THE GALAXY MORPHOLOGY-DENSITY RELATION AT HIGH REDSHIFT WITH CANDELS

by

Dritan Kodra

B.Sc., Aristotle University of Thessaloniki, 2010,

M.Sc., Aristotle University of Thessaloniki, 2012

Submitted to the Graduate Faculty of
the Kenneth P. Dietrich School of Arts and Sciences in partial
fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Dritan Kodra

It was defended on

April 25th 2018

and approved by

Jeffrey Newman, Prof., Physics & Astronomy, University of Pittsburgh

Adam Leibovich, Associate Dean for Faculty Recruitment and Research Development, University of Pittsburgh

Ann Lee, Prof., Department of Statistics, Carnegie-Mellon University

Arthur Kosowsky, Prof., Physics & Astronomy, University of Pittsburgh

Michael Wood-Vasey, Prof., Physics & Astronomy, University of Pittsburgh

Dissertation Director: Jeffrey Newman, Prof., Physics & Astronomy, University of Pittsburgh

THE GALAXY MORPHOLOGY-DENSITY RELATION AT HIGH REDSHIFT WITH CANDELS

Dritan Kodra, PhD

University of Pittsburgh, 2019

One of the biggest open questions regarding the evolution of the galaxy population over time, is how their properties (such as their morphologies) are affected by their local environment, e.g. the density of matter in the region where they are found. In the local universe, studies have shown that elliptical galaxies are found predominantly in the central regions of galaxy clusters where densities are higher, while disk galaxies reside in regions of lower densities such as the edges of clusters and the low-density ”field”. We investigate if this pattern continues to exist at earlier times by using data from the CANDELS collaboration at redshifts up to $z \sim 3$. For this, we make use of photometric redshift probability distributions (photo-z PDFs) for the galaxies observed by CANDELS. This required the development of new statistical methods to improve the quality of the PDFs measured by the CANDELS team, described in the thesis. We have used 3D-HST grism redshifts as well as spectroscopic redshifts where available, to test and optimize techniques for combining PDFs determined from multiple methods. We use morphological catalogs provided by the CANDELS team to select galaxies from three main categories: spheroid, disk, and irregular galaxies. We investigate the relative clustering of these different morphological types by estimating two-point cross correlation functions of each type with the full sample of CANDELS galaxies. Our results show that spheroid galaxies still cluster more strongly than disk galaxies at small separations at higher redshifts, while at larger separations the difference in their clustering amplitudes is not statistically significant. At the highest redshifts studied, clustering measurements are too noisy to detect differences in clustering strength, if any persist.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Our view of the universe has drastically changed over the past one hundred years or so. For millennia ancient civilizations had developed various mythological theories regarding the start and evolution of the cosmos as was known to them. For most of the recorded history humankind believed that our planet was the center of the universe, and all objects on the night sky revolved around it. Then, with the help of Galileo, Copernicus, and Kepler, astronomers shifted their theories towards a Heliocentric universe. Although this was a huge step in the right direction of the true picture, up until the beginning of the twentieth century, astronomers believed that all celestial objects that were visible at the time were part of our Galaxy, the Milky Way. Then, it became clear that the "nebulae" which were known to them, were too far away to belong to the Milky Way (Hubble (1926), Hubble (1936)), and extragalactic astronomy was born. Soon after, with the help of technological advancements other galaxies were observed and cataloged, which led to a better understanding of the cosmos.

Today we know that our galaxy is only a drop in the vast ocean of galaxies in our universe, and the visible components of galaxies represent a small fraction of the 5% of the total mass-energy content of the Universe that consists of ordinary or "baryonic" matter. Another 25% of the total mass-energy content is made up of a component known as dark matter; it interacts with baryonic matter primarily or exclusively via the gravitational force. Finally, the nature of the remaining 70% is also unknown to us today; it appears to be responsible for the observed accelerating expansion of the universe, and is generally labelled as "dark energy" (Mo et al. (2010)). Given that we cannot observe dark matter and dark energy directly, any information we have about them is obtained by observing what we can, the baryonic matter.

Due to gravitational instabilities, baryonic matter collapses wherever the density is sufficiently high, forming dense structures such as galaxies, which in turn can group together to make up the largest gravitationally bound structures in the universe, called galaxy clusters. The mechanisms responsible for the formation of galaxies are yet to be fully understood; how they evolve with time is less well known. In order to make any progress towards such discoveries, we first need to identify and classify the objects of interest, which is why we present a brief introduction of the wide variety of galaxies in the following section.

1.1 CLASSIFYING GALAXIES

Galaxies have been observed to have a wide range of properties such as different ages, sizes, shapes, and colors even at constant mass; the reasons for this large variety are still not very clear. One thing that is apparent is that these properties are not entirely independent from one another. That is, the most massive galaxies tend to be older, redder, larger in size, and more commonly elliptical in shape. On the other hand, lower-mass galaxies tend to be younger, bluer, somewhat smaller in size, and disk shaped or irregular/peculiar. It is fairly certain that these properties of galaxies and how they evolve with time depend not only on baryonic matter and its distribution in the universe, but also on the properties of dark matter and dark energy, which is why the study of galaxies is of utmost importance for the study of the universe itself and therefore, cosmology. It is only natural for their study to start with some of their most obvious properties, such as their shape and color.

1.1.1 Galaxy Morphology

Since before realizing that the nebulae observed in the night sky were individual galaxies themselves, astronomers were fascinated with the shapes of these large structures of the universe. They appeared to have very noticeable differences, but at the same time many of them appeared to be very similar. Therefore it was clear that they can be divided into groups of similar appearances, an endeavor that occupied some of the brightest astronomers of the twentieth century, such as Wolf, Hubble, de Vaucouleurs, Reynolds, etc. While different scientists used different ways of categorizing galaxies based on morphology, the most commonly known is the Hubble Sequence (HS; Hubble (1926), Hubble (1936), Holmberg (1958), de Vaucouleurs (1958), van den Bergh (1960)), which initially separated them into three major categories: ellipticals (E), spirals (S), and irregulars/peculiars (Irr) [Figure 1.1].

We continue with a brief description of the main morphological types of galaxies; for more details we refer the reader to Section 2.3 of Chapter 2 of Mo et al. (2010), and Chapter 3 of Schneider (2006).

Elliptical galaxies received their name due to their isophotal curves being ellipses. They are generally very smooth systems with lack of structures and brightness profiles that decline continuously with distance from the center, described by the de Vaucouleurs profile. They appear to have various ellipticities, which is the reason why their labelling E is followed by the integer n, taking values between 1 and 7. This integer represents the galaxy's ellipticity and is given by $n = (1 - b/a) \cdot 10$, where b/a is the minor-to-major axis ratio of the galaxy.

Spiral galaxies show very clear features known as spiral arms, hence the name of their category. The spiral arms may spring from a central nucleus in which case they are referred to as normal spirals and labelled S, or the arms could start from a central region that resembles a bar, in which case galaxies are called barred spirals and they are labelled SB. Members of both of these two subgroups are further subdivided into smaller groups based on the state of their spiral arms, with the three main subgroups being a, b, and c. Sa and SBa galaxies have closely coiled spiral arms, Sb and SBb ones have more open arms, and Sc and SBc ones have quite open arms. Later updates to the initial Hubble Sequence by Hubble himself, de Vaucouleurs, and Sandage, added more subclasses such as Sd, SBd, etc., which was possible due to greater details in images obtained from better observations.

Another subgroup of the HS is the one containing the so called *S0* galaxies, which have similarities both with ellipticals and spirals. They have a lenticular shape and have sometimes been speculated to be galaxies in a transitional phase from spiral to elliptical galaxies. In the HS they appear to the right of E7 galaxies, and mark the spot of the separation between the ”normal” and ”barred” spirals. The S0 galaxies provide a sense of continuity on the HS, where the shifts from one subgroup to the next is somewhat smooth with the members

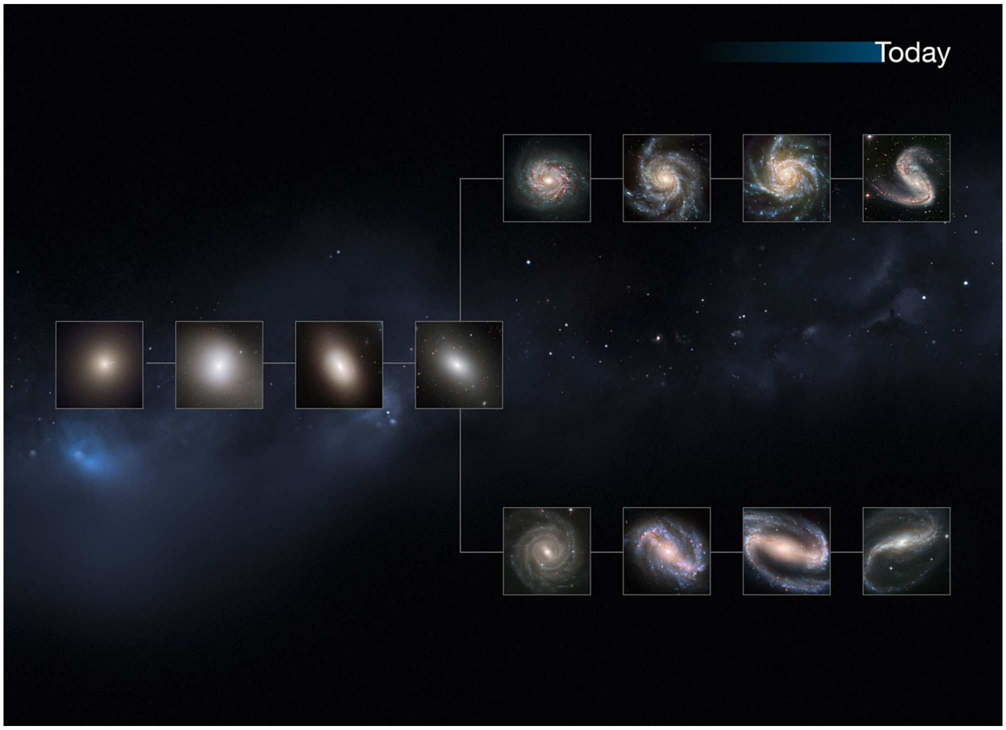


Figure 1.1 Hubble Sequence (HS) today, with elliptical galaxies on the left, and spiral galaxies on the right; S0 galaxies in the middle. Spiral galaxies separate into the regular (upper row) and barred (lower row) groups. As we go from left (early-types) to right (late-types), galaxies appear to be bluer in color due to young massive blue stars formed in spiral galaxies. Credit: NASA, ESA, M. Kornmesser

having plenty of similarities among them.

Finally, there is the group of *irregular* galaxies (Irr). They are galaxies with irregular shapes (hence the name of their class), meaning that they do not show any elliptical or spiral symmetries of any sort, and therefore constitute a separate group from them. This class often contains galaxies which are characterized as peculiar, usually consisting of galaxies that are in a merging state with one or more other galaxies and hence appear abnormal in shape.

An additional set of terminology used to describe galaxy types is a temporal one. Going from left to right on the Hubble sequence they are labeled as early-type, intermediate-type, and late type. This language is misleading, since it is not connected with the ages of galaxies. Instead, this terminology was borrowed from the classification of stars; short-lived, massive, blue stars placed on the left of the Hertzsprung-Russell (HR) diagram are considered early-type ones, whereas longer-lived, lower mass, red stars that appear on the right of the HR diagram are referred to as late-type stars. The early, intermediate, and late terms were based on early theories of the stages of the lives of stars. It is ironic to note that early-type galaxies (E/S0) are usually made exclusively of old, red, late-type stars, whereas late-type galaxies (Sc, SBc, Irr) often contain many young, blue, early-type stars. This shows again the unfortunate nature of this temporal nomenclature for galaxies; we emphasize that this characterization does not indicate in any way an evolution or a time scale of any kind.

1.1.2 Galaxy Color

Another important visual property of galaxies is their color, which is closely related to their constituents, i.e., dust, gas, and stellar populations. When a galaxy is primarily made of old, red stars, it will also appear reddish in color when observed. In contrast, if another galaxy includes young, blue stars, it will then appear to be more blue when observed than the previous case.

The color of a galaxy in Astronomy is defined as the difference in magnitudes as measured in two different spectral bands, e.g. B-V where B is the magnitude as measured in the B-band and V is the one measured in the V-band. An interesting segregation is noticed when the colors of galaxies are plotted against their absolute magnitudes, an important visual tool that is usually referred to as the color-magnitude diagram for galaxies, and is analogous to the H-R diagram for stars (de Vaucouleurs (1961)). In this diagram galaxies separate into two very distinct groups, with one group favoring bluer colors (i.e. lower values of B-V), and the other consisting of galaxies with redder colors (i.e., higher values of B-V).

Given that the calculation of the absolute magnitude of a galaxy requires the knowledge of its distance, the color-color plot is in some cases easier to construct in order to see this bimodality of galaxies. For it we plot the colors of galaxies in two separate bands versus the colors in two other bands, e.g., U-B vs. B-V. Since colors are constructed from the differences between magnitudes, which in turn are related logarithmically to the ratio of fluxes, any dependence on the distance cancels out, and therefore it is not needed. This of course is only true for nearby galaxies, since for distant galaxies their colors are shifted to the redder wavelengths, as we will discuss in a later section, in which case we need to estimate the rest-frame colors of galaxies. In the color-color plot the two separate groups of galaxies are again distinguishable, with galaxies belonging to the "blue cloud" or the "red sequence", depending on their respective colors (Strateva et al. (2001); Blanton et al. (2005); Schawinski et al. (2014)).

With the recent increase of data, scientists have also shown interest in region of the color-magnitude diagram that is referred to as the green valley, which is the region of transition between the blue cloud and red sequence. Recent studies have suggested (Licquia et al. (2015)) that our own galaxy may be a member of this population.

Additionally, it has been noticed that the colors of galaxies are closely linked to other properties of them, such as their morphologies (Strateva et al. (2001)). For instance, galaxies in the red sequence tend to be early-type ones (E/S0), whereas the blue cloud usually consists of late-type ones (S/SB & Irr).

Besides morphology and color, galaxies can be separated based upon a variety of other properties, such as their star-formation rates, masses, sizes, etc. As we have already mentioned, it is clear that these properties are not completely independent from one another, and their correlation is of great interest to the scientific community, since it could shed light towards the mechanisms that drive galaxy formation and evolution. One important aspect regarding galaxy properties is how they are affected by the environment where different galaxies are observed. It appears that different types of galaxies of different characteristics are found in different regions of the universe; some favor low-density regions (the 'field') whereas others favor groups and clusters of galaxies. The trends observed in the local universe may change significantly as we look at more distant objects, since larger distances correspond to earlier times when structure and galaxies were less developed. Therefore it is crucial not only to study the dependence of galaxy properties on environment but also how it evolves with time.

Before we start our discussion regarding the structure of the universe and its evolution, we must first define one of the most important quantities in Astronomy that allows us to describe large distances and past times. This quantity is generally known as the *redshift*; we briefly introduce this concept in the following section.

1.2 REDSHIFT

As the name suggests, the term 'redshift' refers to a shift of the emitted light of distant objects to larger (corresponding to redder) wavelengths. This can happen due to the Doppler effect when objects are moving away from us with peculiar velocities, or due to the expansion of the universe resulting in the stretching of wavelengths of light. The former can be significant only for the very nearby objects whereas the latter dominates for larger distances. Mathematically, redshift is defined as $z = (\lambda_o - \lambda_e)/\lambda_e$, where λ_o is the observed wavelength and λ_e is the emitted wavelength. In the case of nearby objects, this can be approximated by $z \approx v/c$, where v is the object's velocity with respect to the observer, and c is the speed of light. In this work we focus on distant galaxies, for which a more useful definition is the one related to the scale factor $a(t)$. The scale factor is a function of time and determines the emitted and observed wavelengths mentioned above according to $\lambda_o/\lambda_e = a(t_o)/a(t_e)$, where $a(t_o)$ is the scale factor at the time of observation and $a(t_e)$ the one at the time of emission of the photons. This in turn defines redshift as $z = [a(t_o) - a(t_e)]/a(t_e)$, which combined with the fact that at the present time that we observe the universe the scale factor is defined to be unity, leads to $z = [1 - a(t_e)]/a(t_e)$ (Hogg (1999)).

One way to estimate the redshift of a galaxy is by analyzing its spectrum, and observing the shifts in emission and/or absorption lines from their restframe wavelength, which in general yields very accurate results (Newman et al. (2013); Weiner et al. (2005)). Unfortunately, the acquisition of the spectra of galaxies is not an easy task, since it requires long exposure times, while the number of objects that are targeted during each observation is limited, making it practically impossible for surveys such as LSST (Ivezic et al. (2009)). In addition, the most distant galaxies are too faint to be observed spectroscopically, since their light cannot be analyzed in detail.

On the other hand, distant galaxies can be observed using broadband photometry, which is the measurement of the fluxes of all objects in the field of view of the instrument using different filters (Nayyeri et al. (2017); Stefanon et al. (2017); Guo et al. (2013); Galametz et al. (2013)). With this method, many objects can be observed at the same time, and measurements are feasible even for faint objects, thus making it the only solution for large surveys such as the aforementioned LSST. Additionally, the flux measurements can be accomplished with much shorter exposure times than what is required for detailed spectra. The caveat to this is the fact that the information is less detailed; as a result the photometric redshifts of galaxies are not as precise as their spectroscopic counterparts.

