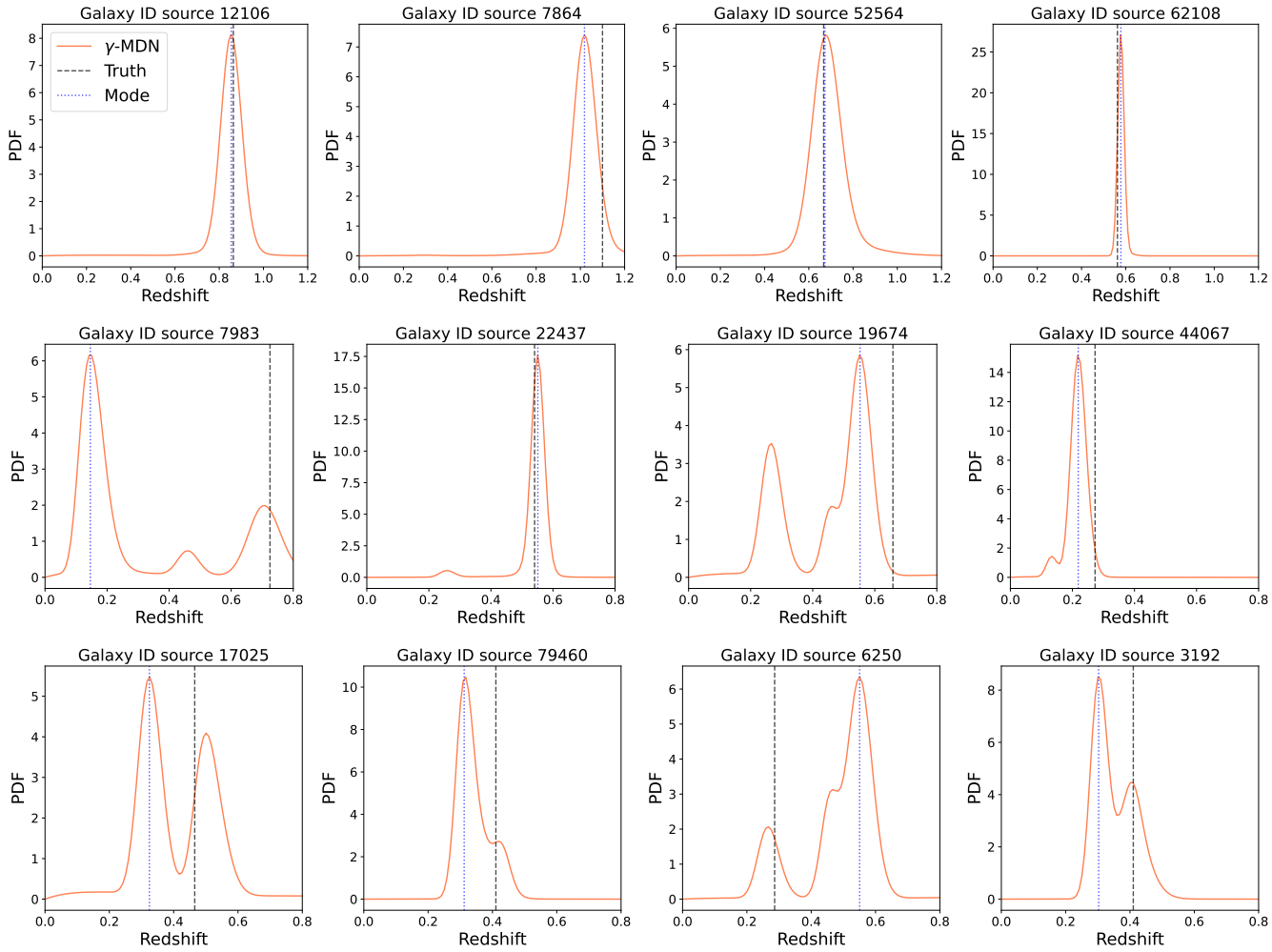

Figure 9: Examples of individual photo-z PDFs predicted by γ-MDN for the following cases: in the first row, random examples of unimodal distributions; in the second row, random examples of multimodal distributions; and in the third row, a selection of galaxies where the true value is near a maximum that is not the mode. For a clearer analysis of the results, the mode of the PDF is given in blue, and the spectroscopic redshift is shown in black.

In Figure 10 we compare the performance metrics of the results obtained by *DNF* (purple) and γ-MDN (green). Overall, we observe that our bias values are better than those of *DNF*, especially at low redshifts, although they worsen slightly at high redshifts. For the dispersion, σ_{68}, the results are very promising as we either improve upon *DNF* or, in the worst case, achieve similar performance. However, for the standard deviation and outliers, the results are worse, with an improvement only observed for intermediate redshifts between 0.6 and 0.85 in the case of outliers.

7.1.3 Application of γ-MDN to Maglim

The γ-MDN model was successfully applied to the entire MagLim subset of DES. As a result, we now have a vast dataset with our own photo-z predictions and their corresponding PDFs. We consider these PDFs to be reliable and highly valuable, given the model's good performance. Moreover, these results have been used to generate the background PDFs used in the Cluster-GMDN method, leveraging the clustering information along the line of sight of each galaxy in the test spectroscopic sample.

7.1.4 Analysis of Cluster-GMDN Applied to DES

In this section, we analyze the results of the Cluster-GMDN method compared to γ-MDN and *DNF*. Figure 10 compares the statistical metrics obtained with Cluster-GMDN against the other two models for a *HEALPix* pixel resolution of 128. Overall,

the bias we obtain is quite similar to what was already achieved with γ-MDN, with a slight degradation at high redshifts. The dispersion also behaves in a very similar manner, showing a slight improvement at high redshifts. The Cluster-GMDN model achieves significantly better results for outliers at high redshifts, outperforming both γ-MDN and *DNF*. Additionally, a considerable improvement is observed in the standard deviation. Overall, Cluster-GMDN seems to improve the γ-MDN results, indicating that by incorporating clustering information from the background, we can improve the photo-z estimates, as we expected.

These metrics were also calculated for pixel resolutions of 64 and 256 (results shown in the Appendix, Figure 23). We chose to stick with a resolution of 128 as it provides an intermediate option, and there wasn't a significant difference in the results among the three resolutions. However, using a resolution of 64 does show some improvement in outliers for both high and low redshifts.

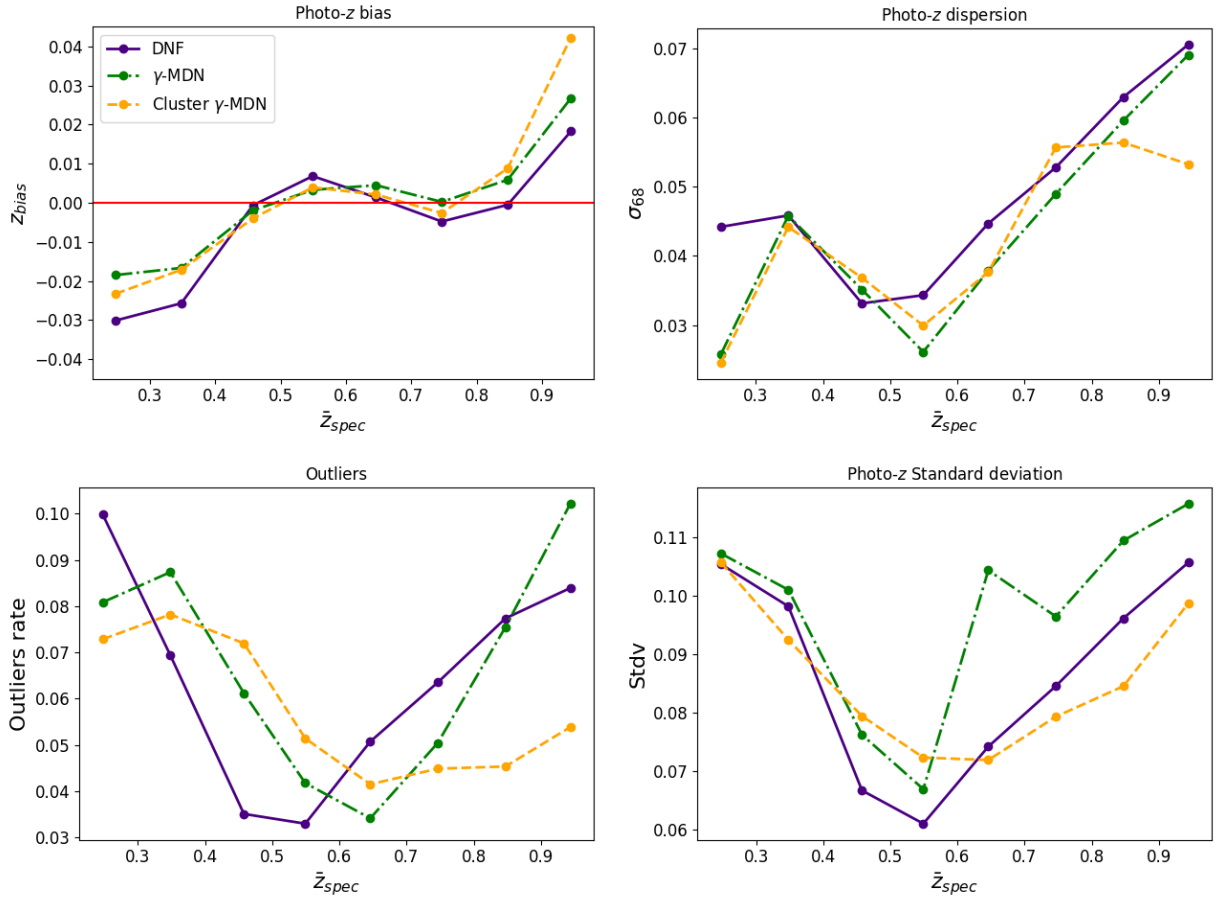

Figure 10: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for *DNF* (purple), γ-MDN (green) and Cluster-GMDN (orange). Cluster-GMDN outperforms γ-MDN for certain redshift ranges.