Having defined redshifts, which serve as the primary proxy for time in our work, we continue by describing the formation of galaxies, their temporal (or redshift) evolution, and how this relates to the environment a galaxy is found in in the following section.

1.3 OVERVIEW OF GALAXY FORMATION AND EVOLUTION

There has been a long history of research related to the formation and evolution of galaxies and the physical processes responsible for their vast variety we observe in the local and distant universe today. The picture is still being developed, although great progress has been made during the past few decades in this field (Dressler (1980); Postman & Geller (1984); Goto et al. (2003); Sheth et al. (2006); Skibba et al. (2013)). In this section we provide a brief summary regarding the main mechanisms we now believe are playing important roles in shaping the structure of the universe into what is found everywhere in the cosmos.

1.3.1 Formation

We start this section with an overview of structure formation; for more details we refer the reader to Section 1.2 of Chapter 1 of Mo et al. (2010), the basics of which we summarize here.

According to the standard model of structure formation (Peebles (1980)), during the first stages of the universe the slightly overdense regions of dark matter collapse due to self-gravity forming small structures, usually referred to as dark matter halos (or simply halos). Because dark matter is believed to be collisionless and interacts only gravitationally, this process happens very early in cosmological time, and the regions with higher density will collapse and evolve faster than the less dense ones. According to the hierarchical scenario (Frenk & White (2012)), which is the most commonly accepted one, smaller sized halos form initially and they act as hosts for baryonic matter to also collapse gravitationally and form galaxies. This happens when the gas in the progenitor clouds cools and the pressure gradient cannot sustain hydrostatic equilibrium, therefore leading to the collapse (Dalgarno & McCray (1972)).

Which mechanisms are predominantly responsible for this cooling depends on the characteristics of the environment and more specifically on the virial temperature ($T_{\rm vir}$) of the dark matter halo, which in turn depends on its mass ($M_{\rm h}$). For massive halos with high temperatures $T_{\rm vir} \geq 10^7$K gas is fully ionized, making Bremsstrahlung emission of free electrons the main mechanism of cooling. In the regime of $10^4\text{K} < T_{\rm vir} < 10^7\text{K}$ cooling happens mainly through atoms that end up on the ground states by radiation emission or via the capture of free electrons by ions to form atoms followed by radiation emission. For lower temperatures $T_{\rm vir} < 10^4$K, atoms are no longer ionized and cooling becomes very inefficient, though the presence of large molecules can lead to emission via vibrational or rotational de-excitation, while the presence of heavy elements can lead to cooling through the fine and hyperfine structure lines of de-excitation. In addition to these temperature-dependent methods of cooling, at very early times in the universe, hot halos of gas can cool through inverse

Compton scattering from the interaction of the free electrons with the cosmic microwave background photons (Kraljic (2014)).

When gas cools sufficiently, it will tend to gravitationally collapse, potentially leading to the eventual formation of a galaxy (Greif et al. (2008)). While the details are still not entirely clear, it is understood that variations in the properties of the host dark matter halos as well as the influence of other galaxies will guide the formation of new galaxies and determine their characteristics such as their shapes, colors, star formation activities, etc. Based on observations of the early universe it is believed that the majority of the first galaxies formed were small in size and irregular in shape, with a small portion of them being spheroids, which could explain the existence of dwarf spheroidal galaxies we observe today, which contain populations of old stars.

If the initial cloud of gas within a dark matter halo had a significant angular momentum, the cooling and collapse of the gas could lead to the formation of a disk (Firmani & Avila-Reese (1999); Firmani & Avila-Reese (2003)). Further collapse of gas clouds within a disk can ignite star formation; the stars themselves can influence the surrounding gas by ejecting into it material and heavier elements produced in their interiors, a process called feedback. This process has been proposed to overcome difficulties that arise regarding the formation of disk galaxies, such as the angular momentum catastrophe (Maller & Dekel (2002)) and the overcooling problem (Benson et al. (2003)). The former refers to the fact that mechanisms such as the dynamical friction between baryonic and dark matter can lead to a significant angular momentum loss therefore the disk would not be sustainable, whereas the latter is related to the first calculations that showed that in an environment where the cooling of gas is efficient, it would lead to its accumulation in the central region of the new formed galaxy, resulting in a bulge rather than a disk shape. Feedback processes (especially feedback from early supernovae, or SNe) can overcome both of these difficulties, since the energy ejected can heat the gas and slow down its collapse in order to preserve the disk, while it can also eject low angular momentum material outside of the region of the forming galaxy, thus maintaining the required angular momentum. While supernova feedback (Meier (1999, 2001); Schawinski et al. (2007)) in combination with the photoionization due to the UV radiation from massive early stars and quasars (Quinn et al. (1996); Gnedin (2000)) can prove to overcome some of the difficulties, it cannot completely solve the issues related with the formation of disks in galaxies.

The existence of the other morphological types observed in the universe is believed to be due to the evolution of early galaxies into other types following various physical mechanisms which will be described in the subsection below. After that we will compare their relative strengths and regions of dominance of these different mechanisms.

1.3.2 Evolution

As described above, the hierarchical bottom-up model suggests that smaller structures formed first. Dark matter halos host galaxies which during the early dense stages of the universe will attract each other into forming pairs, groups, or even galaxy clusters, which are the largest virialized structures of ordinary matter in the universe, with densities much higher than the average. In these environments, galaxies can undergo a variety of processes which can evolve them into becoming members of different morphological types, or they can suppress the formation of new stars, as well as strip galaxies off their gas contents. We proceed with a brief overview of the main mechanisms driving this evolution. For a more detailed discussion we refer the reader to Boselli & Gavazzi (2006); we summarize the mechanisms described in that work below.

One of the most important processes believed to be responsible for galaxy evolution is mergers between galaxies, in which two (or occasionally even more) galaxies can combine together, forming a single larger object. If the two initial galaxies involved are of similar mass, the merger is called major, and studies have shown that such a process can lead to the formation of elliptical galaxies, regardless of the morphology of the progenitor galaxies (White (1978), Hernquist et al. (1993)). If on the other hand, one of the two merging galaxies is significantly more massive than the other, then the process is usually referred to as a minor merger, and the final product can have properties similar to the larger-mass object (Diaferio et al. (2001)). Mergers are more likely to take place when galaxies are found in groups rather than clusters, as in the latter the number densities are not significantly larger but the peculiar velocities can be very high. These groups may eventually become part of a cluster, something that is supported by various observations indicating that clusters have substructures resembling groups of galaxies within them (Colless & Dunn (1996); Rines et al. (2003)). The stage of galaxies interacting and merging with each other before becoming members of a cluster is sometimes referred to as the preprocessing stage (Okamoto & Nagashima (2003); Mulchaey et al. (2005)).

During the interaction of galaxies with each other, tidal effects can significantly disturb them, influencing their properties. If a galaxy is not very compact but has a relatively large radius compared to its distance to other members of the cluster, its interactions with the other members can cause part of the gas from the outer regions of a galaxy to be removed from it, which would then lead to a decrease in its star formation (Byrd & Valtonen (1990), Okamoto & Nagashima (2001), Diaferio et al. (2001)). This process in combination with minor mergers could be responsible for the formation of lenticular galaxies in clusters, though it still cannot account for the observed lenticular galaxies in the field (Springel et al. (2001)).

In a massive galaxy cluster, the gravitational potential of the cluster itself can also have tidal effects on its members, causing disturbances to their structures and internal processes (Merritt (1984), Miller (1986)). It has been suggested that these tidal interactions can also cause part of the interstellar gas to be ejected from the galactic plane of a galaxy, though the fraction of gas actually ejected is usually not very significant. On the contrary, in such situations a large portion of the gas is expected to gather in the central regions of a galaxy, potentially forming bars, as well as inducing star formation and generally increasing its nuclear activity (Byrd & Valtonen (1990), Henriksen & Byrd (1996)).

Given that galaxies in dense clusters can have very high orbital velocities, it is very rare for them to actually merge in the central regions of the cluster, unless the galaxies' velocities are reduced due to dynamical friction, which will eventually lead to their infall towards the center of the cluster where the largest galaxy of the structure usually resides. The eventual fate of such galaxies is to be swallowed by the very massive and large central galaxies, a process that is called galactic cannibalism (Nipoti (2017)). If on the other hand galaxies continue to have very high speeds, then their interactions with the other members will be very short-lived. Nevertheless the very frequent interactions in combination with the interaction with the cluster gravitational potential can lead to a change of galaxy properties via a process called galaxy harassment (Moore et al. (1996, 1998)). Depending on the size and mass of the initial galaxies this process has been suggested to be responsible for the evolution of spiral galaxies into S0s, as well as the evolution of lower size galaxies into dE, dSph and dS0 galaxies (Moore et al. (1999)). It should be noted that while the aforementioned processes have been proposed to have an important role in the transformation of spiral galaxies into S0's, it is now believed that other mechanisms that we discuss below play a bigger part in this evolution.

In addition to the primarily gravitational processes described above, galaxies can also undergo a variety of hydrodynamical processes in dense environments such as clusters which can affect their evolution. The high-velocity galaxy members of a cluster can lose their ISM due to the ram pressure from the IGM of the cluster, a process known as ram pressure stripping (Gunn & Gott (1972)). Various simulations show that while this process can effectively strip the outer layers of gas in the disk of a galaxy, it is very unlikely to completely remove it. Its efficiency depends on various parameters with one of the most important being the inclination of the disk with respect to the direction of motion; disks perpendicular to their direction of motion experience greater ram pressure than disks that are parallel to the plane of the trajectory of the galaxy (Balsara et al. (1994); Abadi et al. (1999); Quilis et al. (2000); Vollmer et al. (2000)). This process could also induce starbursts in a galaxy from the compression of the disk gas (Bekki & Couch (2003); Bekki et al. (2003)), which can eventually lead to reduced levels of gas in the galaxy. Therefore this mechanism could have some responsibility for evolving spiral galaxies into S0s as well as dwarf irregulars into dwarf spheroidals by removing the fuel for star formation (Schulz & Struck (2001)). Furthermore, some studies have shown that ram pressure stripping can cause instabilities in the disk of a galaxy that lead to the formation of spiral arms which in turn can carry angular momentum in the outer regions of the disk and form gas rings (Vollmer et al. (2001); Bekki & Couch (2003)).

Another way that the IGM can extract the ISM from a galaxy in a cluster is through a gas transport process, from the surface of contact of the ISM with the IGM. This process is referred to as viscous stripping (Livio et al. (1980); Nepveu (1981); Nulsen (1982)), and has results similar to ram pressure stripping, where the outer layers of the gas disk can be stripped at different rates depending on whether the flow is laminar or turbulent Nulsen (1982). While ram pressure depends strongly on the IGM density and the orientation of the disk with respect to their direction of motion, viscous stripping efficiency is mainly proportional to the IGM temperature and can be efficient even for galaxies with disks that are parallel to plane of their motion. Both ram pressure and viscous stripping require high velocities to significantly remove gas from galaxies, therefore they are believed to be dominant in the central regions of clusters, where orbital velocities of galaxies are sufficiently large. When the temperature of the IGM happens to be very high compared to the velocity dispersion of a galaxy, then its gas contents can heat up and evaporate through a process characterized as thermal evaporation (Gunn & Gott (1972); Cowie & McKee (1977); Cowie & Songaila (1977)). The effects of this process are stronger for higher IGM temperatures, but they are significantly reduced in the presence of strong magnetic fields (Cowie & Songaila (1977); Sarazin (2009)). Unlike ram pressure and viscous stripping, this process can be effective even at typical galactic velocities within a cluster, removing the cold gas which can be fuel for star formation (Boselli & Gavazzi (2006)). This effect can therefore be at least partly responsible for the existence of quenched galaxies in clusters that are observed both locally and at larger distances.

Often galaxies have outer loosely bound halos of gas that act as sources for gas to slowly fall into a galaxy and keep star formation going. However, if during the life of a galaxy within a cluster this supply is removed, this would lead to the eventual end of star formation within a galaxy, a process also known as galaxy starvation or strangulation (Larson et al. (1980); Treu et al. (2003)). This process has been proposed to be responsible for the formation of S0 galaxies from spirals (Balogh et al. (2000); Bekki et al. (2002)), as well as dwarf ellipticals from low-mass spiral/irregular galaxies (Boselli et al. (2008)). The reason for the second case is that dwarf ellipticals have been observed to have low star formation activities while they have stellar distributions (i.e., density of stars as a function of radius) that are quite different from those of massive ellipticals but similar to those observed dwarf spirals and irregular galaxies; therefore it makes sense to consider the latter as the progenitor galaxies of the former (Mayer (2010)).

1.3.3 Comparison of Evolution Mechanisms

The processes described above are not easily disentangled when studying the evolution of galaxies in the large variety of environments where they are observed in the local and distant universe. In many cases all of them may apply in the formation and evolution of galaxies in clusters, and their importance depends on many factors such as the distance from the cluster center; the size and density of the cluster; the size and density of the member galaxies; the IGM and ISM densities; temperatures and chemical compositions; etc. In this section we attempt to briefly separate these processes and their importance in order to gain a clearer picture of how galaxies may have evolved with time.

Processes such as galaxy-galaxy interactions, galaxy-cluster potential interaction, and galaxy harassment which dominate in a dense cluster environment can all heat up the disks of galaxies and cause them to increase in thickness, but they are not very efficient in removing the gas from the disks, except for cases of small, loosely bound objects. They can also cause perturbations that lead to the transfer of gas towards the central regions of galaxies, thus increasing their nuclear activities (Boselli & Gavazzi (2006)). This would in turn enhance the bulge to disk ratio, which renders these processes capable of explaining some of the morphological differences between S0 galaxies and spirals, though not the quiescence of star formation in the S0s. While galaxy-galaxy and galaxy-cluster potential interactions are more dominant in the central regions of dense clusters ($r < 0.2$Mpc), galaxy harassment is believed to be more effective outside of cluster cores ($r > 0.5$Mpc).

On the other hand, processes that involve interactions of galaxies with the hot IGM cannot boost the thickness of disks or increase the bulge-to-disk ratio. They can however strip part or sometimes even all of the gas in disks, which in turn leads to a decrease in star formation. Processes such as thermal evaporation and laminar viscous stripping depend on the same parameters, though the former is about three times more efficient than the latter (Boselli & Gavazzi (2006)). Ram pressure stripping and turbulent viscous stripping also each have similar dependences on a number of parameters related to the gas mass loss rate, and generally dominate with respect to thermal evaporation and laminar viscous stripping, though in some cases thermal evaporation can be more efficient than the other three (Boselli

& Gavazzi (2006)).

As we have mentioned above, ram pressure and viscous stripping are favored in cases of galaxies with larger sizes and higher velocities, which is the case for the central regions of dense clusters. Thermal evaporation, on the other hand, does not require high velocities and can be efficient for smaller-sized galaxies; its effectiveness depends primarily on the IGM temperature (but not density, radius, or velocity). Hence this process is strong wherever the IGM temperature be high. Evaporation can be effective at all radii in a cluster, though it usually dominates outside of the cluster cores, since in the central regions of dense clusters ram pressure and viscous stripping have enhanced effectiveness. Additionally, this process is not expected to be very effective in lower-mass environments such as groups or diffuse clusters where the lower temperatures decrease its efficiency.

In conclusion, hydrodynamical processes are more efficient in regions where the IGM is denser and hotter; i.e., the central regions of a cluster ($r < 0.2$Mpc). Similarly, the interaction of galaxies with each other and with the cluster potential is strongest in the dense central region of the cluster ($r < 0.2$Mpc). On the other hand, interactions that involve galaxy starvation, galaxy mergers, and galactic harassment are all dominant in the outer regions of clusters ($r > 0.5$Mpc) or in group-like environments where the interaction times are longer, as the typical velocities of galaxies are lower at higher radii or in lower-mass systems.

1.4 THE MORPHOLOGY-DENSITY RELATION

It has been known for some time that different types of galaxies are more dominant in different environments. Dressler (1980) used 51 rich galaxy clusters to show that elliptical and S0 galaxies are more prevalent in regions of higher density than spiral galaxies. Postman & Geller (1984) verified this result showing an increase in the fraction of E and S0 galaxies with increasing surface density (as projected in the sky). Bamford et al. (2009) used about 100k galaxies from Galaxy Zoo and found that both the fraction of ellipticals as well as the fraction of red galaxies would increase in higher surface densities. This picture continued to be true even at larger redshifts as was shown by galaxies from the SDSS (York et al. (2000)), where the picture was very similar up to a redshift of ~ 0.5. This observation is commonly known as the morphology-density relation. This suggests that other properties such as star formation and structure, along with their responsible mechanisms, are also correlated with environment.

An interesting recent finding is that while galaxy structure and galaxy morphology are closely related, their dependence on environment and stellar mass differs significantly. van der Wel (2008) used data from the SDSS of 4594 galaxies and showed that morphology, defined by Sérsic index + "Bumpiness" as was defined in Blakeslee et al. (2006), depends on the local environment at fixed mass. On the other hand, structure as defined only by Sérsic index was found to depend on mass and not on environment. Suppression of star formation could be responsible for this behavior as it should primarily affect the morphology rather than the structure of a galaxy. Poggianti et al. (2008) found similar results using data from the ESO Distant Cluster Survey (EDisCS) at higher redshifts $z = 0.4 - 1$. They showed that the morphology-density relation is equivalent to the star formation-density relation, although for each separate morphological class, the star formation properties they took into account did not appear to show any dependence on environment whereas the morphological properties did vary with environment. This is probably due to the declining fraction of mainly late-type spirals (Sc) in denser environments, whereas early-type spirals (Sa, Sb) showed similar distributions to S0 galaxies. Calvi et al. (2012) investigate at low redshift ($0.03 \leq z \leq 0.11$) how the different morphologies are affected by mass and environment. Their findings showed that for mass-limited samples, the fraction of late-type galaxies decreased smoothly when going from single galaxies to dense clusters, while the opposite happened to early-type ones. The morphology of intermediate mass galaxies did not show any dependence on the stellar mass, but rather on the global environment, whereas the most massive galaxies exhibited both a stellar mass and an environmental dependence of their morphologies. Alpaslan et al. (2015) used a volume-limited sample of galaxies extending up to $z = 0.213$, and their results showed that stellar mass played a more important role in galaxy properties than the environment.

Apart from morphology, a relation between density and the colors of galaxies has also been observed. Cooper et al. (2007) used galaxies from the DEEP2 survey (Davis et al. (2003)) to show that for a wide range of redshift the fraction of red galaxies was larger in higher-density regions. Coil et al. (2008) again used data from the DEEP2 survey, but took a different approach by using projected correlation functions, and finding similar results. The existence of a color-density relation is not very surprising since elliptical galaxies tend to be redder in color than disk galaxies. Additionally, bright ellipticals are more massive, do not show any significant star formation, and are generally older. This makes them the best tracers of massive dark matter halos, since they group together in the dense environments of clusters which reside in the most massive dark matter halos. On the other hand, spiral galaxies tend to be more prevalent in the field, i.e., they tend to not be part of clusters. The question then arises whether the morphology-density or the color-density relation is more fundamental and which originated earlier. In order to shed light on these important yet unanswered questions, we need to observe what happens to the aforementioned relations at earlier cosmological times, which we continue to discuss in the following.

1.4.1 Exploring Density Trends at Higher Redshifts

One question of interest in extragalactic astronomy is whether the relations of galaxy properties with environment observed locally still continue to exist at higher redshifts. There is evidence that suggests that at higher redshifts a larger fraction of spiral galaxies appears in denser regions of clusters, which has led to the belief that perhaps elliptical galaxies form from evolving spiral and irregular galaxies (Moran et al. (2007), Poggianti et al. (2008)).

Galaxies change their properties with time, but their evolution time scales are too vastly large for humans to be able to observe this directly. However, due to the fact that light has a finite speed, it takes some time for it to travel from a distant object to us; as a result, the farther away an object is, the longer the light travel time, and therefore the light of objects we observe today has been emitted in the distant past. We can use this to our advantage in order to study galaxy evolution in a statistical sense, i.e., we can observe galaxy populations at different distances (and thus at different look back times) to indirectly observe their evolution. In order to achieve this, a very large number of galaxies needs to be observed, in order to minimize statistical biases and uncertainties. Large surveys such as 2dFGRS (Lahav et al. (2002)), CANDELS (Grogin et al. (2011) and Koekemoer et al. (2011)), and

SDSS (York et al. (2000)), have observed thousands, hundreds of thousands, or even millions of objects in the night sky, while new upcoming surveys, such as LSST (Ivezic et al. (2009)) plan to observe galaxies of the order of billions in number, extending farther and farther away, up to the first stages of the evolving universe.

Apart from a wide variety of studies of galaxy properties in the local universe, there have been numerous papers on the topic of galaxy evolution at higher redshifts, investigating how galaxies have evolved with time; a number of them have explored the evolution in the relationship between galaxy properties and environment. For instance, Tasca et al. (2009) investigated the dependence on environment of galaxy properties up to redshift of $z \sim 1$ with the zCOSMOS redshift survey. Using volume-limited samples with selected luminosities, they found that the galaxy morphology-density relation was already in place at $z \sim 1$, albeit flatter than at lower redshifts. This finding was in agreement with previous works (Capak et al. (2007)), as well as with studies of the color-density relation at similar redshift (Cooper et al. (2007); Coil et al. (2008)). These studies were able to show that around $z \sim 1$ red galaxies clustered more strongly with the overall population of galaxies than bluer ones, as would be expected if they are associated with denser regions. Furthermore, Kawinwanichakij et al. (2017) used the FourStar Galaxy Evolution (ZFOURGE) survey to investigate how environment and stellar mass affect the star formation activity of galaxies in the redshift range $0.5 < z < 2.0$. They showed that both stellar mass and environment play an important role in the quenching of galaxies at such high redshifts.

There are two main conclusions that can be drawn from this large variety of studies. First they make it clear that galaxies evolve differently in different environments, and second they show that the relationship between galaxy properties and the environments where they are found evolves over time. The purpose of this dissertation is to investigate the dependence of average environment on morphology up to high redshifts ($z \sim 3$) using data from the CANDELS Collaboration. More specifically, we want to investigate how the picture seen in the local universe changes with redshift so we can better understand the mechanisms that drive galaxy formation and evolution at different cosmological epochs.

To do this we need to consider the consequences of the hierarchical bottom-up model of formation as well as the mechanisms for galaxy evolution presented in Section 1.3. As dark matter halos come together to form larger ones, and the galaxies they host are also brought closer together in the process to form groups and clusters, it makes sense to consider that different physical processes will play important roles in galaxy evolution at different epochs and in different environments. For instance at high redshifts we expect the gas in galaxies and clusters of a given mass to be denser and hotter; therefore processes such as ram pressure stripping and thermal evaporation could more rapidly deplete galaxies of their gas reservoirs which will in turn quench their star formation. Since thermal evaporation can effectively remove the gas from small-size galaxies provided that the cluster IGM temperature is sufficient, it is expected to play an important role in the depletion of cold gas and the subsequent quenching of star formation at early times, when galaxies are considered to be less massive and much smaller in size compared to their local counterparts.

Although these processes can generally change the gas contents and star formations of galaxies, that can eventually lead to redder colors, they are generally not very efficient at completely changing their morphologies. However, it should be noted that it takes $\sim 1Gyr$ for a galaxy's color to transform after star formation ceases; as a result quenching galaxies may not have had sufficient time to turn red at higher redshifts, therefore galaxies at earlier times are expected to be generally bluer when observed in their restframe.

On the other hand, while processes such as tidal interactions and galaxy harassment can also lead to losses of gas, they can potentially increase the thickness of the disks of galaxies, therefore contributing in the transformation of spiral galaxies into lenticulars. Additionally, mergers of galaxies would tend to lead to higher fractions of ellipticals at later times through major mergers. Finally, galaxy starvation could be responsible for the transformation of dwarf spirals and or irregulars, found in abundance in the early universe, into dwarf spheroidals.

As we will present in detail in the chapters to follow, our work in general is able to separate galaxies into three main morphological types: spheroids, which generally correspond to elliptical galaxies; disks, which would include spirals and lenticulars together; and irregulars which would include galaxies of irregular shape and/or with tidal tails. Comparing the distribution of these three types of morphologies at different redshifts can help us to explore the main processes responsible for galaxy formation and evolution at redshifts up to $z \sim 3$.

We continue with an overview of this work in the section below.

1.5 DISSERTATION OVERVIEW

This work is focused on improvements to photometric redshift probability distributions and on the clustering strength of different morphological types of galaxies as a function of redshift. Photometric redshifts are extremely important for the future of astronomy and improvements to them are of extreme value. In the following chapter we present a statistical method we can apply to photo-z PDFs to improve their overall performance, incorporating both single-valued ”point” estimates of redshift, as well as statistics describing distributions, leading to more representative estimates of errors. This method makes use of the Q-Q plot (Wilk & Gnanadesikan (1968)), and provides information regarding biases, unrepresentative uncertainties, asymmetry problems, and faulty characterization of tails of the photo-z PDFs.

We then continue in the next chapter by applying this method to data from the CANDELS collaboration (Grogin et al. (2011); Koekemoer et al. (2011)). The collaboration has produced six separate estimates of photo-z PDFs; we apply a calibration method based on the work described in chapter 2 to each of the six sets of results independently. We then test their performance using independent spectroscopic redshifts, and we observe a clear improvement in all cases. Furthermore, we present 3 different methods of combining individual PDF estimates for a given object, which in turn lead to even better results when tested with these independent redshifts.

In the next chapter we calculate correlation functions, making use of the CANDELS photometric redshift and morphology catalogs. The redshift catalogs are constructed by using spec-z's whenever available, 3D-HST grism-z's (Momcheva et al. (2016)) if there is no spec-z information, and the improved photo-z's developed in Chapter 3 if neither spec-z's nor grism-z's are available. The morphology catalogs are provided to the collaboration by (Kartaltepe et al. (2015)). We select galaxies belonging to 3 major morphological types: spheroids, which generally represent elliptical galaxies, disks that represent lenticular and spiral galaxies, and finally irregular/peculiar galaxies. We then evaluate cross-correlation functions of each morphology type sample with the full galaxy sample within different redshift bins, in order to investigate how the morphology-density relation evolves with redshift.

The final chapter includes a detailed discussion about our results and findings, and the implications for theories of galaxy formation and evolution. We also present ideas of future work related to this study that can help to unlock the mysteries of the Universe.

2 IMPROVING PHOTOMETRIC REDSHIFT PROBABILITY DISTRIBUTIONS WITH THE QUANTILE-QUANTILE PLOT

2.1 INTRODUCTION

The natures of dark matter and dark energy are still unknown to us today. New and upcoming surveys designed to study these phenomena will characterize very large numbers of objects; for instance, the Large Synoptic Survey Telescope (LSST; Ivezic et al. (2009)) plans to observe billions of galaxies over almost half the sky. We can use the redshift of an object as a proxy for its distance or lookback time; we determine redshifts by evaluating the light we receive from a galaxy.

One way to estimate the redshift of a galaxy is by analyzing a detailed spectrum, which in general yields very accurate results. Unfortunately, the acquisition of galaxy spectra of galaxies is not an easy task, since for faint objects such as those studied by LSST very long exposure times are required to obtain an adequate signal-to-noise ratio, while the number of objects that are targeted during a given observation is limited by technical challenges. As a result it will be practically impossible to obtain spectroscopic redshifts (also referred to as spec-z's) for the great majority of objects studied by LSST and other deep imaging surveys.

Alternatively, distant galaxies can be characterized using broad-band photometry, which provides measurements of the fluxes through a particular filter for all objects in the field of view of an instrument. With this method, many objects can be observed at the same time, and good signal-to-noise is achievable even for very faint objects, making it the only feasible option for studying the objects from large surveys such as the aforementioned LSST. Useful flux measurements can be obtained with much shorter exposure times than what is required for more detailed spectra, even with the need to observe multiple times with different filters.

The downside to this is the much smaller amount of information provided by broad-band imaging. As a result, inferences about the redshifts of galaxies from their photometry – commonly known as photometric redshifts or photo-z's – are considerably less precise than their spectroscopic counterparts.

It is common for photometric redshift algorithms to estimate probability density functions (PDFs) for the photo-z of an object, since PDFs provide considerably more information than a single "point" estimate of redshift or than a point estimate plus assumed-Gaussian errors. PDFs and point estimates are closely related; it is common to estimate point values directly from photo-z PDFs. For instance, one could use the redshift at which the PDF has its greatest value (z_{peak}) or the first moment of the PDF (i.e., the PDF-weighted mean of redshift, z_{weight}; in some cases this may be calculated only using the highest peak of the probability distribution) as a single estimate of the photometric redshift. When either of these point values are compared to independent spectroscopic redshift samples, one generally finds substantial scatter, as well as a significant fraction of "outlier" redshifts which are far from the estimated photo-z. It is not uncommon for photo-z estimates to be biased on average when compared to spectroscopic z's.

Just as point estimates can have imperfections, the PDFs provided by existing photometric redshift codes have proven in the past to be unreliable, not meeting the statistical definition of a probability density function (Fernández-Soto et al. (2002), Hildebrandt et al. (2008), Dahlen et al. (2013), etc.). One way this may be seen is by investigating uncertainty estimates derived from PDF measurements. If credible intervals (the Bayesian equivalent of confidence intervals) are properly constructed, then 68% of spectroscopic redshifts should lie within the 68% credible intervals of the corresponding photo-z PDFs, 95% of spectroscopic redshifts should lie within the 95% credible intervals, etc. The reality can be far from this scenario; in recent tests, the number of spectroscopic redshifts within a given credible region may be much higher or much lower than what would be expected if PDFs have been properly constructed, depending on the particular photo-z code and the size of the credible interval considered (Dahlen et al. (2013)).

In this paper, we present simple methods which can help to correct for low-order deficiencies in the probability density functions output by photometric redshift codes, resulting in better estimates of both point values and credible intervals. This is done using the Quantile-Quantile plot (also known as the Q-Q plot; Wilk & Gnanadesikan (1968)) constructed with a subsample of objects with spectroscopic redshift measurements as well as the corresponding photo-z PDFs of the same objects.

We take advantage of the fact that, if the photo-z PDFs fulfill the standard statistical definition of a probability distribution, then the values of the cumulative distribution functions (CDFs) constructed from the PDFs and evaluated at the actual spec-z's of the objects should follow a uniform distribution from zero to one. When this is not the case, it is an indicator that the photo-z PDFs are imperfect. In this paper, we consider the impacts on the Q-Q plot of a bias in the PDFs (shifting them from the true PDF); an underestimate or overestimate of errors leading to PDFs that are too narrow or too broad; an inaccurate level of asymmetry in the PDFs (i.e., inaccurate skewness); or finally, cases where the tails of the PDFs are too large or too small (i.e., inaccurate kurtosis).

In a recent paper published while this paper was being written, Freeman et al. (2017) used Q-Q plots to investigate the impact of differences between the distribution of properties of objects with spectroscopic redshifts used to train photo-z algorithms and the properties of the objects to which the algorithms are applied, as well as to evaluate their methods for mitigating this effect. In this paper, we consider the utility of the Q-Q plot as a general tool for assessing photo-z PDF accuracy, as well as methods for optimizing PDFs using statistics based on these plots.

In Section 2.2 we explain in detail the methods we use to assess the quality of photo-z PDFs, and present Q-Q results for a variety of simple cases to illustrate the information available. In Section 2.3 we describe how information from Q-Q statistics can be used to calibrate photo-z PDFs to more closely fulfill the statistical definition, and illustrate our methods with mock data. Finally, in Section 5.2 we summarize and discuss the findings of this study.

2.2 METHODS

In order to quantify the quality of photo-z PDFs, we make use of a commonly-used variant of a probability plot, commonly known as the quantile-quantile plot or Q-Q plot (Wilk & Gnanadesikan (1968)). The Q-Q plot is frequently used as a visual technique to assess whether a set of data follows a given distribution, and is constructed by plotting quantiles – that is, the value within a distribution that some particular fraction of all values are below – calculated from data (shown as the y axis) against a set of quantiles expected from theory (used as the x axis). Percentiles or quartiles are examples of quantiles, but in the Q-Q plot the fractions at which quantiles are defined may be distributed continuously.

In an ideal case where the data are drawn from the theoretical distribution and the number of datapoints is large, the data quantiles should equal exactly the theoretical quantiles. In this scenario, since both the x values and y values are exactly the same, the plot should correspond to the unit line (i.e., the line from $(x = 0, y = 0)$ to $(x = 1, y = 1)$); this can be used as a reference line that the Q-Q curve for a particular set of distributions can be compared to. The deviation from the unit line provides useful information about how the data differs from the expected distribution, which can then be used to recalibrate and improve the quality of PDFs, as we show below.

More specifically, the data quantiles we use to construct the Q-Q plot are the values of the photo-z cumulative distribution functions (CDFs) evaluated at the actual spectroscopic redshifts of objects of known z, where the CDFs are derived from the corresponding photometric redshift PDFs (photo-z PDFs), compared to the expected quantiles for a uniform distribution (as this is the distribution expected for data values randomly drawn from the PDF corresponding to a given PDF). The Q-Q plot constructed from the photo-z CDFs evaluated at the spectroscopic redshifts of sample objects provides a test of whether the photometric redshift probability distributions meet the statistical definition of a PDF.

If the photo-z PDFs have erroneous moments (such as their mean, standard deviation, skewness, kurtosis, etc.) this causes signatures in the Q-Q plot. Photometric redshift probability distributions can be off in a variety of ways. For instance, an error in the mean corresponds to the PDFs having a bias such that they are effectively shifted to the left or to the right compared to the true PDF of redshifts. An error in the standard deviation of the PDFs implies that they are too narrow or too broad; this will have effects primarily on the size of credible intervals predicted from a PDF. A difference in the skewness results in the PDFs having a different symmetry (or asymmetry) when compared to the true distributions. Finally, a difference in kurtosis results in the PDFs having too small or too long tails, and therefore predicting more or fewer outliers than they should. All these four cases show unique features in the Q-Q plot, and we investigate them below, following a detailed description of how we construct *Q*-*Q* plots.

2.2.1 Construction of Quantile-Quantile Plots

In order to construct a Q-Q plot, we start by generating mock data for a sample of size of $N_s = 1000$ objects, and for each object we generate a spec-z value from its true photo-z PDF. For simplicity, in this paper we generally assume the true PDF is the same for all objects in a sample and is represented by a perfect Gaussian distribution of mean $\mu_{\rm phot}$ and standard deviation $\sigma_{\rm phot}$. We can make this assumption without loss of generality, as the Q-Q statistics are based upon the distribution of CDF values constructed from a particular object's PDF and evaluated at its spectroscopic redshift; the details of that PDF cease to matter as soon as the CDF is evaluated (we illustrate this with a detailed test below). An example of a simple toy model photo-z PDF and an associated spectroscopic redshift ($z_{\rm spec}$) is shown in Figure 2.1.