In Figure 11 we show the deviation of the values obtained by Cluster-GMDN from the true redshift in the x-axis against the deviation of γ-MDN from the true redshift in the y-axis. This plot aims to compare the performance of the two models more generally. A region of higher density is observed along the bisector of the graph, indicating that for galaxies in this region, both models yield very similar results. Galaxies for which the Cluster-GMDN model has performed better than γ-MDN are located in the upper triangle above the bisector. The most significant improvements are seen for galaxies with $z_{\text{spec}} - z_{\text{Cluster GMDN}}$ values below 0.2 and $z_{\text{spec}} - z_{\gamma\text{MDN}}$ values above 0.5. Conversely, galaxies situated in the lower triangle below the bisector are the ones for which applying the group PDF information results in a worsening of the estimates. Analyzing the overall distribution of galaxies in the plot, it appears that there is a higher density of galaxies in the region where applying Cluster-GMDN worsens the results. However, as a preliminary test of this model, the results are very promising, as there are many galaxies for which the redshift estimates are improved. Therefore, this provides a starting point for further analysis to determine, for example, if

there is a trend in the types of galaxies for which Cluster-GMDN yields good or poor results.

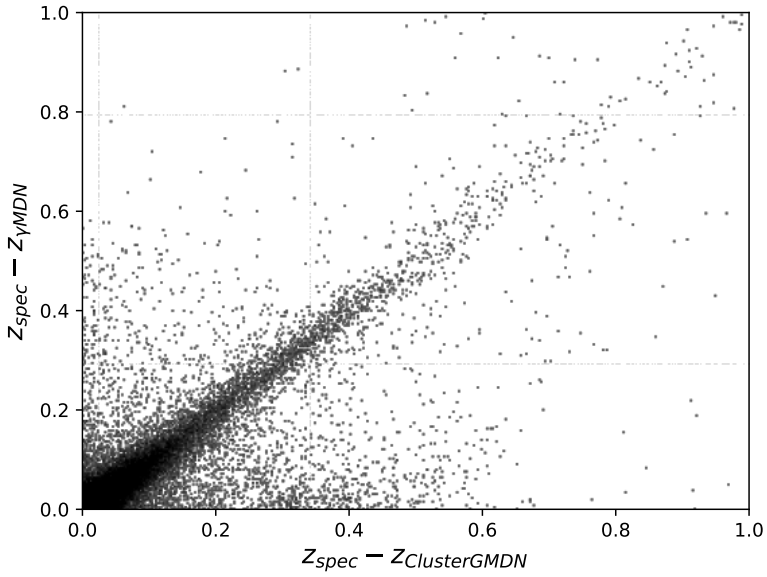

Figure 11: Scatter plot comparing the deviation between the true spectroscopic redshift and the predictions from Cluster-GMDN (x-axis) versus the deviation between the true spectroscopic value and the γ-MDN prediction (y-axis). Due to the construction limitation of the MagLim sample, the results are shown only for galaxies with $z > 0.2$. In the upper-left triangle of the plot, above the diagonal, we have the galaxies for which the Cluster-GMDN provides a more accurate photo-z estimation and improves the results of γ-MDN, with the most significant improvements occurring for small values on the x-axis and large values on the y-axis. However, in the lower-right triangle of the plot, we find the galaxies for which we are worsening the photo-z prediction by incorporating the clustering information from the background.

In Figure 12, a selection of specific galaxies, picked from the previous plot, is presented, to show the performance of the Cluster-GMDN method on individual galaxies. The green line represents the PDF from γ-MDN, the gray line shows the PDF resulting from the background cosmic web, obtained from our predictions for the MagLim sample along the line of sight of that galaxy, and the yellow line represents the PDF resulting from the multiplication of the previous two probability functions. The respective modes of γ-MDN and Cluster-GMDN are highlighted as vertical green and yellow lines, respectively. The spectroscopic position of the galaxy is overlaid in blue. The first row shows two examples of galaxies for which employing the background information significantly improves the results by giving more weight to the peak where the galaxy is actually located and selecting the correct peak of probability. The second row presents the opposite scenario, a case in which a good value was obtained with γ-MDN but worsens when applying Cluster-GMDN.

The photometric characteristics of galaxies at different redshifts can be very similar depending on their spectral type. But in general, we expect more probability of a galaxy to seat over a density peak than in a Void. Therefore, this type of analysis and results are very beneficial for studies like DES, or the future LSST, which cover a wide range of redshifts and spectral types, thus, having a higher probability of encountering galaxies with multimodal PDFs. Specifically, in our analysis, the percentage of multimodal PDFs obtained in DES is 31.59%. The ability of γ-MDN to provide multimodal PDFs as output variables is particularly useful in these cases, as it allows us to consider multiple possibilities and leverage the result using the background clustering, to discern the most likely position of the galaxy in redshift space. This helps us overcome the ambiguity caused by photometric characteristics typically found in photo-z codes and improve our redshift estimations.

To further explore the behaviour of the Cluster-GMDN method, we separated the results into two samples: one with multimodal PDFs and another for unimodal PDFs only, and their metrics were calculated separately. Overall, the results for galaxies with unimodal PDFs behaved similarly to the entire sample. For the multimodal PDFs, although there is more variation, the trend remained the same. In other words, Cluster-GMDN provided significant improvements, especially at high redshifts, in almost all metrics (figures shown in Appendix, Figures 21 and 22).

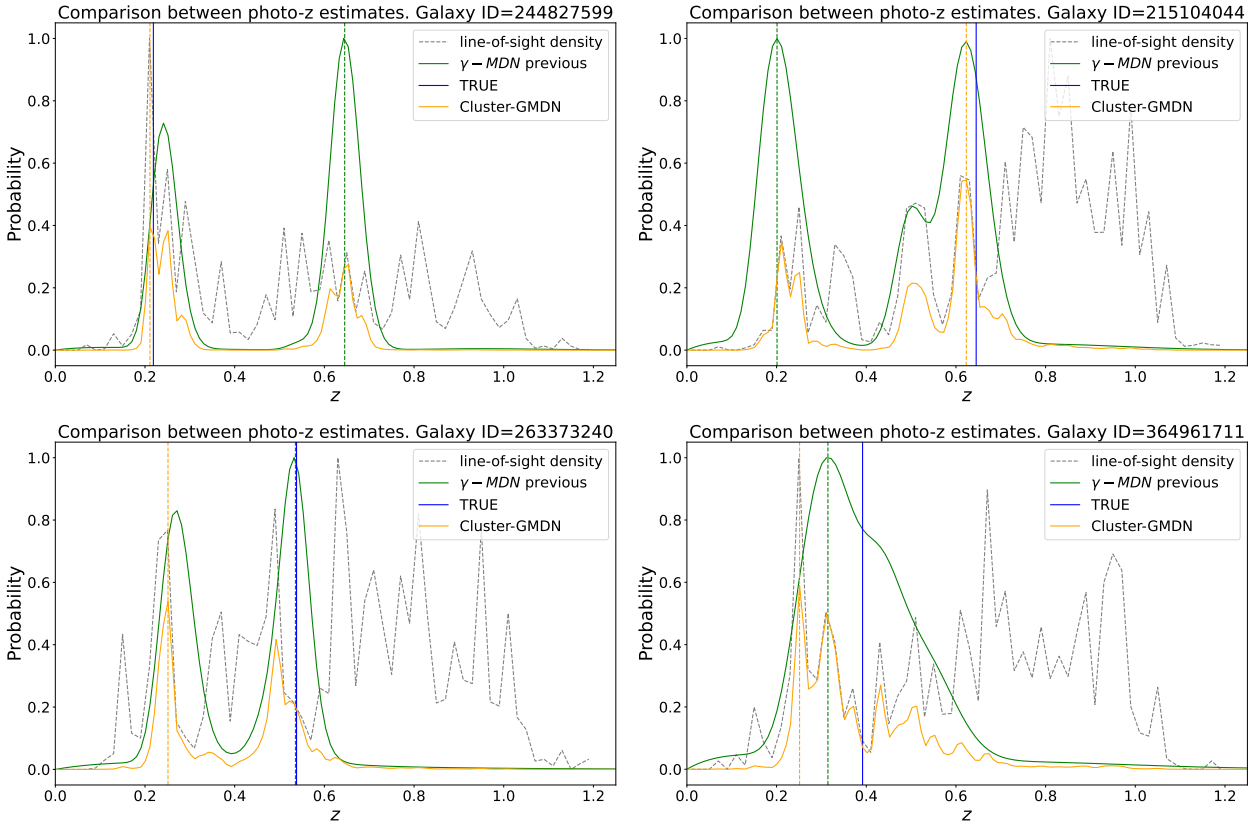

Figure 12: Comparison of photo-z PDFs estimated by γ-MDN and Cluster-GMDN for individual galaxies. The estimated PDF by γ-MDN are shown in green, the background PDF along the line of sight is shown in gray, and the product of the γ-MDN and background PDFs, known as Cluster-GMDN is shown in yellow. Vertical lines depict the values of the true spectroscopic redshift (blue), as well as the modes provided by the γ-MDN (green) and Cluster-GMDN (yellow). These galaxies have been specifically chosen to illustrate cases where Cluster-GMDN improves the predictions (first row) and cases where it worsens (second row).

7.1.5 Cluster-GMDN as a tool for improving datasets

Another potential application of the Cluster-GMDN model is to use it as a tool for creating a cleaner subset by removing outliers based on the information it provides. To accomplish this, Equation 16 was used, identifying galaxies with a difference greater than 0.15 between γ-MDN and Cluster-GMDN. Figure 13 shows the comparison between the original γ-MDN sample and the one where outliers are filtered using the Cluster-GMDN difference cut. The results are very promising as the metrics for the filtered subset are significantly improved with respect to the unfiltered sample, while only removing a few thousand galaxies. It is worth noting that the bias worsens slightly at higher redshifts, which can be attributed to γ-MDN artificially producing a lower bias due to the presence of outliers at these redshifts, while the filtered value might be more realistic.