Using an assumed photo-z PDF (which may or may not correspond to the true PDF from which redshifts were generated) we construct the cumulative distribution function (CDF(z)), from the equation:

$$\mathrm{CDF}\,(z) = \int_0^z \mathrm{PDF}\,(z')\,\mathrm{d}z' \tag{2.1}$$

Additionally, we calculate the data quantile for a given object, $Q_{\rm data}$, using the spec-z value and the CDF:

$$Q_{\rm data} = \mathrm{CDF}\,(z_{\rm spec}) = \int_0^{z_{\rm spec}} \mathrm{PDF}\,(z')\,\mathrm{d}z' \tag{2.2}$$

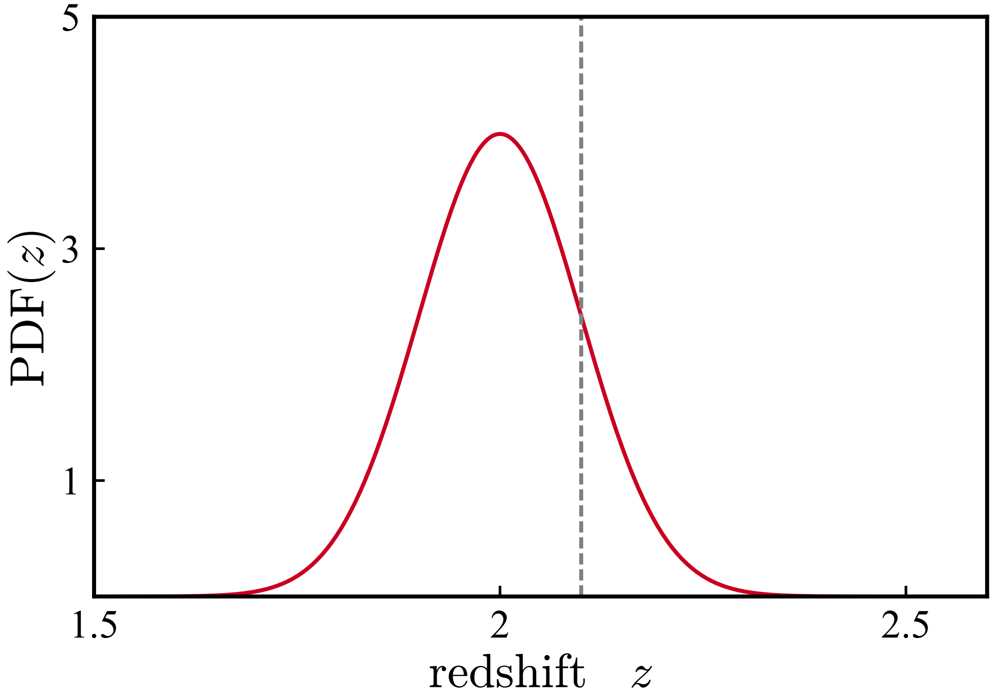


Figure 2.1 True photometric redshift Probability Density Function, PDF(z) of a hypothetical object. The red curve corresponds to a normal distribution with $\mu_{\text{phot}} = 2$ and $\sigma_{\text{phot}} = 0.1$, which we assume to be the photometric redshift PDF for a particular object. The vertical grey dashed line represents the spectroscopic redshift of the same object, which is here assumed to be $z_{\text{spec}} = 2.1$.

as is illustrated in Figure 2.2.

We continue by calculating the quantiles of the CDF values for all the objects in the mock data sample, sorting the values in increasing order. Theoretically we expect the values of the quantiles to follow a uniform distribution from 0 to 1, since we expect 1% of the objects to have values of the CDF less than or equal to 0.01, 2% of them to have CDF values less than or equal to 0.02, etc. We then construct the Q-Q plot by plotting the calculated quantiles from our mock data versus the theoretical quantiles (which just correspond to the fraction of CDF values that are lower than the given value). For instance, if the CDF values are below 0.2 10% of the time, the (Q_{theory}, Q_{data}) pair (0.1,0.2) would be a point along the Q-Q curve, as 0.2 would be the CDF value that corresponds to the 0.1 quantile (or tenth percentile) in the distribution of CDFs.

In the ideal case, the data quantiles should exactly equal the theoretical values, and therefore the plot should correspond to the unit line. In practice, that will never be the case even if PDFs are perfectly known, since the data are randomly sampled, causing deviations of quantiles from the theoretical expectation.

In order to quantify the deviation of the Q-Q plot from the ideal diagonal line, we use the normalized ℓ^2−norm, which is the square root of the sum of the squares of the differences of the plotted Q-Q curve from the diagonal line at each Q_{theory} value (i.e., the Euclidean distance), normalized by dividing by the square root of the number of objects:

$$\text{normalized } \ell^2\text{−norm} = \sqrt{\frac{\sum_{i=1}^{N_s} |y_{i,\text{data}} - y_{i,\text{theory}}|^2}{N_s}}, \tag{2.3}$$

where $y_{i,\text{data}}$ is the y value in a plot for the true curve and $y_{i,\text{theory}}$ is the theoretical expectation (corresponding to the unit line, in our case). The differences between the curve and the reference line at each point are calculated solely in the vertical direction.

This is not the only possible choice of metric, of course. One alternative is the normalized ℓ^1−norm, which is the sum of the absolute values of the differences between the two curves

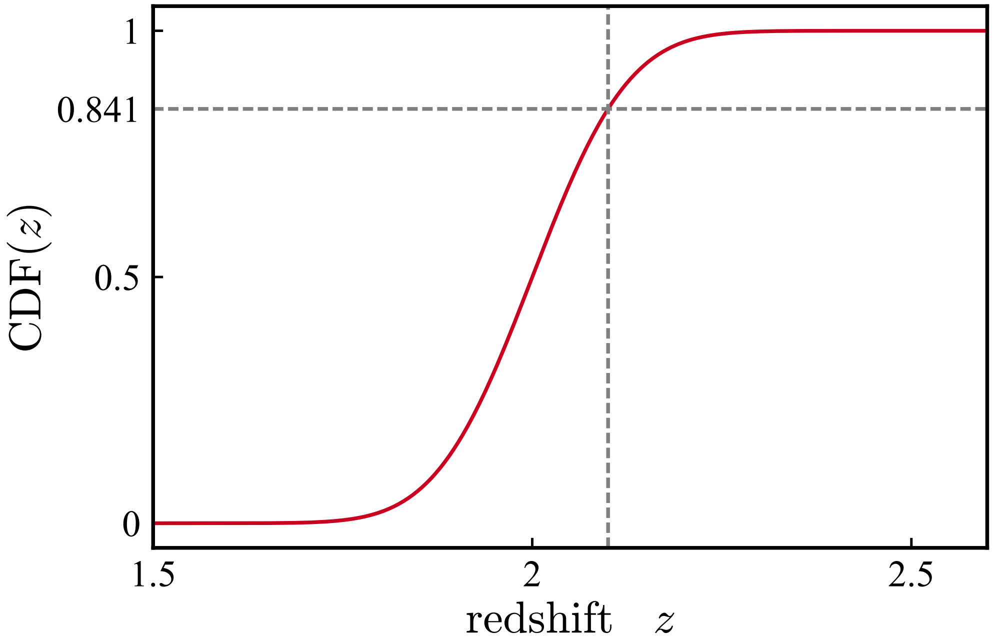


Figure 2.2 Cumulative Distribution Function, $\mathrm{CDF}\,(z)$, for the hypothetical object from Figure 2.1. The red curve indicates the cumulative distribution function corresponding to the PDF in Figure 2.1. The spectroscopic redshift of $z_{\mathrm{spec}} = 2.1$ is indicated by the vertical grey dashed line, while the horizontal grey line shows the quantile value for the given object, $\mathrm{CDF}\,(z_{\mathrm{spec}}) = 0.841$.

at each point, normalized by dividing by the square root of the number of objects:

$$\text{normalized } \ell^1\text{−norm} = \frac{\sum_{i=1}^{N_s} |y_{i,\text{data}} - y_{i,\text{theory}}|}{N_s^{1/2}}. \tag{2.4}$$

Another metric that can be used is the largest difference between the calculated curve and the reference line, which is analogous to the D-value used in the Kolmogorov-Smirnov test (or K-S test for short). The difference here is that the K-S D statistic is generally computed between the empirical CDF and the theoretical CDF, whereas we are comparing quantile-quantile curves to the unit line.

The ℓ^1−norm and ℓ^2−norm are in general more sensitive than the K-S-like D-value, since they take into account information from the entirety of the curve, not only the largest difference. Furthermore, the ℓ^2−norm is more sensitive than the ℓ^1−norm , since larger differences are squared and affect it more, for much the same reasons that the mean of a distribution of points is affected more by outliers than the median.

An additional useful quantity we measure is the fraction of the objects of the sample for which the quantiles are very small ($\text{quantile} \leq 0.0001$), or very large ($\text{quantile} \geq 0.9999$). These values correspond to spec-z's that fall outside of the main regions of their respective photo-z PDFs, such that they might be considered outliers. We label this fraction of objects as f_{op} (the fraction-outside-pdf), and we exclude them from the construction of the Q-Q plot, tracking their numbers separately.

We start with the simplest case possible, where we consider all 1000 objects of our sample to have the same photo-z PDF, such that the spec-z's are all generated from the same Gaussian distribution (Figure 2.3). We then calculate the quantiles for each redshift and after sorting them in increasing order, we plot them against a uniform distribution of values from 0 to 1 (Figure 2.4). The randomness in the generated spec-z's results in the plotted curve deviating slightly from the ideal diagonal line, and the normalized ℓ^2−norm, though very small, is not identically zero. As expected, there are no outliers in this case, so $f_{\text{op}} = 0$.

We should note that while in this paper we generally employ toy models which assume that all objects have the same assumed photo-z PDF and that all spectroscopic redshifts

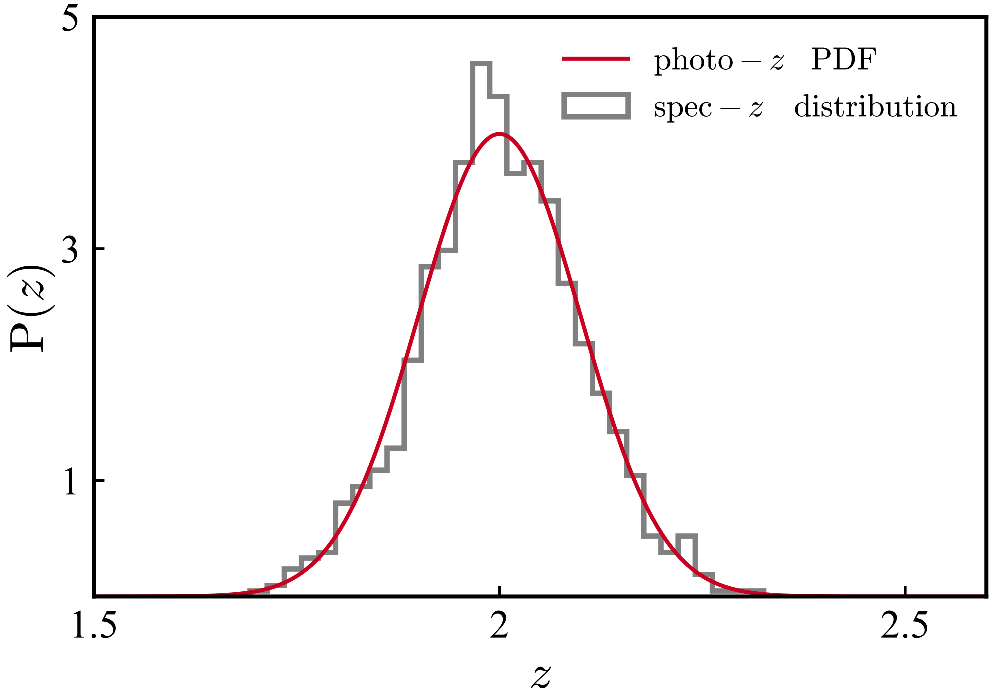


Figure 2.3 Example of the generation of a set of spectroscopic redshifts from photometric redshift PDFs. In this case, the spectroscopic redshifts (grey normalized histogram) are generated from the same distribution as the photo-z PDF (red solid curve), which is a Gaussian with $\mu_{\rm spec} = \mu_{\rm phot} = 2$ and $\sigma_{\rm spec} = \sigma_{\rm phot} = 0.1$. For simplicity, the photo-z PDF is chosen to be the same for all objects of the sample (in this case, $N_s = 1000$). The histogram shows that although there is some randomness in the generation of the spec-z's, such that the histogram has small deviations from the red line, the two distributions generally accord with each other.

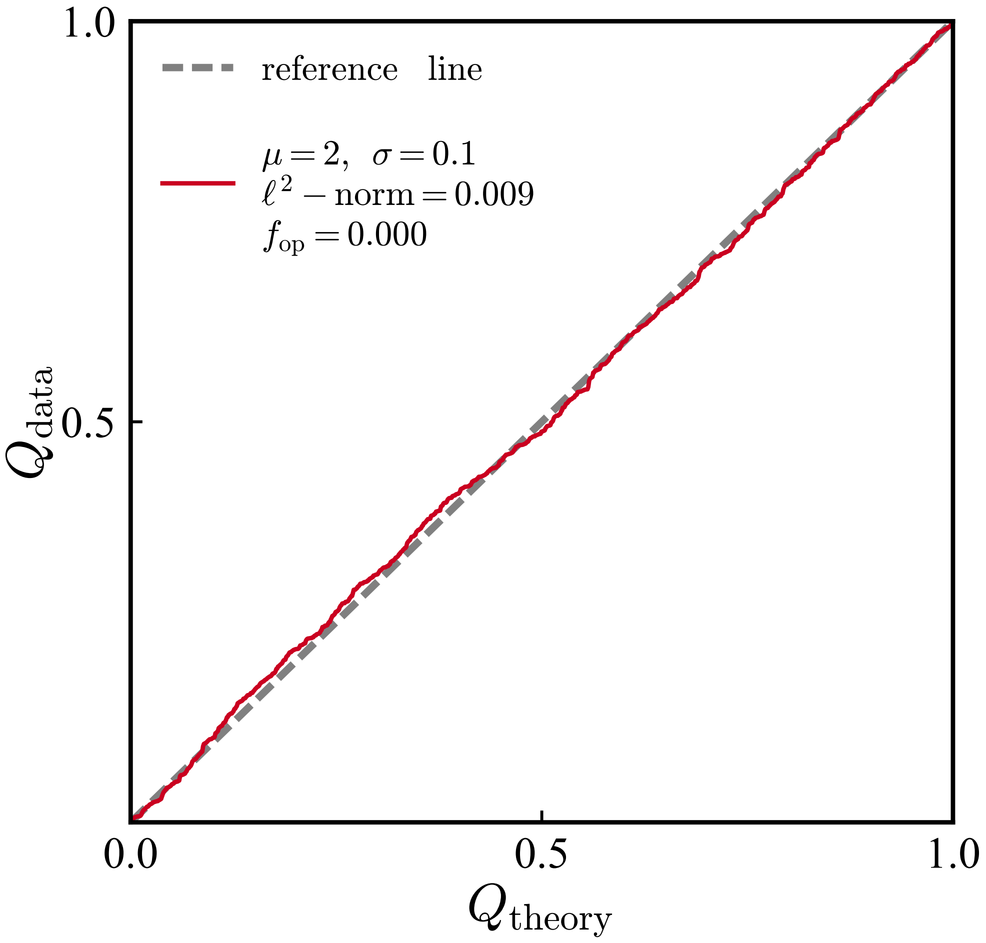


Figure 2.4 The Q-Q plot for the same sample of $N_s = 1000$ objects illustrated in Figure 2.3. The diagonal gray dashed line represents the ideal case, where the distribution of quantiles derived from the assumed photo-z PDF evaluated at the spec-z's is perfectly uniform between zero and one. The red curve is constructed using CDF values from the same distribution actually used to generate the spec-z's, as illustrated in Figure 2.3. As expected, the ℓ^2−norm is very close to zero and the outlier fraction f_{op} is exactly zero for this case.

are truly generated from a single Gaussian distribution, that does not have to be the case. Each object can have its own photo-z PDF – Gaussian or otherwise – with a spec-z may be generated from a correspondingly varying distribution, but the Q-Q results will be the same. To illustrate this, we construct separate photo-z PDFs for each object of a sample with $N_s = 1000$, taking each to be a Gaussian distributions with mean μ_i and standard deviation σ_i randomly chosen from a uniform distribution of values in the ranges $2 \leq \mu_{\rm phot} \leq 4$ and $0.05 \leq \sigma_{\rm phot} \leq 0.5$. We then generate a single spec-z per object from each distribution (with parameters μ_i and σ_i) and calculate its CDF values using a Gaussian of the correct parameters for that object's PDF. We proceed again with calculating the quantiles and construct the Q-Q plot. The results are shown in Figure 2.5; the Q-Q result is indistinguishable from that where all objects have identical PDFs, differing only due to random sampling. This demonstrates that we can explore the behavior of Q-Q statistics using the same photo-z PDF for every object in a sample or using varying PDFs and still obtain equivalent results. In the remainder of this paper we will employ the first of these strategies for simplicity.

Although this case is more realistic than having a single PDF for all objects, the more complicated scenario has no effect on Q-Q statistics. Therefore, in the remainder of this paper we will continue to use a more simplistic model where the assumed photo-z PDF for all objects is identical, and where the spec-z's are generated from a single Gaussian distribution (which may or may not match the assumed PDF).

In the following subsections, we examine the cases where the two distributions (the estimated distribution represented by the photo-z PDF versus the true probability distribution from which the spec-z's are drawn) differ in their mean ($\mu_{\rm spec} \neq \mu_{\rm phot}$); their standard deviation ($\sigma_{\rm spec} \neq \sigma_{\rm phot}$); their skewness ($\gamma_{1,\rm spec} \neq \gamma_{1,\rm phot}$); or their kurtosis ($\beta_{2,\rm spec} \neq \beta_{2,\rm phot}$). In the first two cases, we assume Gaussian distributions for both the assumed photo-z PDF and the spec-z distribution. In the latter two cases we still generate the spec-z's from Gaussian distributions, but the photo-z PDFs are constructed using a mixture model consisting of more than one Gaussian. We will investigate how errors in each of these moments of the PDF affects the shape of the Q-Q plot.

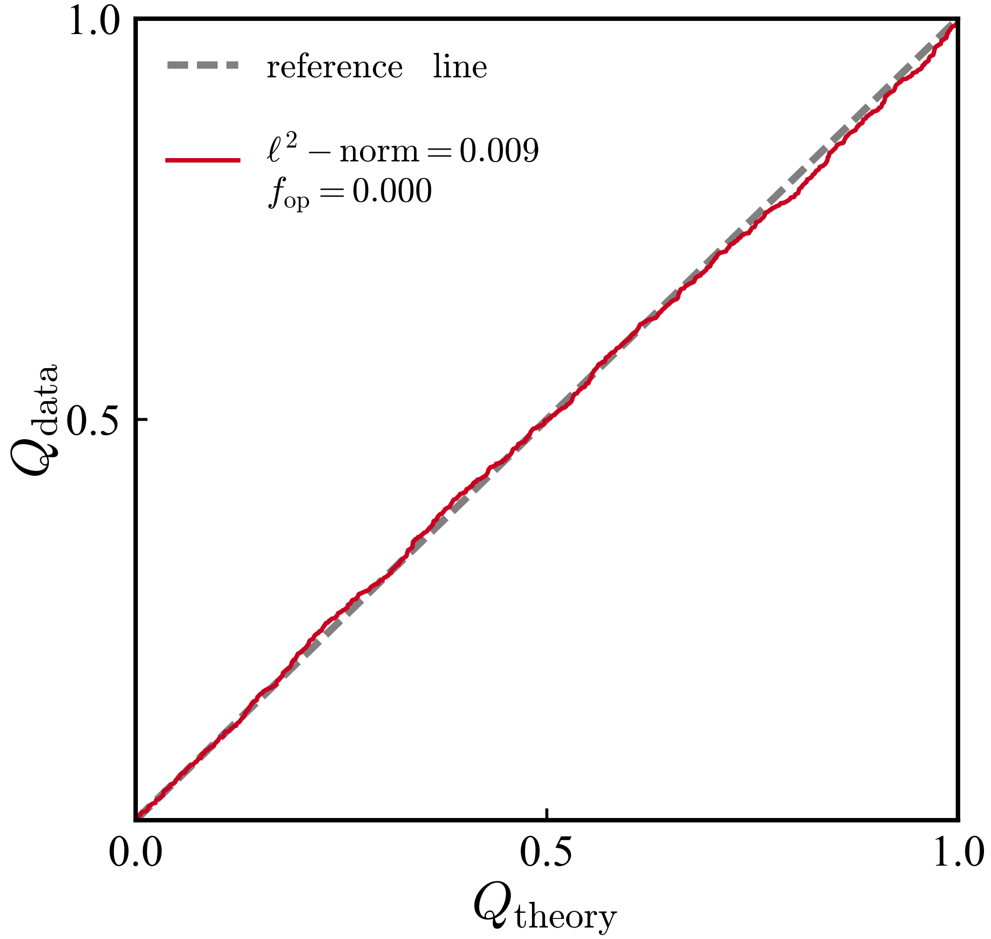


Figure 2.5 Q-Q plot for a sample of 1000 objects, constructed using a different photo-z PDF for each object. Again, the diagonal grey dashed line represents the expectation for the ideal case, while the red solid line is constructed using the mock data. As before, in this case the normalized ℓ^2−norm is close to zero and the outlier fraction $f_{\rm op}$ is exactly zero, even though photo-z PDFs were variable rather than identical between objects. Because the Q-Q curves rely on CDF values, rather than the details of a PDF, we can explore them using simple PDF scenarios and still capture their behavior properly.

2.2.2 Impact of Inaccuracies in the PDF Mean

The first case we consider is the possibility that the first moment – i.e., the mean redshift – of the photo-z PDF is incorrect; this is sometimes referred to as a bias in a set of photo-z estimates. In this scenario we use a mock assumed photo-z PDF which is a normal distribution with the same standard deviation as the spec-z distribution, $\sigma_{\mathrm{spec}} = \sigma_{\mathrm{phot}}$, but having a different mean, $\mu_{\mathrm{spec}} \neq \mu_{\mathrm{phot}}$. Effectively, in this case the photo-z PDFs is shifted relative to the true (spec-z) distribution; we here investigate the signature that such a feature leaves in the Q-Q plot.

We start by generating a set of one thousand spectroscopic redshifts from a Gaussian distribution with $\mu_{\mathrm{spec}} = 2$ and $\sigma_{\mathrm{spec}} = 0.1$, while we consider five different cases of putative photo-z PDFs, with $\sigma_{\mathrm{phot}} = \sigma_{\mathrm{spec}} = 0.1$ and $\mu_{\mathrm{phot}} = 1.9$, 1.95, 2, 2.05, or 2.1 (Figure 2.6). This includes the case where the photo-z PDF is the same as the distribution used to generate the spec-z's ($\mu_{\mathrm{phot}} = \mu_{\mathrm{spec}} = 2$ and $\sigma_{\mathrm{phot}} = \sigma_{\mathrm{spec}} = 0.1$), shown as the red solid curve. We proceed with the construction of the Q-Q curves (shown in Figure 2.7) and the evaluation of the normalized $\ell^2\mathrm{-norm}$ and the f_{op} for each scenario, using the same analysis methods as in Subsection 2.2.1.

As was seen in the previous subsection, when the assumed probability distribution for redshifts matches the actual distribution, the Q-Q curve is very close to the reference line, with small deviations from it only due to random sampling. In contrast, when the assumed photo-z PDFs are biased compared to the true distributions (so $\mu_{\mathrm{phot}} \neq \mu_{\mathrm{spec}}$), the curves start to deviate more and more from the diagonal as the difference in the means of the two distributions becomes larger. More specifically, when $\mu_{\mathrm{phot}} = 1.9$ or 1.95 the curves are above the reference line, with larger deviations from diagonal when the mean shift is greater. On the other hand, for $\mu_{\mathrm{phot}} = 2.05$ or 2.1 the curves are below the reference line, with, again, larger deviations for greater inaccuracies in the mean redshift of the PDF. This is also reflected in the value of the normalized $\ell^2\mathrm{-norm}$ for each case, as shown in Table 2.1. As seen in the table the f_{op} values in all cases are very low, less than 1%. This is because not many spec-z's fall extremely far into the tails of the photo-z PDFs for the relatively modest biases in the mean used here. If we assume a much larger difference in the mean of the

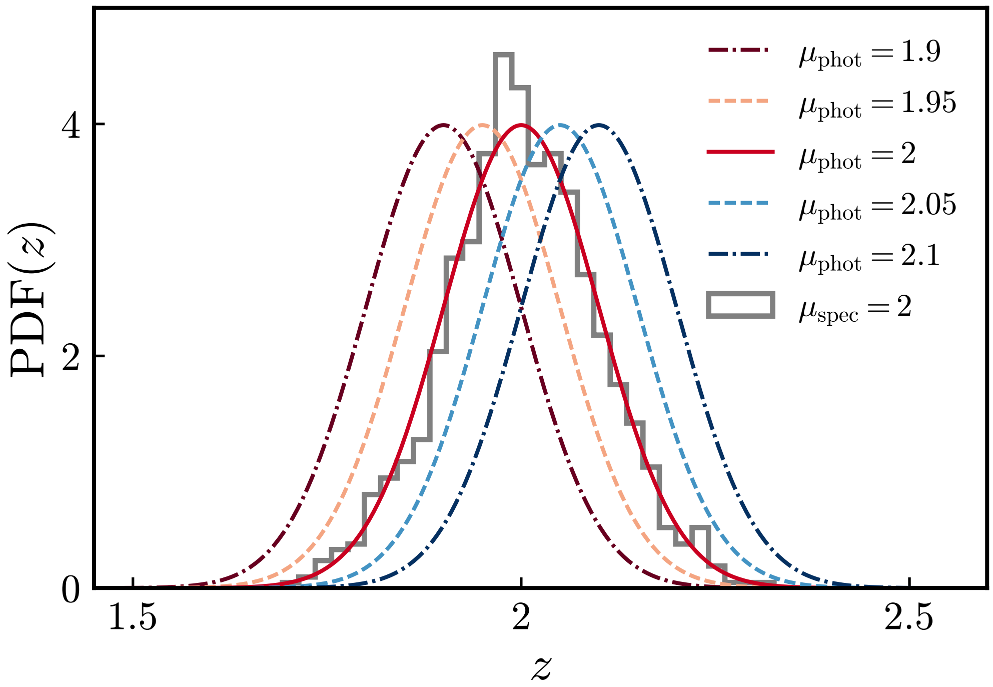


Figure 2.6 Five different assumed photo-z PDFs used to assess the impact on the Q-Q plot of biases in the mean redshift of the PDF. The solid curves correspond to Gaussian distributions with means μ_{phot} = 1.9, 1.95, 2, 2.05, 2.1 and standard deviation σ_{phot} = 0.1. Also shown in black is the normalized histogram of mock spectroscopic (true) redshifts, which are generated from a Gaussian distribution with μ_{spec} = 2 and σ_{spec} = 0.1; this distribution, in combination with the cumulative distribution functions defined by each PDF in this plot, is used to generate the QQ curves in Figure 2.7.

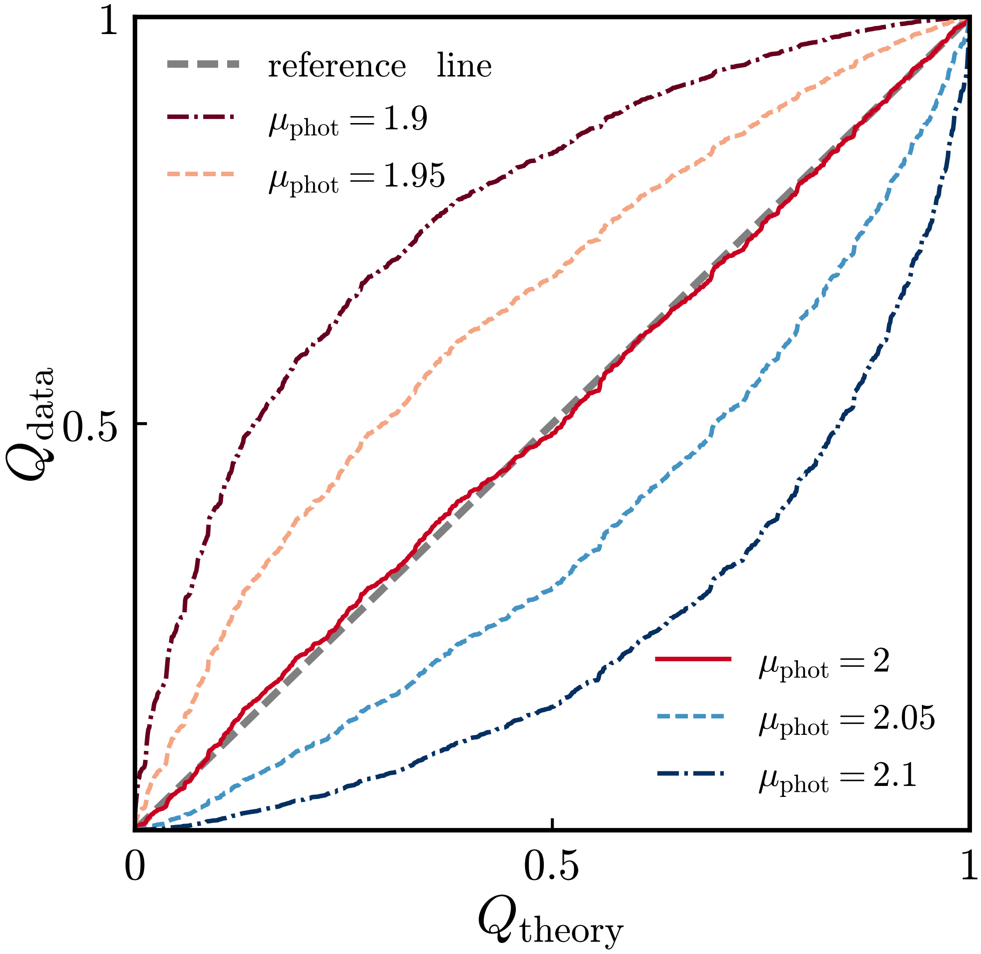


Figure 2.7 Q-Q plot for the photo-z PDFs shown in Figure 2.6. The case where the assumed PDF has the same mean as the spec-z distribution (so $\mu_{\rm phot} = \mu_{\rm spec} = 2$) is shown by the red solid curve, which differs from the ideal case (indicated by the dashed line) only due to random sampling. For the other four curves, the larger the difference in μ, the larger the deviation of the plotted curve from the reference line; the sign of a photo-z PDF's bias compared to the true distribution can be determined from whether the curve lies above or below the diagonal.

Table 2.1 Table with values of ℓ^2−norms and $f_{\rm op}$'s for photo-z PDFs which are Gaussians with varying means ($\mu_{\rm phot}$), compared to a true, spectroscopic redshift distribution which is Gaussian with mean $\mu_{\rm spec} = 2$. Both the ℓ^2−norm and the $f_{\rm op}$ are lowest when the assumed photo-z PDF matches the distribution from which the spectroscopic redshifts are drawn, and grow larger the greater the difference between the two distributions.

$\mu_{\rm phot}$	normalized $\ell^2 -$ norm	$f_{\rm op}$
1.9	0.288	0.003
1.95	0.153	0.001
2	0.009	0
2.05	0.150	0
2.1	0.284	0.003

two distributions (such as $\mu_{\rm phot} - \mu_{\rm spec} = 2.5 - 2 = 0.5$), the $f_{\rm op}$ values increase by an order of magnitude or more.

2.2.3 Impact of Inaccuracies in the PDF Standard Deviation

We next investigate the signatures of photo-z PDFs which have a different standard deviation from the true Gaussian distribution of the spec-z's. This could result from estimated probability distributions which are overconfident, such that they have a smaller standard deviation than the true spec-z distribution (i.e., $\sigma_{\rm phot} < \sigma_{\rm spec}$), or when they are underconfident, with $\sigma_{\rm phot} > \sigma_{\rm spec}$.

We again generate sets of one thousand spectroscopic redshifts from a normal distribution with $\mu_{\rm spec} = 2$ and $\sigma_{\rm spec} = 0.1$, which we compare to five different photo-z PDFs, all of which have the same mean ($\mu_{\rm phot} = \mu_{\rm spec} = 2$) but with varying standard deviations $\sigma_{\rm phot} = 0.02, 0.05, 0.1, 0.15$, or 0.2. The PDFs used are depicted in Figure 2.8. For each photo-z PDF considered, we construct a Q-Q curve; these are shown in Figure 2.9. The case where the assumed photo-z PDF matches the spec-z distribution, so $\mu_{\rm phot} = \mu_{\rm spec} = 2$ and $\sigma_{\rm phot} = \sigma_{\rm spec} = 0.1$, is again shown as a red solid curve.

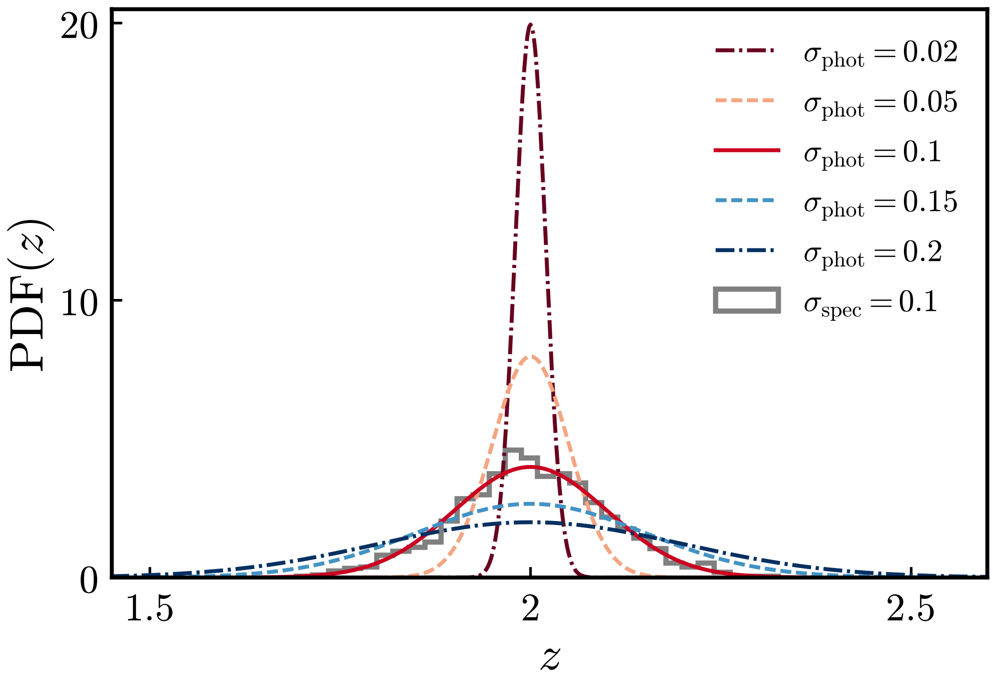


Figure 2.8 The five putative photo-z PDFs used to investigate the impact of inaccuracies in the PDF standard deviation. The curves are all Gaussians with $\mu_{\rm phot} = 2$ and $\sigma_{\rm phot} = 0.02, 0.05, 0.1, 0.15$, or 0.2. Also shown is the normalized histogram of spec-z's generated from a Gaussian distribution of mean $\mu_{\rm spec} = 2$ and standard deviation $\sigma_{\rm spec} = 0.1$ which is used to determine the Q-Q curves for each PDF.