As an additional test, the threshold used to filter the outliers was changed, with values <0.12 and <0.17. It was found that there was minimal variation in the results. This is shown in Figure 24 in the Appendix

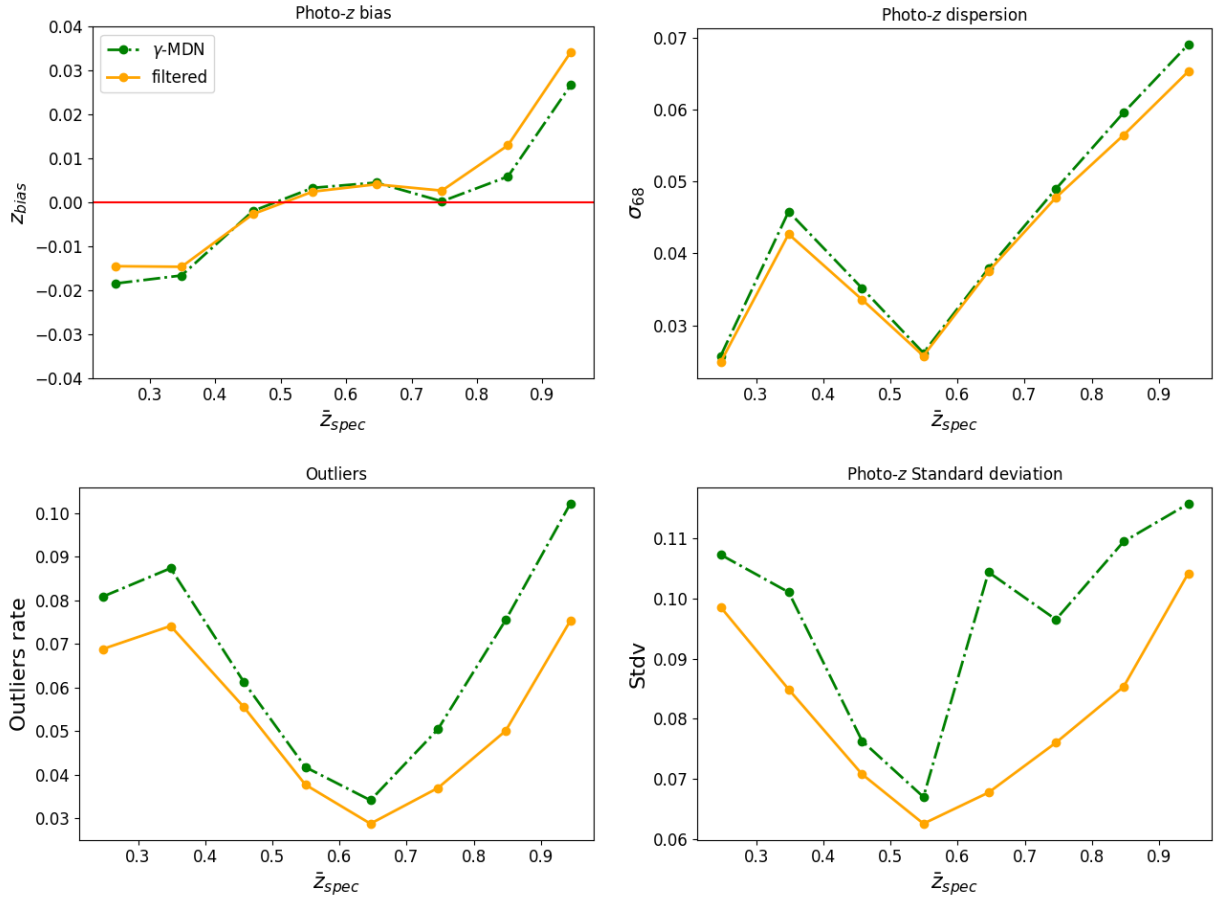

Figure 13: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for the predictions of the γ-MDN model (green) and the outlier-cleaned γ-MDN model (orange) using the Cluster-GMDN information.

7.2 Application to 2MPZ

7.2.1 Assessing Model Performance: Comparative Analysis of γ-MDN and the 2MPZ Reference Method

Similar to the analysis conducted for DES, this section examines the results of applying γ-MDN to the 2MPZ dataset and compares them with the photo-z results provided by the *ANNz* method. We start with a general analysis of the results and then proceed to examine examples of individual galaxies. A more in-depth analysis will be carried out using the statistical metrics.

Figure 14, on the right side, depicts the mode of the predicted PDFs by γ-MDN plotted against the spectroscopic values, using colors and magnitudes as input features. On the left side, the results obtained by the *ANNz* method. Comparing the predictions of both models, we can see that our results appear slightly more scattered. However, *ANNz* only provides a single value as the output variable, while we provide PDFs spanning the entire range of redshift for the sample, which can be considered reliable. Therefore, although our results may be slightly worse, they contain more statistical information that can be highly valuable for the scientific community.

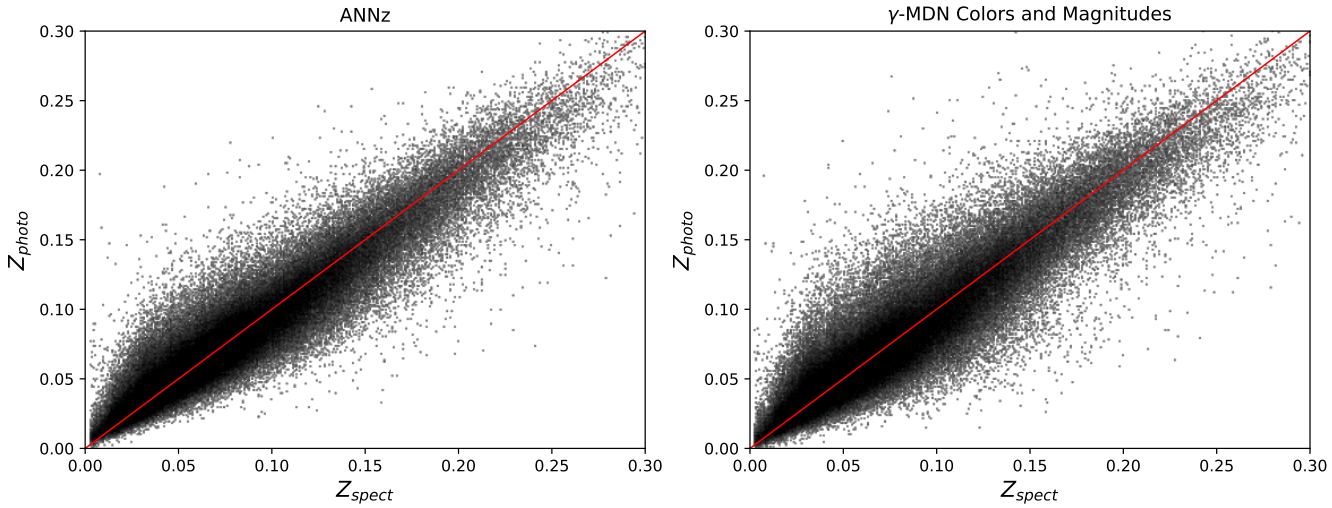

Figure 14: Scatter plot of the spectroscopic values, z_{spect}, versus the photometric predictions, z_{photo}, obtained by *ANNz*, used by 2MPZ, (left) and γ-MDN (right). For clarity, the bisector is shown in red.

Resulting PDFs for individual galaxies were selected and showcased in Figure 15. The first row presents four randomly selected galaxies with unimodal PDF. The next row features another random selection of four examples with multimodal PDFs. The third row displays four examples where the true position of the galaxy is near another local maximum that is not the mode. Finally, we have added an additional row showing galaxies with PDFs exhibiting high variance and asymmetry. Since the 2MPZ sample has a much narrower redshift range compared to DES (with a maximum redshift of 0.3 compared to DES's 1.4), the percentage of multimodal PDFs is much smaller. We obtained a multimodal percentage of 2.07%. Therefore, for this dataset, it may not be that important to generate multimodal PDFs. Nonetheless, the use of mixture Gamma functions allows us to generate asymmetric PDFs that are neither Gaussian nor perfectly Lorenztian. This provides valuable statistical information as well.

In Figure 18 we show the comparison between γ-MDN (in green) and *ANNz* (in purple). It is evident that both models yield fairly similar results, and if our results appear slightly worse, it may be due to the fact that our model was not trained on the same training sample as *ANNz*. We were unable to reproduce the same exact dataset they propose since, as we explained before, the version of the data provided is not the same as the one they used to train their model, and it is limited to magnitudes up to 14. Additionally, they train using the W1-W3 color index. These factors can affect the comparison, potentially resulting in differences coming from issues different than the photo-z code itself. Furthermore, it is possible that we are using galaxies in the test sample that they used for training, where the *ANNz* value will be exactly its spectroscopic redshift. This can lead to a better metric compared to our estimate. Without knowledge of the specific train sample they used, it is difficult to perform an exact comparison.

In Figure 16, the metrics for galaxies identified by γ-MDN as unimodal, with a variance lower than 0.00015, are also shown. The low-variance multimodal predictions from *ANNz* are shown in purple, while the mode and mean of our predictions are shown in green and blue, respectively. In Figure 17, we did a similar study but for galaxies with a high variance. The separation between the mode and the mean, in the results, was done because, despite being mostly unimodal, these galaxies exhibit a significant degree of asymmetry, and the mean value may provide a closer estimate to the true value. In fact, a general observation is that the mean tends to provide better statistical metrics. However, no clear conclusions can be drawn from both figures. The differences between σ_{68} and the standard deviation appear to be larger for galaxies with high variance, indicating a potentially higher level of asymmetry in the PDFs. However, these observations are inconclusive. Further studies are necessary to explore the results provided by γ-MDN for PDFs with a high degree of asymmetry.

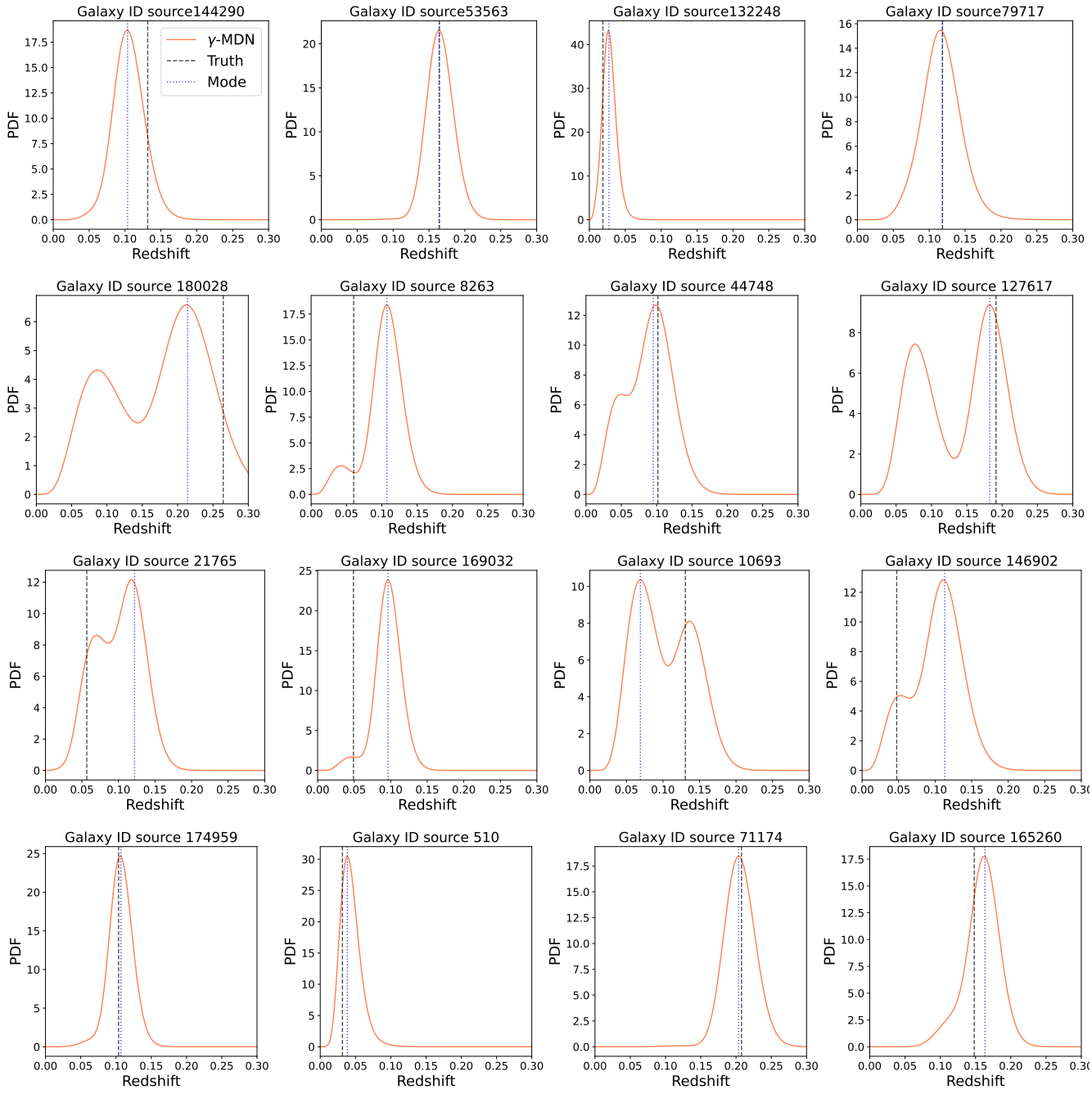

Figure 15: Examples of individual photo-z PDF predicted by the γ-MDN code for the following selection of galaxies: in the first row, random examples of unimodal distribution; in the second row, random examples of multimodal distribution; in the third row, a selection of galaxies where the true value is near a maximum that is not the mode; and in the fourth row, a selection of unimodal galaxies with high variance presenting asymmetries. For a clearer analysis of the results, the mode of the predicted PDF is overlaid in blue, and the spectroscopic redshift is shown in black.

Finally, Table 1 shows the results of applying the statistical metrics used by 2MPZ (Table 1 in Bilicki et al. [2013]) to our results, both for the mode and the mean. In general, the agreement is good, showing that even if some differences were found in the previous metrics, the global statistics are very similar, even within two orders of magnitude for the mean value of the spectroscopic redshifts of the 2MPZ test sample.

Method	$\text{Mean}_{\text{phot}}$ $\langle z \rangle$	$\text{Median}_{\text{phot}}$ $\bar{z}$	1σ scatter $\sigma_{\delta z}$	scaled MAD	net bias $\langle \delta z \rangle$	median error	% of outliers
2MPZ	0.078934	0.0707	0.016221	0.012750	-4.6e-5	12.5%	2.99%
mode γ-MDN	0.078663	0.0694	0.018358	0.014381	-0.000317	14.2%	3.04%
mean γ-MDN	0.079725	0.0705	0.018055	0.014529	0.000745	14.2%	2.73%
z spectroscopic	0.078979	0.079725	-	-	-	-	-

Table 1: Statistics employed by 2MPZ for their photometric redshift estimates, including the predictions from *ANNz* used by 2MPZ, as well as the mean and mode predictions from γ-MDN.

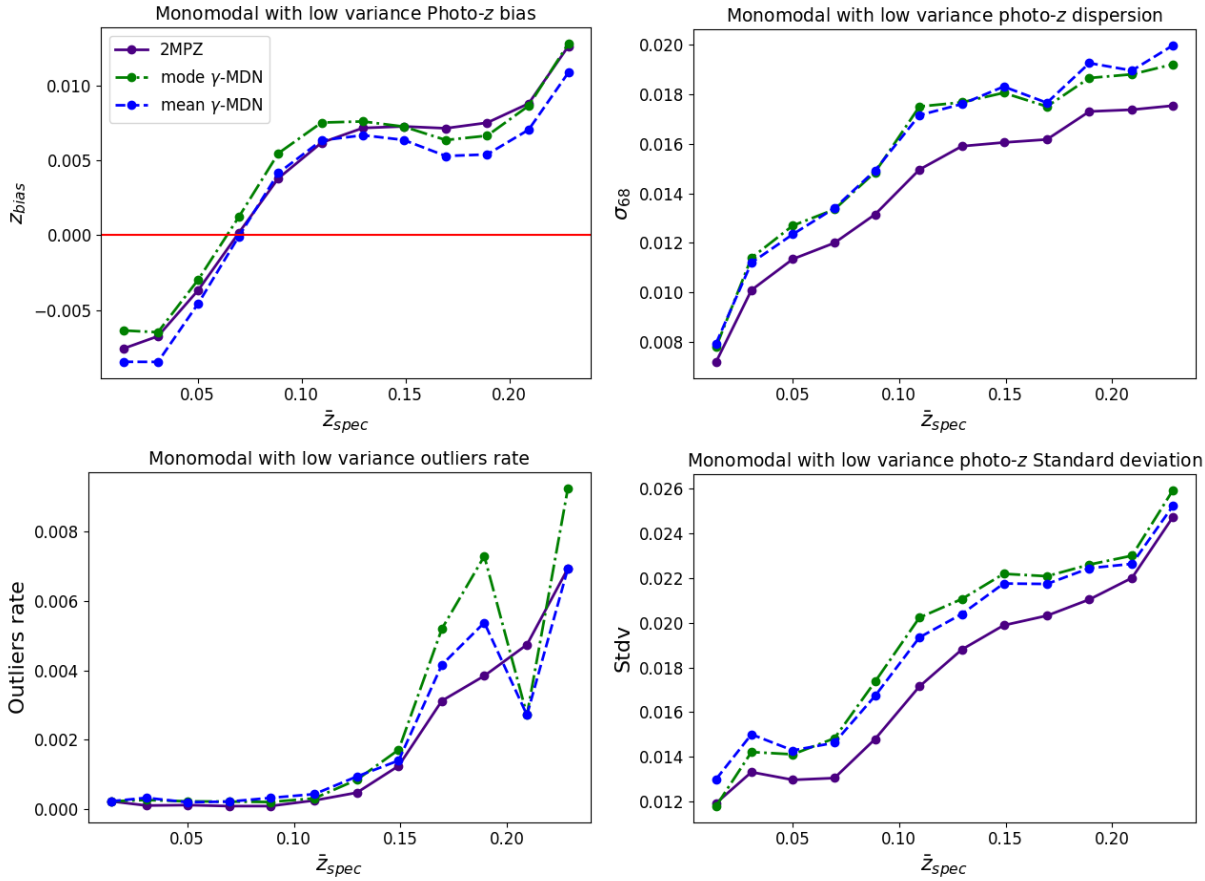

Figure 16: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for the predictions of *ANNz* (purple) and γ-MDN. The γ-MDN predictions are shown only for unimodal PDFs with a variance lower than 0.00015. The γ-MDN predictions are given in two ways, using the PDF mode (in green) and using the mean of the distribution (in blue). Differences between both indicates the PDFs are not symmetric, justifying the use of Gamma functions instead of, for example Gaussian distributions.