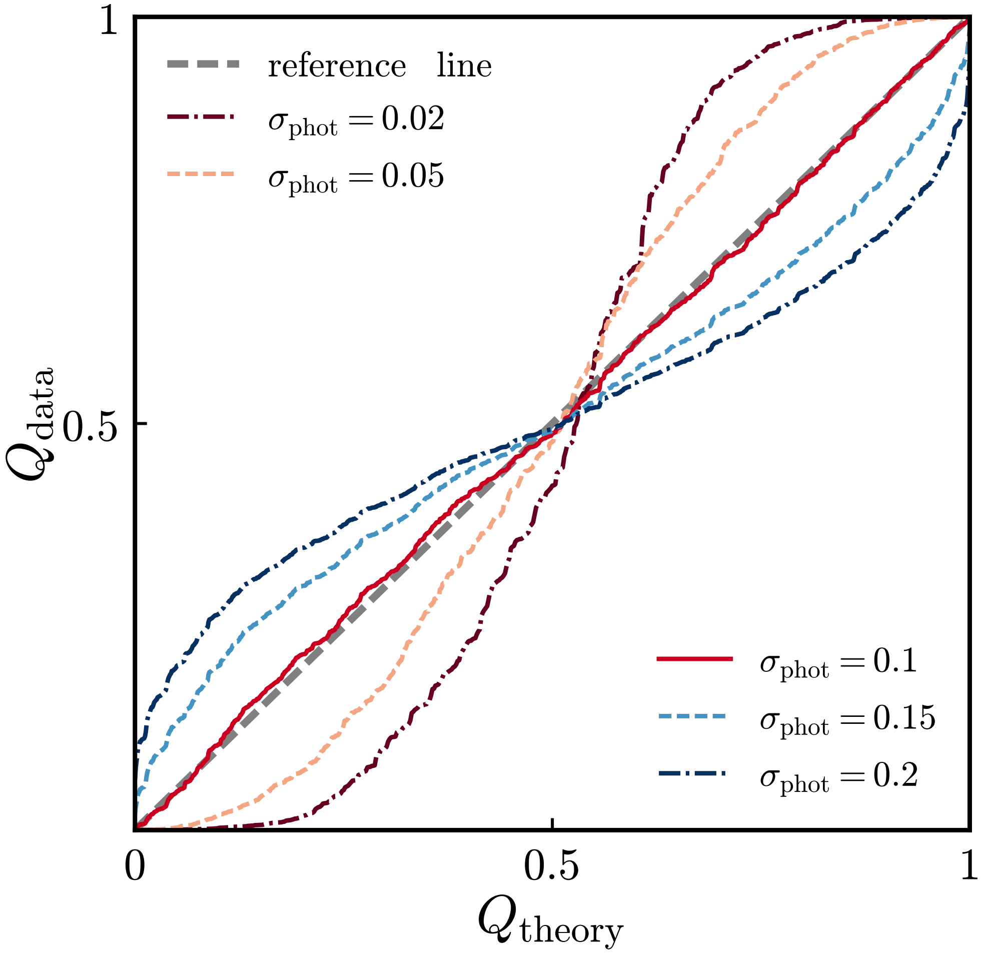


Figure 2.9 Q-Q plot for the photo-z PDFs shown in Figure 2.8, which share the same mean as the spec-z distribution ($\mu_{\mathrm{phot}} = \mu_{\mathrm{spec}} = 2$), but have varying standard deviations σ_{phot}. The case where the PDF has the same standard deviation as the spec-z distribution is shown by the red solid curve. For the other four cases, the larger the difference in σ is, the larger the deviation of the plotted curve from the reference line. Whether the σ is too high or too low can be inferred from the pattern of the Q-Q curve.

Table 2.2 Table with values of ℓ^2−norm and f_{op}, for the different cases of σ_{phot}. Both the ℓ^2−norm and f_{op} are lowest when $\sigma_{\text{phot}} = \sigma_{\text{spec}} = 0.1$.

σ_{phot}	normalized ℓ^2 − norm	f_{op}
0.02	0.147	0.435
0.05	0.093	0.071
0.1	0.009	0
0.15	0.074	0
0.2	0.117	0

We see clearly from the Q-Q plot that when the standard deviations of the PDFs differ from that of the spec-z distribution by a larger factor, the curves deviate more from the reference line. We note that when $\sigma_{\text{phot}} < \sigma_{\text{spec}}$ the curves start below the reference line, then cross it near the center of the plot and remain above the diagonal until $Q_{\text{theory}} = 1$, while the opposite happens when the standard deviation of the photo-z PDF is larger than the one from the spec-z distribution. Since the distance of the Q-Q curves from the reference line is larger when the σ is misestimated by a larger amount, the values of the normalized ℓ^2−norm also grow accordingly, as shown in Table 2.1. The trend of the fraction of spec-z's that lie outside the photo-z PDF is somewhat different, however. Where the PDF is underconfident (so σ is too large), f_{op} must be zero, since for photo-z PDFs with tails as wide or wider than the distribution of spec-z's, the values of the CDFs evaluated at the spectroscopic redshift will always be greater than 0.0001 and lower than 0.9999, because the spec-z's cannot not fall beyond the photo-z PDF in that case. On the other hand, when the photo-z PDFs have a standard deviation which is too small, more and more spec-z's will be at redshifts beyond the range where the photo-z PDF is significant, yielding CDF values at those redshifts that are either lower than 0.0001 or larger than 0.9999. This is a strong effect; out of the cases considered, f_{op} becomes as large as 46.2% for $\sigma_{\text{spec}} = 0.1$ and $\sigma_{\text{phot}} = 0.02$, making this statistic useful for testing for overconfident (i.e., overly tight) PDFs.

2.2.4 Impact of Inaccuracies in the PDF Skewness

In this subsection we want to investigate the case where a photo-z PDF differs from the spec-z distribution in its skewness (γ_1), while having an identical mean, standard deviation, and kurtosis (β_2).

We continue to generate the spec-z's from a normal distribution of $\mu_{\text{spec}} = 2$, $\sigma_{\text{spec}} = 0.1$. For a normal distribution the skewness is always $\gamma_1 = 0$, while the kurtosis is always $\beta_2 = 3$. In order to construct photo-z PDFs that have $\mu_{\text{phot}} = 2$, $\sigma_{\text{phot}} = 0.1$, and $\beta_{2,\text{phot}} = 3$, but $\gamma_{1,\text{phot}} \neq 0$, we use a set of Gaussian mixture models (other options tested, such as varying forms of skew Gaussians, were unable to match the kurtosis and standard deviation of the normal distribution while varying the skewness). The selected models are constructed to have PDFs with the desired values of the relevant moments by using a probability distribution which is the weighted sum of two or more Gaussian distributions:

$$\text{PDF}(z) = \sum_i c_i \, p_i(z), \tag{2.5}$$

where c_i are a set of weighting coefficients that satisfy the condition $\sum_i c_i = 1$ and $p_i(z)$ are a set of normal distributions of varying mean μ_i and standard deviation σ_i, $\mathcal{N}(\mu_i, \sigma_i^2)$.

The problem is then one of finding sets of values for the parameters c_i, μ_i, and σ_i which correspond to PDFs which have varying skewness but match the other moments of the normal distribution of spectroscopic redshifts. Via hand-tuning we have developed sets of parameters which yield four different values of non-zero skewness, $\gamma_1 = -1.002$, -0.493, 0.508, and 1.004, while keeping the other moments constant to better than 0.2% (the parameters of the mixture models and their first four moments are specified in Table 2.3). The residual variations in kurtosis between these distributions are insignificant for the purposes of this study and can safely be ignored.

The adopted photo-z PDFs with varying skewnesses are shown in Figure 2.10. We then evaluate the Q-Q curve for each of these PDFs, again including the case where the PDF is the same as the one from which the spec-z distribution is drawn (i.e., $\mu_{\text{phot}} = \mu_{\text{spec}} = 2$, $\sigma_{\text{phot}} = \sigma_{\text{spec}} = 0.1$, $\gamma_{1,\text{phot}} = \gamma_{1,\text{spec}} = 0$, and $\beta_{2,\text{phot}} = \beta_{2,\text{spec}} = 3$, shown as the red solid curve) in Figure 2.11.

Table 2.3 Parameters of Gaussian mixture models used to explore the impact of PDFs with varying skewness (γ_1). See Equation 2.5 for the definition of these models. For a photo-z PDF with $\gamma_1 = -1.002$ and $\gamma_1 = 1.004$, only two Gaussian distributions are needed to obtain the desired moment values (specified in the top row), whereas for $\gamma_1 = -0.493$ and $\gamma_1 = 0.508$, a combination of three Gaussians are needed, with varying weights c_i, means μ_i, and standard deviation parameters σ_i.

moments of final PDF	$\mu = 2$ $\sigma = 0.1$ $\gamma_1 = -1.002$ $\beta_2 = 3.005$	$\mu = 2$ $\sigma = 0.1$ $\gamma_1 = -0.493$ $\beta_2 = 3$	$\mu = 2$ $\sigma = 0.1$ $\gamma_1 = 0$ $\beta_2 = 3$	$\mu = 2$ $\sigma = 0.1$ $\gamma_1 = 0.508$ $\beta_2 = 3.001$	$\mu = 2$ $\sigma = 0.1$ $\gamma_1 = 1.004$ $\beta_2 = 3.002$
c_1	0.242	0.243	1	0.147	0.242
μ_1	1.843	1.980	2	2.019	2.158
σ_1	0.056	0.059	0.1	0.051	0.056
c_2	0.758	0.318	-	0.418	0.758
μ_2	2.050	1.911	-	2.070	1.950
σ_2	0.042	0.092	-	0.096	0.042
c_3	-	0.439	-	0.435	-
μ_3	-	2.075	-	1.926	-
σ_3	-	0.056	-	0.055	-

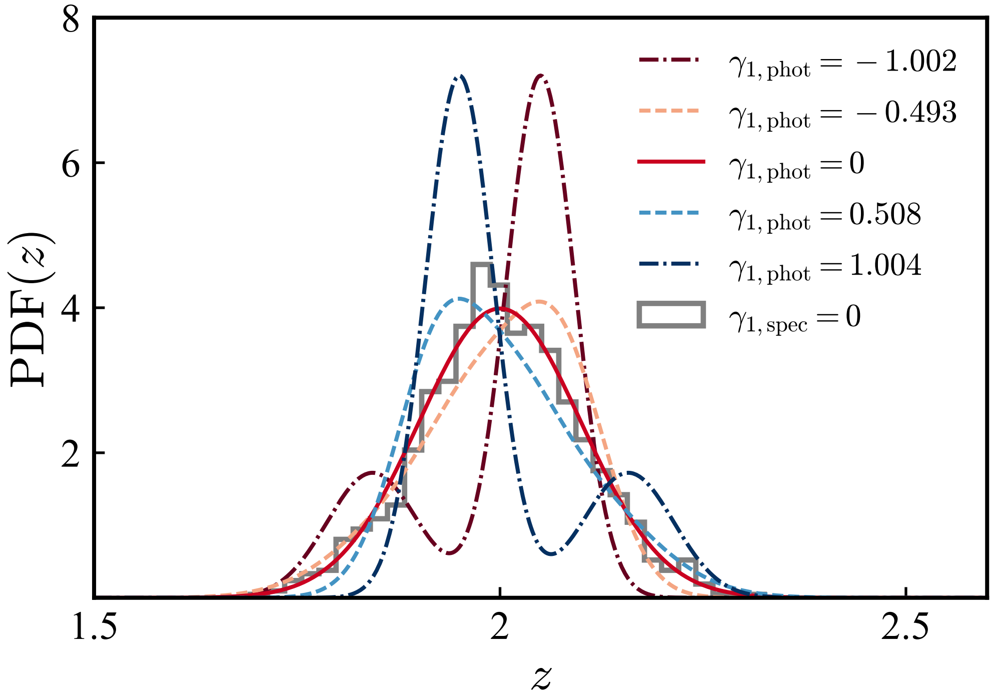


Figure 2.10 Five different cases of photo-z PDFs used to test the impact of varying skewness, with $\gamma_{1,\mathrm{phot}} = -1.002, -0.493, 0, 0.508, 1.004$ but fixed kurtosis $\beta_{2,\mathrm{phot}} = \beta_{2,\mathrm{spec}} = 3$. Also shown is the normalized histogram of spectroscopic redshifts generated from a Gaussian distribution with moments $\mu_{\mathrm{spec}} = 2$, $\sigma_{\mathrm{spec}} = 0.1$, $\gamma_{1,\mathrm{spec}} = 0$, $\beta_{2,\mathrm{spec}} = 3$.

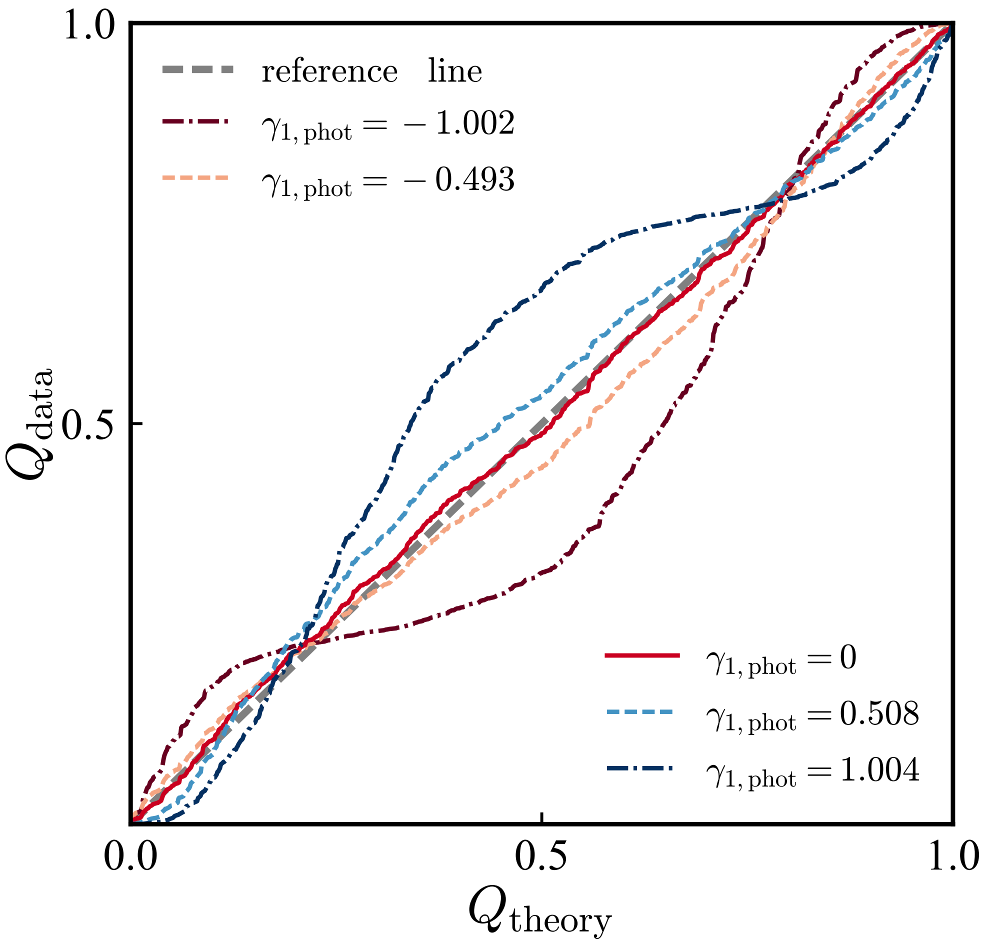


Figure 2.11 Q-Q plot for the five photo-z PDFs depicted in Figure 2.10, which have skewness $\gamma_{1,\mathrm{phot}} = -1.002, -0.493, 0, 0.508, 1.004$ and kurtosis $\beta_{2,\mathrm{phot}} = \beta_{2,\mathrm{spec}} = 3$. The case where the PDF has the same skewness as the spec-z distribution ($\gamma_{1,\mathrm{phot}} = \gamma_{1,\mathrm{spec}} = 0$) is shown by the red solid curve, whereas the reference curve corresponding to a perfect quantile-quantile match is shown as a dashed black line. As differences in γ_1 grow larger, so do deviations of the Q-Q curves from the reference line.

Table 2.4 Table of values of ℓ^2−norm and f_{op}, for the photo-z PDFs of varying skewness depicted in Figure 2.10. Both the ℓ^2−norm and the f_{op} are lowest when the skewness γ_1 is the same for both the true spectroscopic distribution and the assumed photo-z distribution.