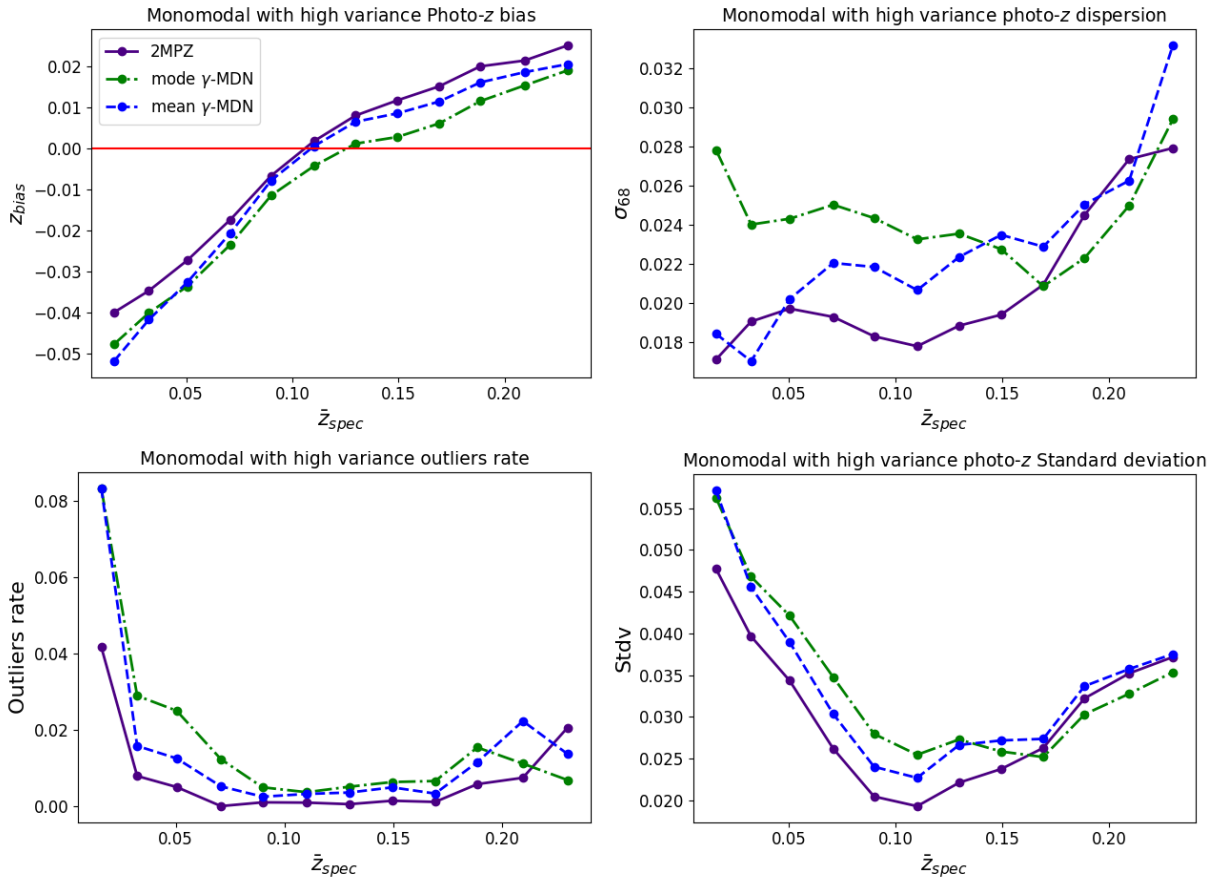

Figure 17: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for the predictions of *ANNz* (purple) and γ-MDN. The γ-MDN predictions are shown only for unimodal PDFs with a variance higher than 0.00015. The γ-MDN predictions are given in two ways, using the PDF mode (in green) and using the mean of the distribution (in blue). Differences between both indicates the PDFs are not symmetric, justifying the use of Gamma functions instead of, for example Gaussian distributions.

7.2.2 Analysis of Cluster-GMDN Applied to 2MPZ

The Cluster-GMDN model was applied to the 2MPZ data for two different resolutions, 64 and 128. The results of this are shown in Figure 18. These results are plotted together with γ-MDN and the proprietary 2MPZ model. In general, incorporating background information worsens the results. This may be due to the fact that these data is constrained to a relatively low redshift range, therefore the effect of clustering along the line of sight is likely not relevant, resulting in poorer metrics. Initially, the method was tested with a pixel resolution of 128, but we decided to decrease it to 64 due to the selection of few galaxies per pixel at this resolution.

In Figure 19, the performance of Cluster-GMDN is compared to γ-MDN. It can be observed that the majority of differences with respect to the spectroscopic value are concentrated between 0 and 0.05. There is a higher dispersion of results below the bisector line, indicating that we are degrading the results more than improving them when using Cluster-GMDN.

Finally, we have selected a pair of galaxies for which we observe an improvement in photo-z estimation with Cluster-GMDN, represented in the first row of Figure 20, and another pair for which the performance worsens, shown in the second row. These plots clearly demonstrate the potential of the Cluster-GMDN model. Further analysis and refinement of the model are necessary to enhance its performance and generalize its effectiveness across a wider range of galaxies.

Additionally, we attempted to use the results from Cluster-GMDN for filtering purposes, but there was hardly any difference observed in the metric results, as was the case with DES. The findings of this analysis are presented in Figure 28 in the Appendix. It is possible that because Cluster-GMDN is not an effective method for catalogs at low redshifts, it may not be suitable for cleaning outlier samples either.

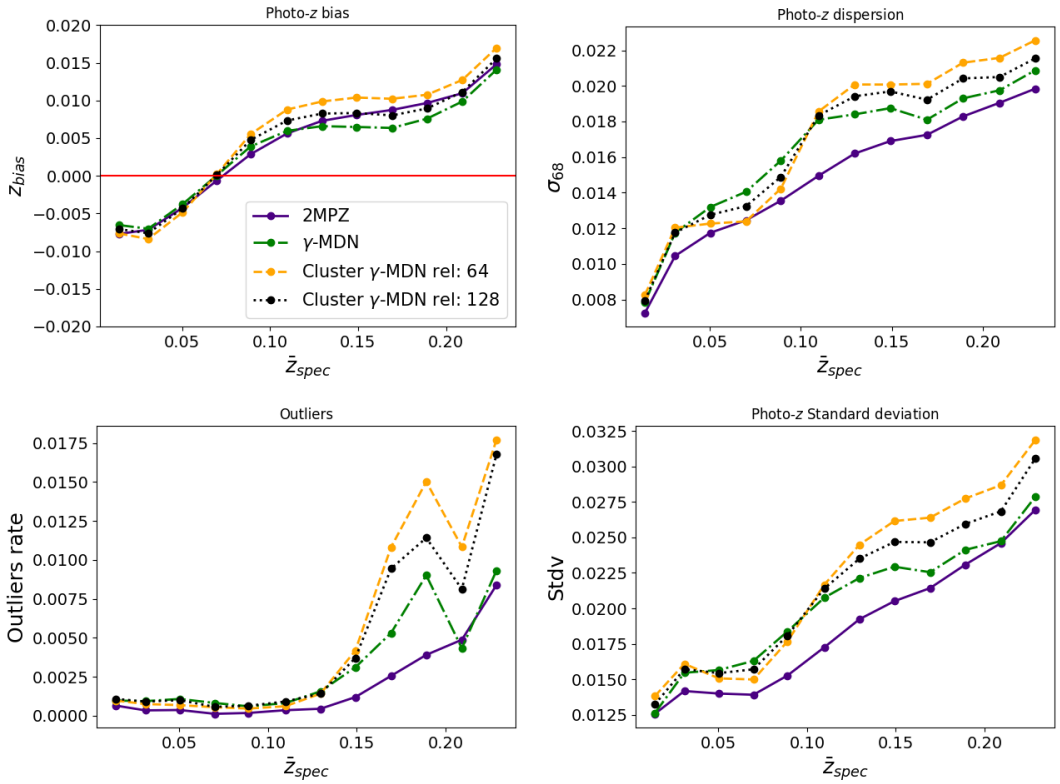

Figure 18: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for *ANNz* (purple), γ-MDN (green), the Cluster-GMDN estimate with a resolution of 64 (orange), and Cluster-GMDN with a resolution of 128 (black). In general, applying the Clustering technique does not improve the results, likely due to the small redshift range of the 2MPZ sample.

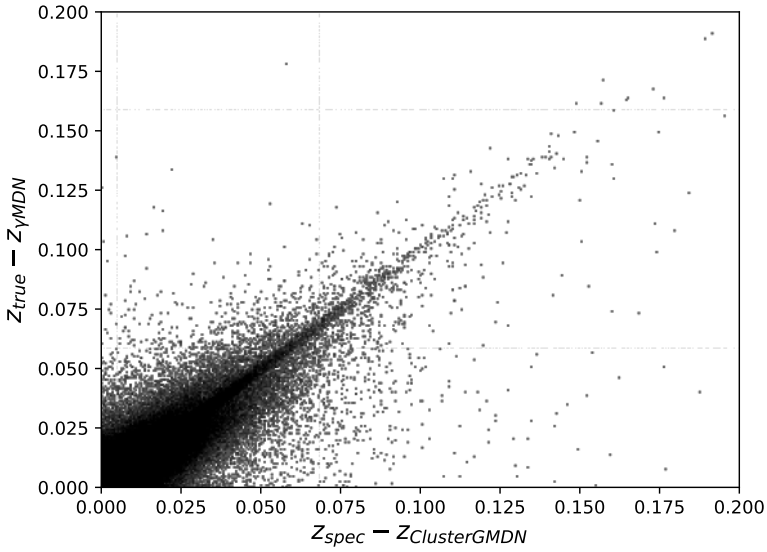

Figure 19: Scatter plot comparing the deviation between the true spectroscopic redshift and the predictions of the Cluster γ-MDN model (x-axis) with the deviation between the true spectroscopic value and the predictions of the γ-MDN model (y-axis). The test sample of 2MPZ has been used for this plot. In the upper-left triangle of the plot, above the diagonal, we have the galaxies for which the Cluster-GMDN provides a more accurate photo-z estimation and improves the results of γ-MDN, with the most significant improvements occurring for small values on the x-axis and large values on the y-axis. However, in the lower-right triangle of the plot, we find the galaxies for which we are worsening the photo-z prediction by incorporating the clustering information from the background.

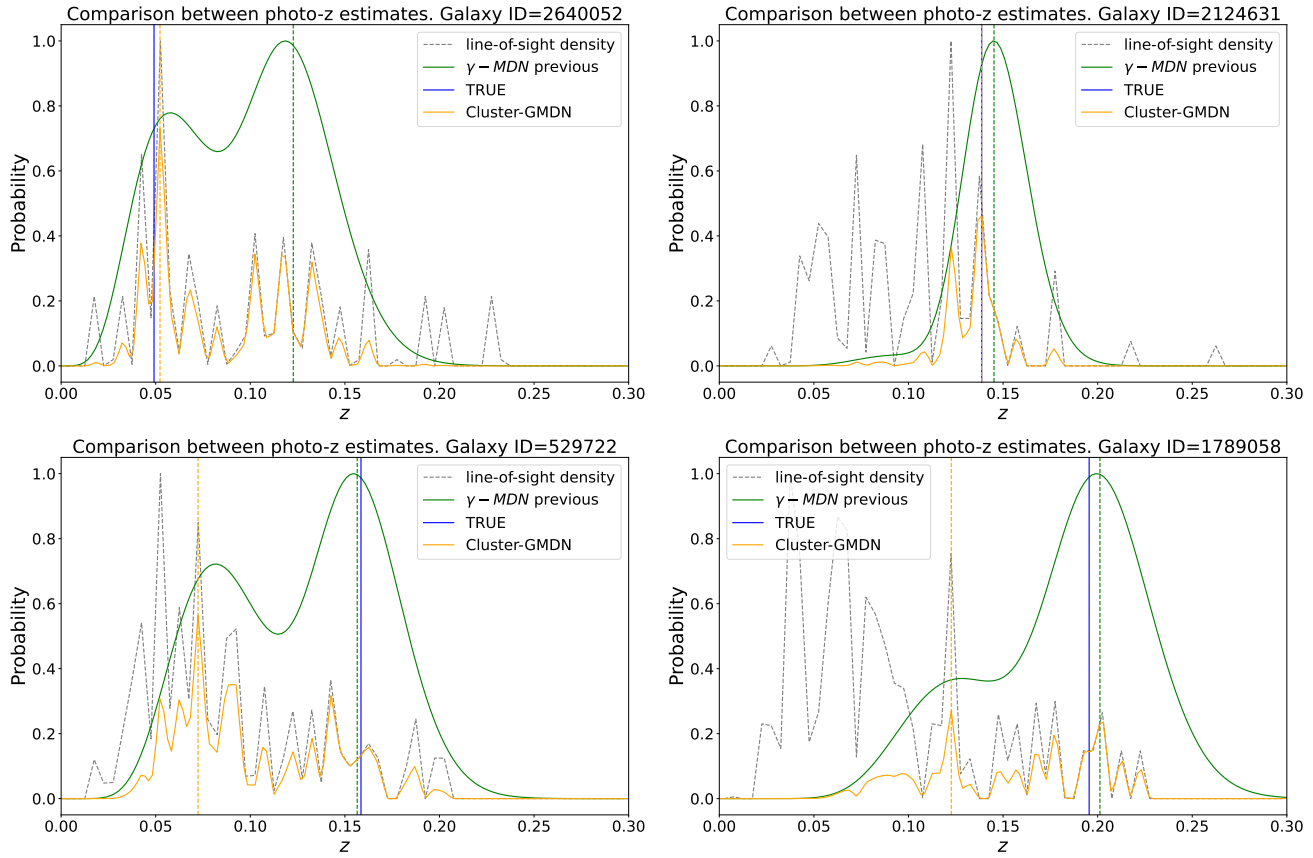

Figure 20: Comparison of photo-z estimations and predicted PDFs by the γ-MDN and Cluster γ-MDN models for individual example galaxies from the 2MPZ dataset. The estimated PDF by the γ-MDN model is shown in green, the background PDF along the line of sight to the selected galaxy is shown in gray, and the product of the γ-MDN and background PDFs, known as Cluster-GMDN, is shown in yellow. Vertical lines depict the values of the true spectroscopic redshift (blue), as well as the modes provided by the γ-MDN (green) and Cluster-GMDN (yellow) PDFs. These galaxies have been specifically chosen to illustrate cases where Cluster-GMDN improves the predictions (first row) and cases where it worsens them (second row).

8 Conclusions and future prospects

The main goals of this work were to test and evaluate the performance of γ-MDN on the DES and 2MPZ samples, as well as to test the new model called Cluster-GMDN. These objectives were successfully achieved, showing that γ-MDN is a highly competent model beyond its use to DES, comparable in precision with other photo-z codes. Its advantage over other models lies in its hability to provide reliable probability density functions (PDFs), which adds significant information to the results. In this work, we demonstrated how to use this extra information. This is where the Cluster-GMDN model comes into play. We find that combining the background clustering with the individual galaxy PDFs, we can refine the photo-z estimation. It is important to note that our algorithm does not rely on any fiducial cosmology, making it a completely free model and avoiding possible biases arising from assuming an incorrect cosmological framework.

As secondary objectives, we successfully applied the γ-MDN model to the DES MagLim galaxy sample, resulting in an new photo-z estimation including PDFs. Additionally, we discovered that Cluster-GMDN is a useful tool for creating outlier-free sub-samples, without removing an excessive number of galaxies.

Throughout this study, we obtained several intriguing findings. Particularly, we discovered that for samples with a broader redshift range, γ-MDN yields a higher percentage of multimodal PDFs, which is logical as the uncertainty increases with greater distances along the line of sight to the galaxy. We achieved a multimodal PDF percentage of 31.59% for DES and 2.07% for 2MPZ, with DES extending up to z=1.4 and 2MPZ up to z=0.3. Additionally, we observed that Cluster-GMDN performs better in surveys with galaxies spread across a wider redshift range, for the same reason. If the redshift ranges we worked with are too narrow, the significance of galaxy clustering diminishes, as there may be fewer galaxies between our target galaxy and us.

Finally, considering the novelty of the Cluster-GMDN model and the time constraints associated with a master's thesis, there remains a considerable amount of work to be done. The following future steps, building upon this study as a starting point, are suggested:

- One suggestion is to implement recursion in the Cluster-GMDN model, using the new PDFs to improve the accuracy of the γ-MDN to be used again in the Cluster-GMDN PDF, until convergence.
- An analysis of potential redshift biases that may be overlooked when generating group PDFs with Cluster-GMDN and selecting the most probable peak should be conducted. Introducing randomness when dealing with multimodal PDFs could be one possible solution. Alternatively, a statistical analysis of probability differences among peaks should be performed.
- A comprehensive analysis is crucial to determine whether there are patterns among galaxies where the implementation of Cluster-GMDN may yield worse results. It is essential to identify situations where its usage is advantageous and when it is preferable to solely apply γ-MDN in order to determine the optimal conditions for its application and achieve the best possible outcomes.
- Incorporating the concept of conformity [Kerscher, 2018] when generating group probability density functions (PDFs) by analyzing the properties of galaxies within clusters would be interesting. This approach could provide insights into, for example, the stronger gravitational connections of red galaxies to clustered environments. This consideration has the potential to improve the performance of Cluster-GMDN.

This work is expected to yield a minimum of two papers, showcasing its two most significant contributions: the application of γ-MDN to 2MPZ and the introduction of the Cluster-GMDN model to the scientific community.

References

Abbott, T., Abdalla, F., Aleksić, J., Allam, S., Amara, A., et al. (2016). The dark energy survey: more than dark energy–an overview. *Monthly Notices of the Royal Astronomical Society*, 460(2):1270–1299.

Abbott, T. M. C. et al. (2021). The Dark Energy Survey Data Release 2. *ApJS*, 255(2):20.

Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., and Steward, L. (2013). Two micron all sky survey photometric redshift catalog: a comprehensive three-dimensional census of the whole sky. *The Astrophysical Journal Supplement Series*, 210(1):9.

Bishop, C. M. (1994). Mixture density networks. *Aston University, Dept. of Computer Science and Applied Mathematics,*. Tech. Rep. NCRG/94/004.

Cerdán, L. (2020). Prediction of galaxy distances using mixture density networks in photometric surveys and its application to the year 3 dark energy survey data. Master's thesis, Universidad Carlos III, Madrid.

Chiu, I.-N., Klein, M., Mohr, J., and Bocquet, S. (2023). Cosmological constraints from galaxy clusters and groups in the erosita final equatorial depth survey. *Monthly Notices of the Royal Astronomical Society*, 522(2):1601–1642.

Collister, A. A. and Lahav, O. (2004). ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *PASP*, 116(818):345–351.

De Vicente, J., Sanchez, E., and Sevilla-Noarbe, I. (2016). Dnf–galaxy photometric redshift by directional neighbourhood fitting. *Monthly Notices of the Royal Astronomical Society*, 459(3):3078–3088.

Dodelson, S. and Schmidt, F. (2021). *Modern Cosmology*. Elsevier, London.

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(2):243–268.

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M. (2005). Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759.

Kerscher, M. (2018). Spatial range of conformity. *Astronomy and Astrophysics*, 615:A109.

Kuruoğlu, E. E., Molina, C., and Fitzgerald, W. J. (1998). Approximation of α-stable probability densities using finite gaussian mixtures. In *9th European Signal Processing Conference (EUSIPCO 1998)*, pages 1–4.

Padmanabhan, N., Xu, X., Eisenstein, D. J., Scalzo, R., Cuesta, A. J., Mehta, K. T., and Kazin, E. (2012). A 2 per cent distance to z= 0.35 by reconstructing baryon acoustic oscillations – i. methods and application to the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 427(3):2132–2145.

Ponce Aguilar, R. (2015). *Magnificación cósmica en el Dark Energy Survey*. PhD thesis, Universidad Complutense de Madrid.