γ_1	β_2	normalized ℓ^2 − norm	f_{op}
−1.002	3.005	0.104	0.024
−0.493	3	0.031	0.003
0	3	0.009	0
0.508	3.001	0.033	0.002
1.004	3.002	0.103	0.023

Again, the Q-Q plot shows us that when the two distributions (spec-z distribution and photo-z PDF) are the same (corresponding to the red solid curve), the curve is very close to the dashed reference line, whereas when there is a difference in skewness ($\gamma_{1,\text{phot}} \neq \gamma_{1,\text{spec}}$), the curves deviate increasingly as the difference becomes larger. This can be seen by eye in the Q-Q plot or quantitatively from the values of the normalized ℓ^2−norm in Table 2.4. Additionally, the f_{op} values increase to about 2% for the PDFs where the difference in skewness is largest. This is expected since one of the tails of the PDF becomes smaller when the skewness differs significantly from zero, as can be seen in Figure 2.10. This results in some of the spec-z's falling beyond where the PDF is significant, such that the CDF evaluated at the spectroscopic redshift will be lower than 0.0001 or larger than 0.9999.

It is worth noting that whereas differences in the mean between the assumed photo-z PDF and the true spectroscopic redshift distribution causes no crossings of the diagonal reference line between the endpoints of the Q-Q plot, differences in the standard deviation cause one crossing, and differences in skewness cause two. In this way, the nature of deviations between photo-z PDFs and reality may be read off the Q-Q plot directly. In each case, the sign of those deviations is also reflected directly in whether the Q-Q curve is above or below the diagonal reference line at the lowest Q_{theory} values.

2.2.5 Impact of Inaccuracies in the PDF Kurtosis

Finally, we investigate the case where the photo-z PDFs differ from the spec-z distribution in kurtosis (β_2), but have the same mean, standard deviation, and skewness (γ_1). Again, we generate the spec-z's from a normal distribution with $\mu_{\text{spec}} = 2$ and $\sigma_{\text{spec}} = 0.1$, whereas the photo-$z$ PDFs are constructed using Gaussian mixture models of the type described previously. The values of the parameters in this case are presented in Table 2.3; again there are some small deviations in the values of the moments of the final PDFs, but these are small enough to have no significant impact on this qualitative investigation.

Plots of the photo-z PDFs with varying kurtosis used for this study are shown in Figure 2.12. We then evaluate the Q-Q curve for each PDF as plotted in Figure 2.13; the case where the PDF is the same as the spec-z distribution (so $\mu_{\text{phot}} = 2$, $\sigma_{\text{phot}} = 0.1$, $\gamma_{1,\text{phot}} = 0$, and $\beta_{2,\text{phot}} = 3$) is again shown as a red solid curve.

As with the other moments, the Q-Q plots in Figure 2.13 show that the larger the difference in kurtosis between the assumed photo-z PDF and the spec-z distribution, the greater the deviation of the curve from the reference line. In contrast, the deviation is minimal when $\beta_{2,\text{phot}} = \beta_{2,\text{spec}} = 3$, simply reflecting sampling noise. Continuing the pattern from the previous moments, the Q-Q curves cross the unit line three times (apart from their endpoints) when the assumed photo-z PDF differs from the spec-z PDF in its kurtosis (compared to zero crossings for a bias in the mean, one for an error in the standard deviation, and two for inaccuracy in the skewness).

We present the values of the normalized ℓ^2−norm and the f_{op} for each assumed photo-z PDF in Table 2.6. As expected, the lowest value of the normalized ℓ^2−norm is for $\beta_{2,\text{phot}} = \beta_{2,\text{spec}} = 3$; the largest is for $\beta_{2,\text{phot}} = 1.498$. This PDF has substantially smaller tails than the spec-z distribution; as a result, f_{op} is also larger, $\sim 3\%$. In contrast, when the PDF kurtosis is larger than for the spec-z's, f_{op} is zero, as the photo-z PDF has longer tails in this case.

Table 2.5 Parameters used for Gaussian mixture models having different values of kurtosis, β_2, while matching the mean, standard deviation, and skewness of a normal distribution. See Equation 2.5 for the definition of these models. Only two Gaussian distributions are needed to construct the final PDFs in this case.

moments	$\mu = 2.001$	$\mu = 2.001$	$\mu = 2$	$\mu = 2$	$\mu = 2$
of	$\sigma = 0.1$	$\sigma = 0.1$	$\sigma = 0.1$	$\sigma = 0.1$	$\sigma = 0.1$
final	$\gamma_1 = 0.001$	$\gamma_1 = 0$	$\gamma_1 = 0$	$\gamma_1 = 0$	$\gamma_1 = 0$
PDF	$\beta_2 = 1.498$	$\beta_2 = 2$	$\beta_2 = 3$	$\beta_2 = 4$	$\beta_2 = 4.503$
c_1	0.497	0.477	1	0.305	0.363
μ_1	1.907	1.913	2	2	2
σ_1	0.036	0.052	0.1	0.036	0.025
c_2	0.503	0.523	-	0.695	0.637
μ_2	2.093	2.081	-	2	2
σ_2	0.037	0.056	-	0.118	0.124

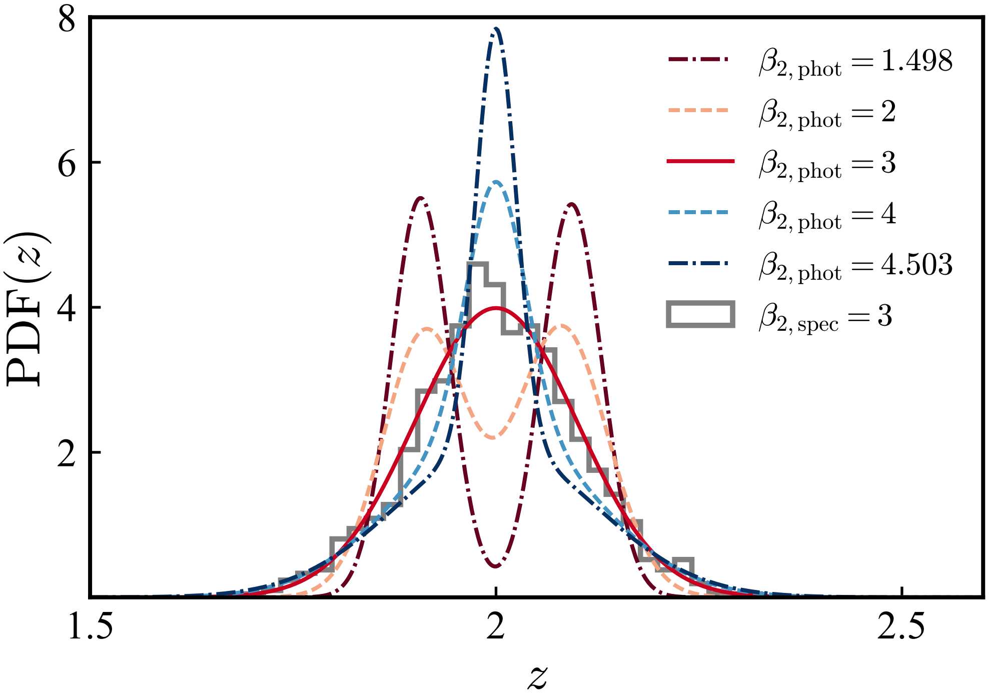


Figure 2.12 Five different cases of photo-z PDFs, with varying kurtosis values $\beta_{2,\mathrm{phot}} = 1.498, 2, 3, 4,$ and 4.503 but skewness $\gamma_{1,\mathrm{phot}} = 0$, along with the normalized histogram of spec-z's generated from a Gaussian distribution with $\mu_{\mathrm{spec}} = 2$, $\sigma_{\mathrm{spec}} = 0.1$, $\gamma_{1,\mathrm{spec}} = 0$, and $\beta_{2,\mathrm{spec}} = 3$. Like the normal distribution of spec-z's, these PDFs are all symmetric about their mean, but they have significantly different degrees of central concentration, reflecting the variations in kurtosis.

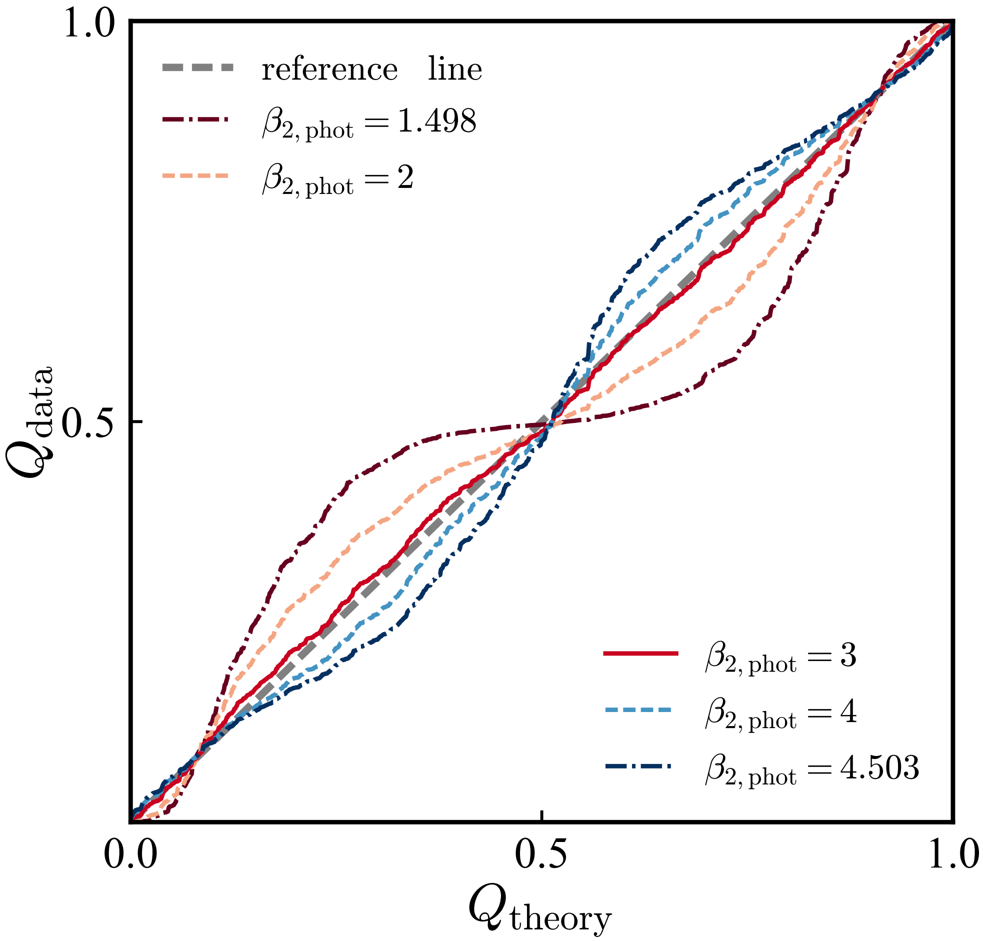


Figure 2.13 Q-Q curves for the five photo-z PDFs, with kurtosis $\beta_{2,\mathrm{phot}}$ = 1.498, 2, 3, 4, 4.503 and skewness $\gamma_{1,\mathrm{phot}}$ = 0, which were depicted in Figure 2.12. The case where the PDF has the same kurtosis as the spec-z distribution $\beta_{2,\mathrm{phot}}$ = 3 is shown by the red solid curve. For the other 4 cases, the bigger the difference in β_2, the larger the deviation of the plotted curve from the reference line.

Table 2.6 Table of values of $\ell^2-\mathrm{norm}$ and f_{op} for photometric redshift PDFs with varying kurtosis, $\beta_{2,\mathrm{phot}}$. Both the ℓ^2−norm and the f_{op} are lowest when the assumed photo-z matches the true spec-z distribution, $\beta_{2,\mathrm{phot}} = \beta_{2,\mathrm{spec}} = 3$.

$\gamma_{1,\mathrm{phot}}$	$\beta_{2,\mathrm{phot}}$	normalized $\ell^2-\mathrm{norm}$	f_{op}
0.001	1.498	0.097	0.029
0	2	0.050	0.005
0	3	0.009	0
0	4	0.025	0
0	4.503	0.047	0

2.3 CALIBRATING PHOTMETRIC REDSHIFT PDFS WITH Q-Q STATISTICS

We can use information from the Q-Q plots and related statistics to optimize photo-z PDFs for objects in a sample so that they better meet the standard statistical definition of a probability density function. In real data sets, the highest-quality spectroscopic redshifts are found to be very reliable (cf., e.g., Newman et al. (2013)) and hence can be considered to represent the true redshifts of the objects they are measured for, enabling tests of PDF accuracy.

Photo-z PDFs can be inaccurate in a wide variety of ways, including biases/shifts, overconfidence/underconfidence, inaccurate levels of symmetry/asymmetry, or having tails that are too long or too short. These issues correspond to PDFs with incorrect values of their first four moments, i.e., the mean, standard deviation, skewness, and kurtosis, respectively. Since it is nontrivial to construct PDFs with arbitrary values of skewness and/or kurtosis while keeping the other moments constant, in this section we focus only the first two of these issues, but similar tests would be possible for other aspects of a PDF. Keeping in mind that for a properly-calibrated PDF the distribution of CDF values evaluated at the spectroscopic redshifts of objects should be uniform, we can use Q-Q statistics measured for objects with spec-z's to recalibrate photo-z PDFs so that they more closely meet the statistical definition.

In this section, as a toy model we consider the case of identifying and removing differences in the mean and in the standard deviation between Gaussian photo-z PDFs and spec-z distributions. However, since Q-Q statistics are based only on CDF values, the same methods can be used to evaluate improvements to non-Gaussian PDFs which vary from object to object. One can minimize the same statistics used here to identify the best shift or change in standard deviation (or equivalently, power to which a PDF is raised) that should be applied to PDFs to enable them to better meet the standard definition of a probability density function.

In the case of a bias, we assume that the mean of the photo-z PDFs differs from the mean of a properly-calibrated PDF (and hence the spec-z distribution) by an amount b, which we will call the bias; $\mu_{\text{phot}} = \mu_{\text{spec}} + b$. The bias can be positive or negative, corresponding to $\mu_{\text{phot}} > \mu_{\text{spec}}$ or $\mu_{\text{phot}} < \mu_{\text{spec}}$, respectively; these biases correspond to PDFs which are shifted to the right or to the left with respect to a properly-calibrated distribution. We will use the Q-Q statistics defined in the previous section as a tool for estimating the bias.

Given an estimate for b, which we label as β, we can correct for it and make the means of the two distributions closer to each other; the shifted probability distribution will have a mean $\mu'_{\text{phot}} = \mu_{\text{phot}} - \beta = \mu_{\text{spec}} + b - \beta$. When the true bias b and the estimated bias β are the same, then this procedure will cause the two means to become exactly equal and the bias will be removed completely (so $\mu'_{\text{phot}} = \mu_{\text{spec}}$). When that happens, the Q-Q curve will lie as close as possible to the diagonal reference line in the Q-Q plot, so the normalized ℓ^2-norm must be minimized. Hence, we can optimally estimate the value of β by identifying the shift that minimizes the normalized ℓ^2-norm in the Q-Q plot. Of course, due to sampling noise in the distribution of spec-z's the value of β will randomly scatter about the true bias of the photo-z PDFs, b; we investigate the amplitude of this scatter below. After determining the value of β which minimizes the distance of the Q-Q curve from the diagonal, we can improve photo-z PDFs by applying a transformation that removes its effect (i.e., shifting all PDFs in redshift by $-\beta$).

We note that instead of using the ℓ^2-norm as the metric of mismatch between the distributions, we could use the ℓ^1-norm or the maximum deviation between the Q-Q curve and the diagonal (a quantity analogous to the K-S test D-value, so we label it as D below). As can be seen in Figure 2.14, all of these metrics would indicate similar offsets for retrieving the true mean μ. In general, all three metrics have similar performance (i.e., minimizing each metric yields similar scatter between β and b); we explore this in detail in Subsection 2.3.1.

We can instead consider the case where the photo-z PDFs have an erroneous standard deviation, corresponding to either overconfident or underconfident PDFs. This difference can be represented by a positive factor a, such that $\sigma_{\text{phot}} = a \times \sigma_{\text{spec}}$. If a > 1, then $\sigma_{\text{phot}} > \sigma_{\text{spec}}$ and the photo-z PDFs are broader than they should be, providing larger uncertainties and therefore underconfident errors. Contrastingly, when $a < 1$ then $\sigma_{\text{phot}} < \sigma_{\text{spec}}$, leading to very narrow photo-z PDFs and small uncertainties, corresponding to overconfident errors.

Given the dependence of the toy model Gaussian photo-z PDFs on the standard deviation parameter, $\mathbf{PDF}(z) \propto \exp\left[-1/\sigma^2_{\text{phot}}\right]$, we can see that raising $\mathbf{PDF}(z)$ to a power γ, such that the new PDF is given by $\mathbf{PDF}'(z) = [\mathbf{PDF}(z)]^\gamma$, and then renormalizing so that the integral of $\mathbf{PDF}'(z)$ is one, is equivalent to changing the standard deviation parameter of the normal distribution to $\sigma'_{\text{phot}} = \sigma_{\text{phot}}/\sqrt{\gamma}$. We note that taking the PDF to a power and renormalizing is a general procedure that can be applied to *any* photometric redshift PDF to make it more or less overconfident. When γ is greater than one, peaks in a PDF become tighter and valleys between them (if present) become deeper; if it is less than one, the opposite is true. Thus, although we focus on tests of changing the standard deviation parameter of a Gaussian distribution here, similar procedures can be applied to improve arbitrary PDFs.