Porredon, A., Crocce, M., Fosalba, P., Elvin-Poole, J., Carnero Rosell, A., Cawthon, R., Eifler, T. F., Fang, X., et al. (2021). Dark energy survey year 3 results: Optimizing the lens sample in a combined galaxy clustering and galaxy-galaxy lensing analysis. *Phys. Rev. D*, 103:043503.

Rosell, A. C., Rodriguez-Monroy, M., Crocce, M., Elvin-Poole, J., Porredon, A., Ferrero, I., Mena-Fernández, J., Cawthon, R., De Vicente, J., Gaztanaga, E., et al. (2022). Dark energy survey year 3 results: galaxy sample for bao measurement. *Monthly Notices of the Royal Astronomical Society*, 509(1):778–799.

Sanchez, C., Kind, M. C., Lin, H., Miquel, R., Abdalla, F. B., Amara, A., Banerji, M., Bonnett, C., Brunner, R., et al. (2014). Photometric redshift analysis in the dark energy survey science verification data. *Monthly Notices of the Royal Astronomical Society*, 445(2):1482–1506.

Sherwin, B. D. and White, M. (2019). The impact of wrong assumptions in bao reconstruction. *Journal of Cosmology and Astroparticle Physics*, 2019(02):027.

Soch, J. (2020). Proof: Cumulative distribution function of the gamma distribution. The Book of Statistical Proofs. [online] `https://statproofbook.github.io/P/gam-cdf.html`.

Thom, H. C. (1958). A note on the gamma distribution. *Monthly weather review*, 86(4):117–122.

Wiemann, P. F., Kneib, T., and Hambuckers, J. (2021). Using the softplus function to construct alternative link functions in generalized linear models and beyond. *arXiv preprint arXiv:2111.14207*.

Željko Ivezić, Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., , et al. (2019). Lsst: From science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111.

A Appendix

A.1 DES

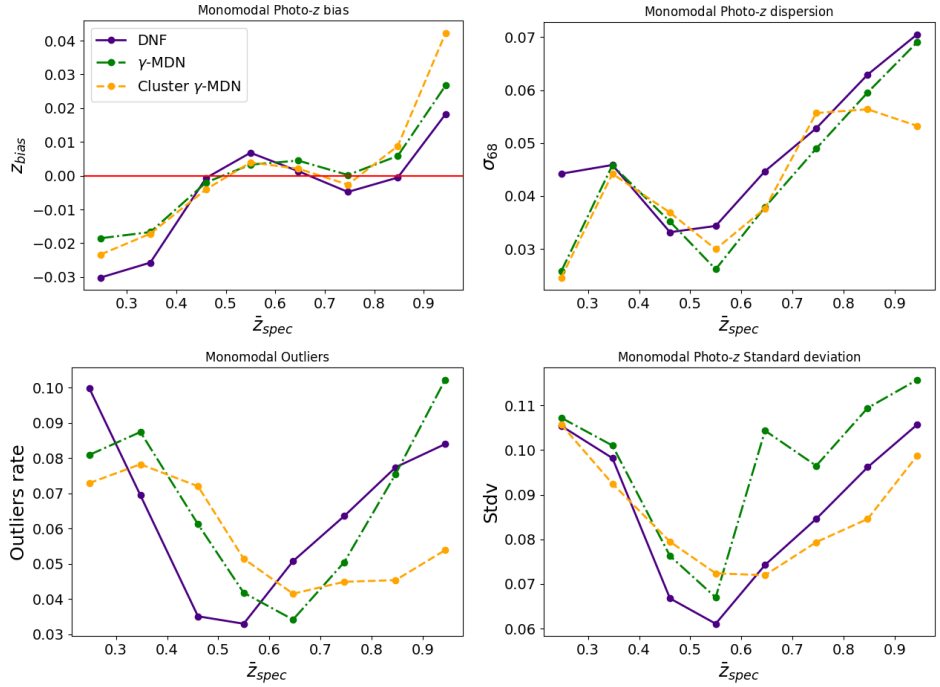

Figure 21: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for *DNF* (purple), γ-MDN (green), and Cluster-GMDN (orange) for galaxies with unimodal PDF distributions as measured by γ-MDN.

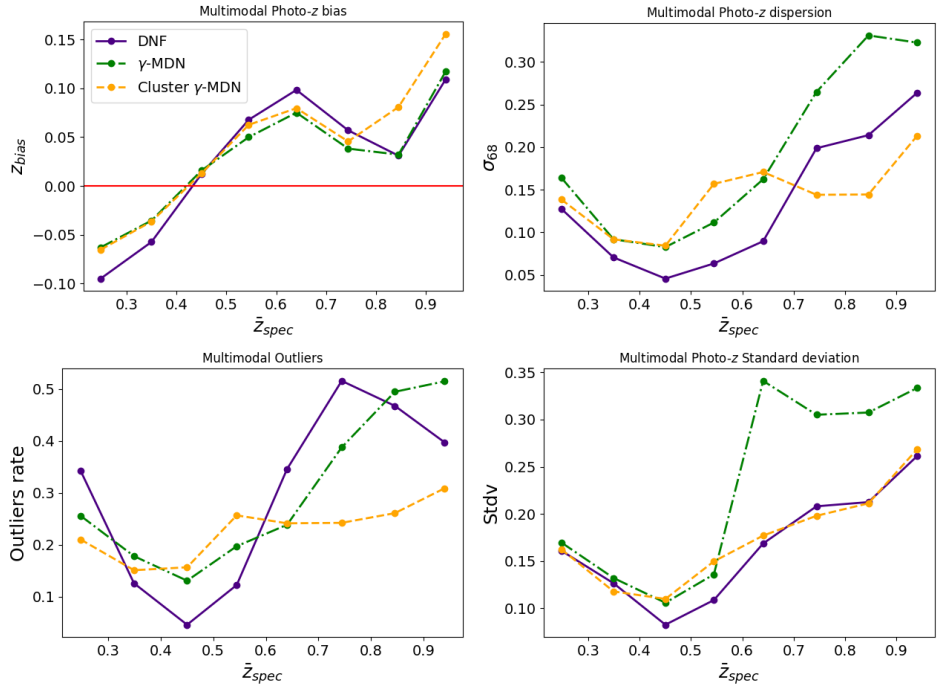

Figure 22: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for the predictions of *DNF* (purple), γ-MDN (green), and Cluster-GMDN (orange), for galaxies with multimodal PDF distributions as measured by γ-MDN.

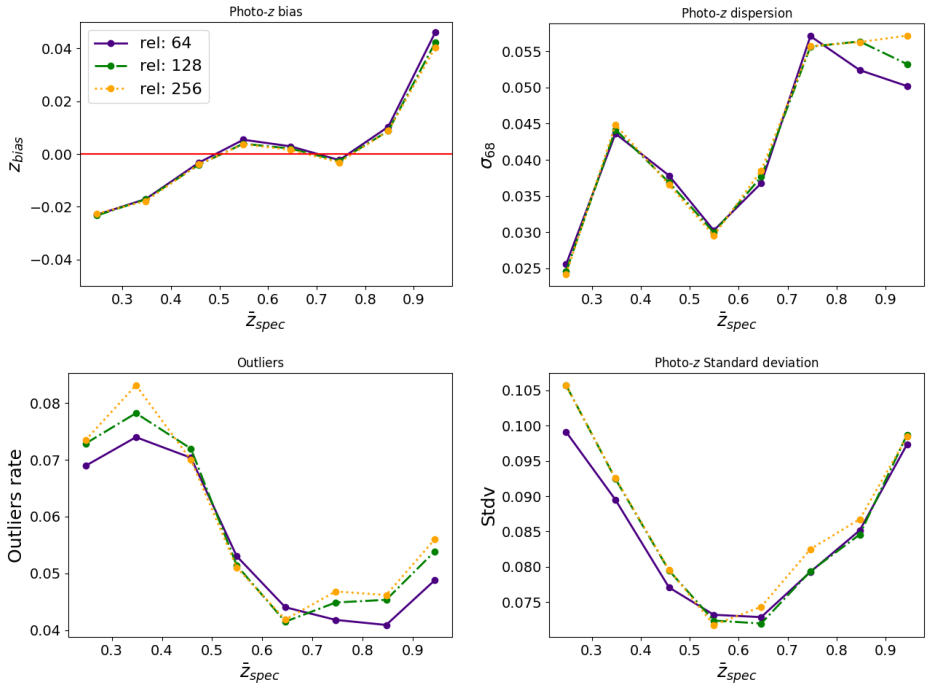

Figure 23: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for Cluster-GMDN at different *HEALPix* resolutions.

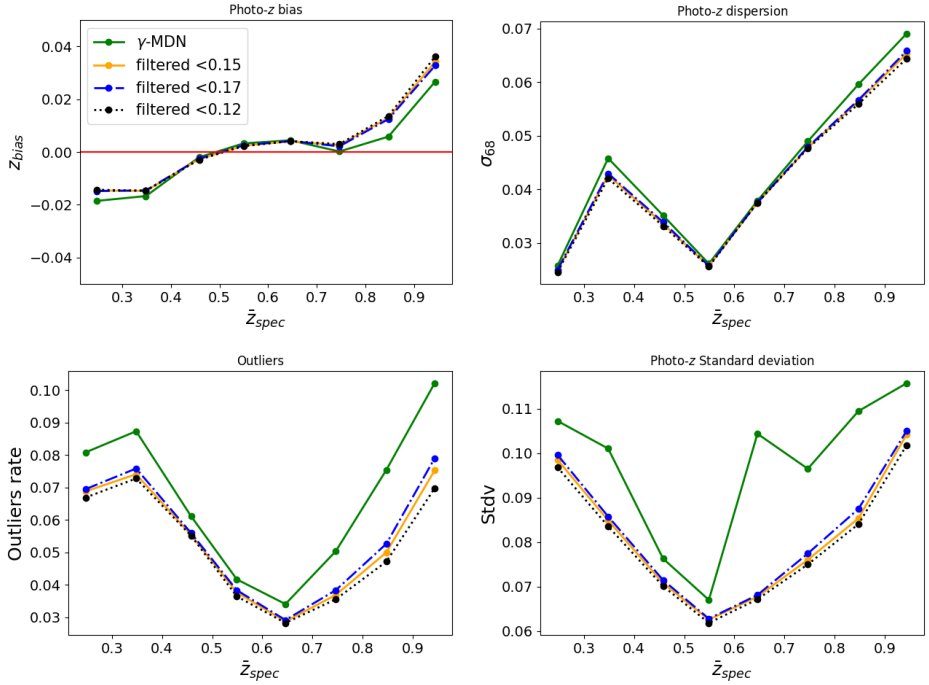

Figure 24: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for γ-MDN (green) and the outlier-cleaned γ-MDN using Cluster-GMDN, for different outlier thresholds.

A.2 2MPZ

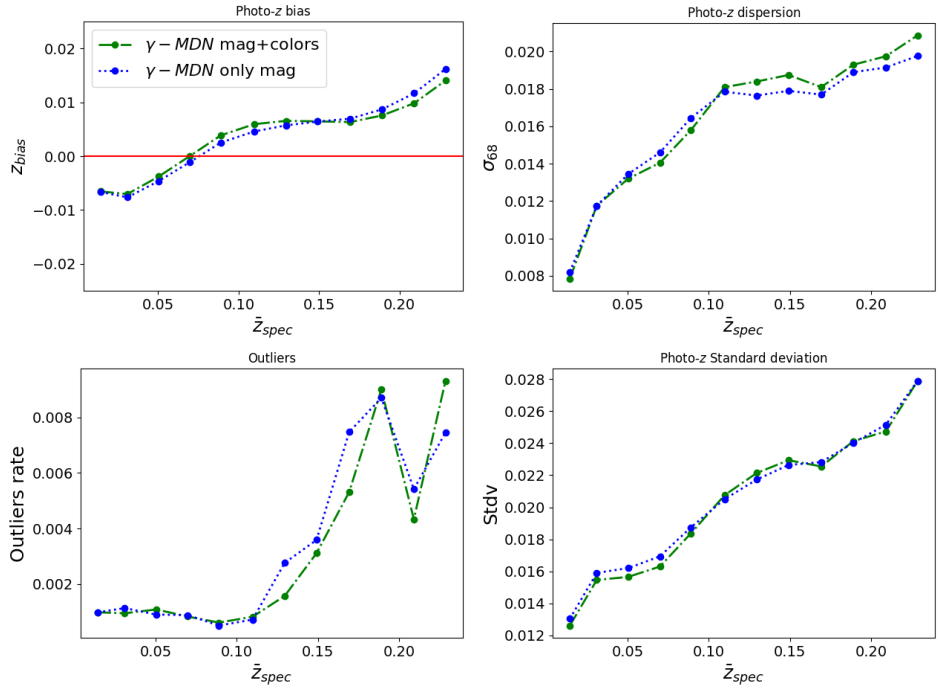

Figure 25: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for γ-MDN using only magnitudes (blue) and using magnitudes and colors (default configuration, green).

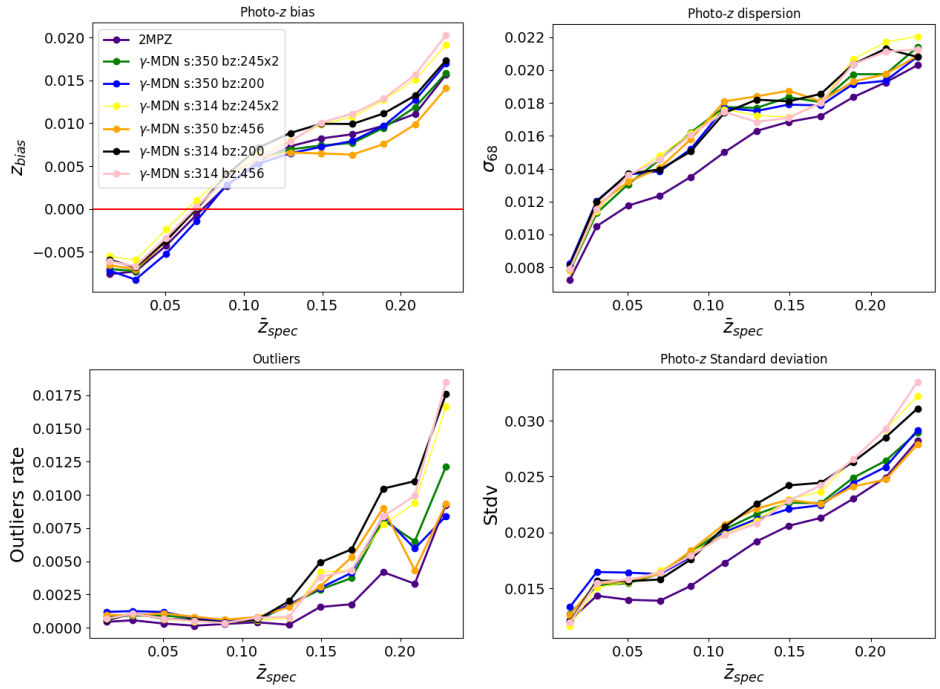

Figure 26: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for γ-MDN for different random seeds and batch sizes.

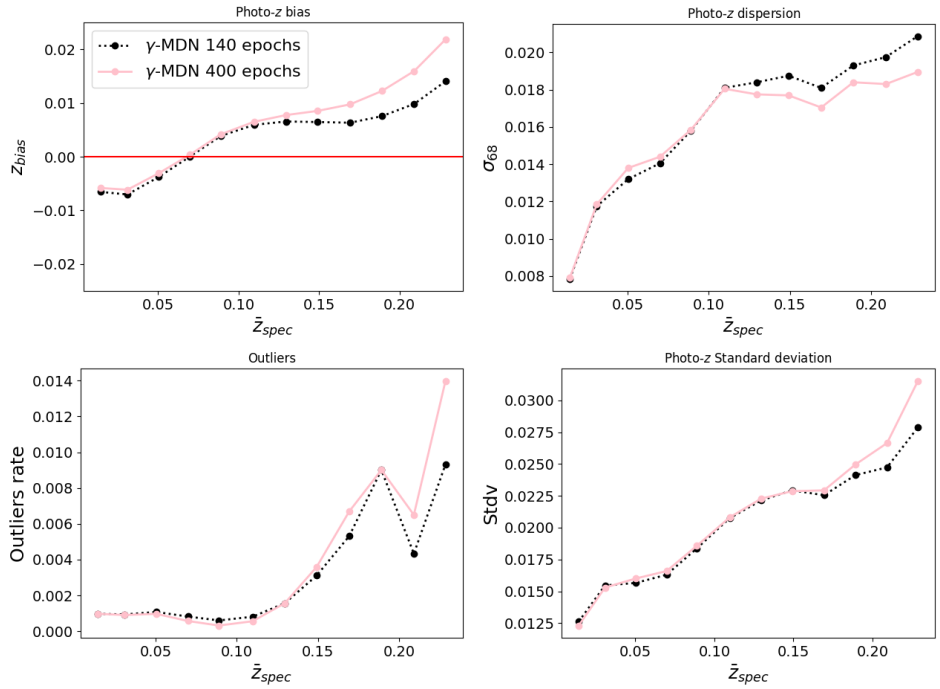

Figure 27: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for γ-MDN, trained up to 140 epochs (pink) and trained up to 400 epochs for the same batch size.

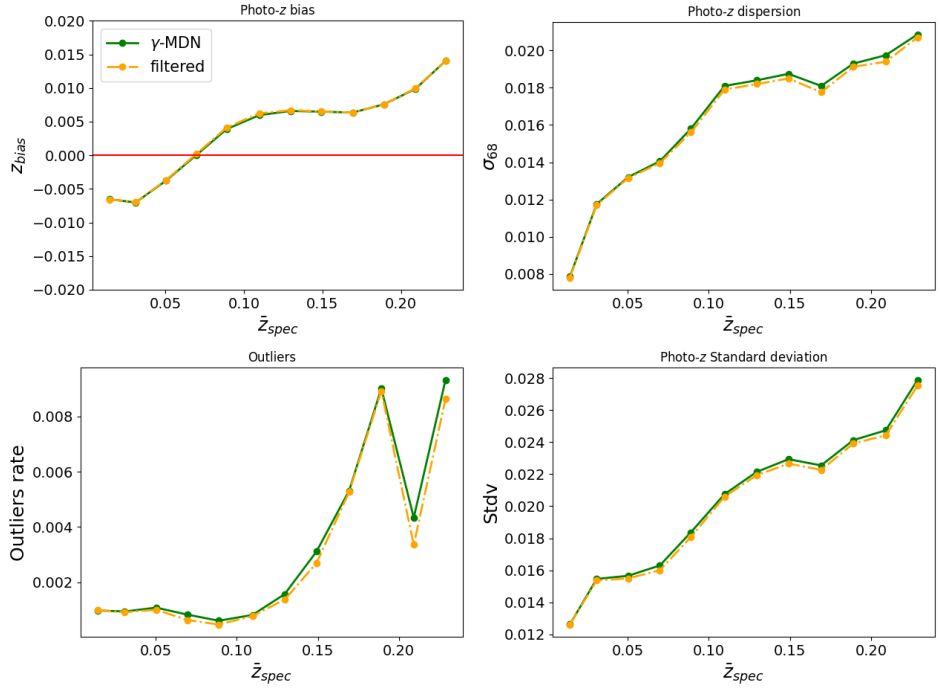

Figure 28: Photo-z bias (top left), σ_{68} (top right), outliers rate (bottom left), and standard deviation (bottom right) as a function of true redshift for γ-MDN model (green) and the outlier-cleaned γ-MDN model using the Cluster-GMDN results (orange).