Setting $\gamma = \alpha^2$, we find $\sigma'_{\text{phot}} = (a/\alpha) \times \sigma_{\text{spec}}$. When $a = \alpha$, the standard deviations of the two distributions are exactly equal ($\sigma'_{\text{phot}} = \sigma_{\text{spec}}$); in this case the Q-Q curve should lie along the reference line and our Q-Q distance metrics should all be small. Therefore, we can improve the photo-z PDFs by identifying the value of α that minimizes our distance metrics and then applying a transformation that removes its effect (i.e., taking the PDF to a power $1/\alpha^2$ and then renormalizing).

In Figure 2.15 we show all three Q-Q distance metrics as a function of α for a case where the photo-z PDF has $\sigma_{\text{phot}} = 0.5$, matching the spec-$z$ distribution, for a sample of one thousand mock redshifts. In this case minimizing the $\ell^2\mathbf{-norm}$ performs better than using the other two metrics; this is not universally true (due to sampling noise), but is the

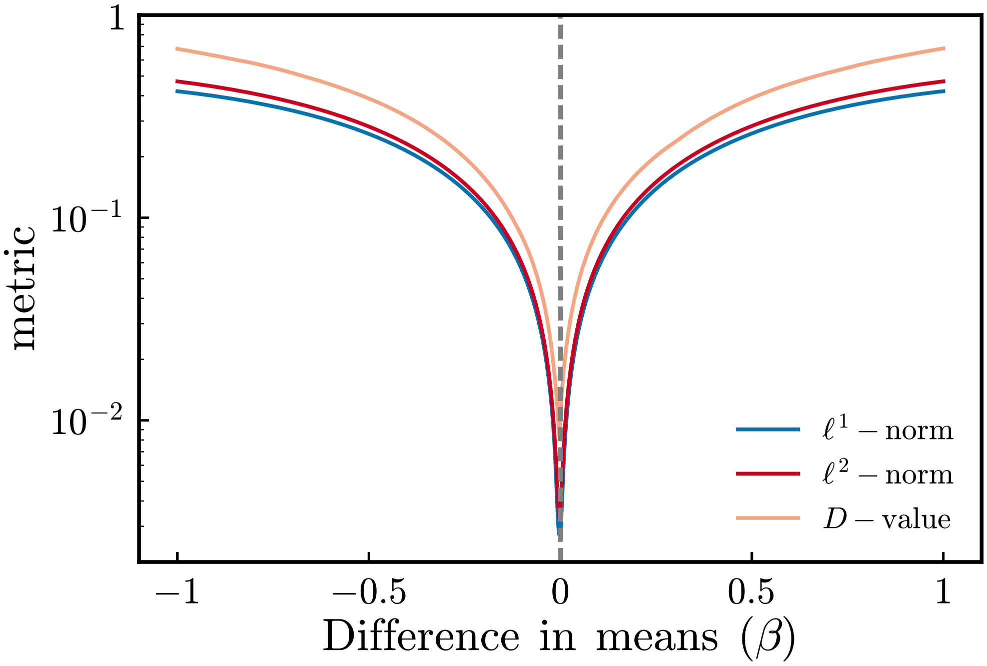


Figure 2.14 Three different metrics of distance between the Q-Q curve and the diagonal – the ℓ^1–norm, ℓ^2–norm, and maximum deviation (analogous to the K-S D-value), plotted as a function of possible values of the bias β between the mean of the assumed photo-z PDF and the true distribution. The intrinsic redshift distribution adopted is a Gaussian with mean $\mu_{\text{spec}} = 2$ and standard deviation $\sigma_{\text{spec}} = 0.5$. All three curves were generated using a dataset of one thousand redshifts and with a photo-z PDF which was unbiased relative to the true distribution; in this case, all curves are minimized at a bias of $\beta \approx 0$, successfully recovering the correct answer to better than 0.01 in z.

most common case. We explore the performance of each statistic in Subsection 2.3.2.

In the following subsections, we investigate in detail how well we are able to retrieve the values of b and a for different values of the parameters defining the true redshift probability distribution and for different sizes of the sample of spectroscopic redshifts. More specifically, we will explore how the retrieval of the true bias, b (or the true standard deviation ratio, a) depends on the value of the true mean redshift (which we would not expect to make a difference), the underlying standard deviation of the distribution, and the size of the sample of objects used in the analysis.

2.3.1 Identifying a PDF bias

We first consider the case where we wish to identify the overall bias (i.e., difference in mean redshift) between a photo-z PDF and an ideal distribution. We begin by creating a sample of objects of size N_s, for which we generate spectroscopic redshifts from a Gaussian distribution of μ_{spec} and σ_{spec}. We then adopt a nominal photo-z PDF described by a Gaussian distribution with mean μ_{phot} and σ_{phot}, which is the same for all objects of the sample, but differs from the PDF from which spectroscopic redshifts are generated by having a different mean; $\mu_{\text{phot}} = \mu_{\text{spec}} + b$. We then consider a series of photo-z PDFs each described by a Gaussian distribution with mean $\mu_{\text{test}} = \mu_{\text{spec}} + \beta$ and standard deviation σ_{test}, keeping $\sigma_{\text{test}} = \sigma_{\text{spec}}$ but varying β over a grid of 101 values from $-6\sigma_{\text{spec}}/\sqrt{N_s}$ to $+6\sigma_{\text{spec}}/\sqrt{N_s}$.

Using the same set of spec-z's, we construct the Q-Q curve for the PDF with each value of β on the grid and determine all three distance metrics (normalized ℓ^2-norm, normalized ℓ^1-norm, and D-value) as a function of β. The optimal value of β will be the one that minimizes the chosen Q-Q distance metric. We repeat the same process with different randomly-generated spectroscopic redshift samples a number of times N_r and calculate the quantity $\text{RMS}(\beta_{\text{opt}} - b)$, where β_{opt} is the value of β for which the chosen metric is minimized in a given random realization. This quantity measures the uncertainty in retrieving the bias in the photo-z PDF. We note that if b were non-zero, the only consequence would be a shift in the β value where the minimum occurs by that amount. As a result, we adopt $b = 0$ (i.e., a nominal photo-z PDF which has zero bias compared to the true distribution) as our test

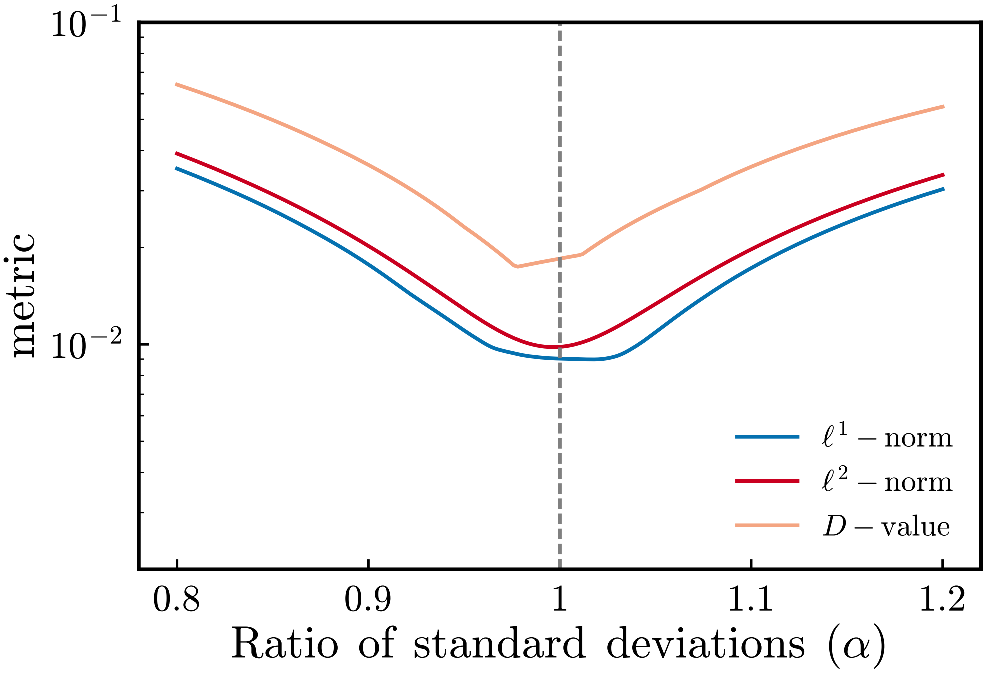


Figure 2.15 Three different metrics of distance between the Q-Q curve and the diagonal – the ℓ^1–norm, ℓ^2–norm, and maximum deviation (analogous to the K-S D-value), plotted as a function of possible values of the ratio between the standard deviation of the assumed photo-z PDF and the true distribution, α. Again we adopt a Gaussian with mean $\mu_{\text{spec}} = 2$ and standard deviation $\sigma_{\text{spec}} = 0.5$ for the true redshift distribution. All three curves were generated using a dataset of one thousand redshifts and with a photo-z PDF which was unbiased relative to the true distribution; in this case, the minimum of the ℓ^2–norm occurs closest to the true value (corresponding to $\alpha = 1$).

case for simplicity; $\mathrm{RMS}(\beta_{\mathrm{opt}} - b)$ should not be affected by this choice (we demonstrate that the error in bias recovery does not depend upon the mean redshift below).

In general, we would expect the uncertainty in the ordinary mean of the redshift of a Gaussian-distributed spec-z sample to be given by $\mathrm{RMS}_{\mathrm{mean}} = \sigma_{\mathrm{spec}}/\sqrt{N_s}$, the typical standard error formula. This would correspond to the error in recovering b resulting from comparing the mean redshift predicted from the photo-z PDF to the mean redshift of the sample to determine the bias. In our case, we are minimizing a very different quantity – the average distance of the Q-Q curve from the diagonal, rather than the sum of the squares of the deviations of the redshifts in the sample from a value – but it is still reasonable to expect a similar scaling of errors in this case. Given this, we make the assumption (which we test below) that $\mathrm{RMS}(\beta_{\mathrm{opt}} - b) = c_\mu \times \mathrm{RMS}_{\mathrm{mean}} = c_\mu \times \sigma_{\mathrm{spec}}/\sqrt{N_s}$, where c_μ is a coefficient by which the value of RMS obtained by minimizing a Q-Q distance metric differs from the standard error ($c_\mu = \mathrm{RMS}_{\mathrm{data}}/\mathrm{RMS}_{\mathrm{mean}}$).

For the scenario we consider in our toy model the ordinary mean minimizes the RMS error in the redshift bias by construction, as it is a least-squares estimator. However, it is important to recognize that it is optimal only under relatively strong assumptions (of Gaussianity and uniform errors) which do hold in the case of our mock data but are rarely if ever true for the ensemble of photo-z PDFs in a real sample. Even though we are using Gaussian distributions for our mock data in this study for simplicity, we are using those distributions to explore methods which are applicable to the arbitrarily variable PDFs which occur in real data.

We first test the dependence of the error in $\beta - b$ on the size of the spectroscopic sample used to determine the Q-Q curve. For the largest spectroscopic sample size we consider, with $N_s = 3200$ objects per realization, we create $N_r = 400$ realizations of spectroscopic samples of size N_s and then calculate $\mathrm{RMS}(\beta - b)$ from the distribution of values of this quantity resulting from the realizations. However, we also have tested the recovery of β with smaller samples for which $\mathrm{RMS}(\beta_{\mathrm{opt}} - b)$ will be larger (as the standard error of the mean is proportional to $\frac{1}{\sqrt{N}}$). The standard error in the standard deviation (or, in this case, the RMS) for normally-distributed data is given by $\sigma_{\mathrm{RMS}} = \mathrm{RMS}/\sqrt{2N}$, where N is the number of data points from which the RMS is calculated. In our application, N corresponds to the number of realizations, N_r.

As a result, we can prevent the uncertainty in $\mathrm{RMS}(\beta_{\mathrm{opt}} - b)$ from depending on the size of the spectroscopic sample by altering the number of realizations such that $N_s \times N_r$ remains constant. Therefore for sample sizes $N_s = 50, 100, 200, 400, 800, 1600,$ and 3200 we use $N_r = 25600, 12800, 6400, 3200, 1600, 800,$ and 400 realizations, respectively. This would be expected to yield *uncertainties* in $\mathrm{RMS}(\beta_{\mathrm{opt}} - b)$ which are the same at all N_s.

The results of this analysis are shown in Figure 2.16. In this plot, the blue points with error bars represent the measured RMS error in recovering b by minimizing the normalized $\ell^2\mathrm{-norm}$ between the Q-Q curve and the reference line from the N_r different realizations of a spectroscopic dataset. The red dashed line corresponds to a curve of the form $\mathrm{RMS}(\beta - b) =< c_\mu > \times \sigma_{\mathrm{spec}}/\sqrt{N_s}$, with $< c_\mu >$ given by the mean value of the coefficient across the different sample sizes. In parentheses we provide the uncertainty in this mean value, which is given by the standard deviation of the set of values of c_μ from each sample size (calculated with one degree of freedom) divided by $\sqrt{N_v}$, where N_v is the number of different values of c_μ which were averaged (seven in this case).

We next investigate the case of changing the value of μ_{spec} – which is equivalent in effect to changing b by $-(\mu_{\mathrm{spec}} - 2)$ – while keeping the other two quantities unchanged (at $\sigma_{\mathrm{spec}} = 0.5$, and $N_s = 3200$). So long as $\mu_{\mathrm{spec}}/\sigma_{\mathrm{spec}}$ is large (so that the unphysicality of redshifts below zero does not matter), we expect $\mathrm{RMS}(\beta - b)$ not to depend on the mean, since in the analogous case of the standard error the uncertainty only depends on the standard deviation of the distribution and the sample size, but not the sample mean. This expectation is borne out in Figure 2.17. Again we have plotted the values of $\mathrm{RMS}(\beta - b)$ as blue points with error bars, this time as a function of the mean redshift of the spectroscopic sample. The red dashed line represents the curve $< c_\mu > \times \sigma_{\mathrm{spec}}/\sqrt{N_s}$, where $< c_\mu >$ is the mean value of the coefficient c_μ determined from each plotted point, as in the previous case. We again determine the uncertainty in $< c_\mu >$ via the standard error formula; in this case, $N_v = 10$ have been averaged. As seen in the figure, this simple model is consistent with all of the simulation samples at $< 2\sigma$, suggesting that our assumption that the uncertainty in recovering b should be proportional to the ordinary standard deviation of the mean.

Finally, we investigate the case of changing the standard deviation of the distribution of

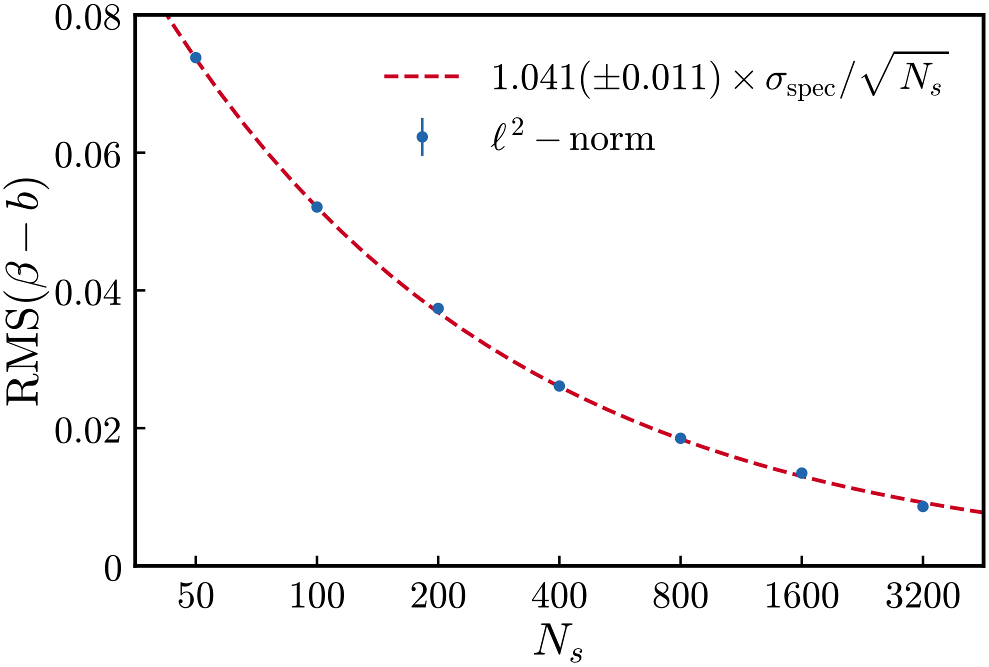


Figure 2.16 **RMS** values of the difference between the recovered PDF redshift bias and the true bias ($\beta - b$) resulting from Q-Q analysis, as a function of the size of the spectroscopic sample used to construct the Q-Q curve, $\mathrm{N_s}$ (blue points with error bars). The plotted points are the results from Monte Carlo simulations done using a Gaussian distribution of redshifts with mean $\mu_{\rm spec} = 2$ and standard deviation $\sigma_{\rm spec} = 0.5$, testing against Gaussian photo-z PDFs with the same standard deviation ($\sigma_{\rm phot} = 0.5$) but varying means ($\mu_{\rm phot} = \mu_{\rm spec} + \beta$), and then determining the value of β which minimizes the normalized ℓ^2-norm between the Q-Q curve and the unit line. As one would expect, the uncertainties are reduced as the size of the spectroscopic sample used to estimate the photo-z PDF bias increases. The red dashed line represents the curve $< c_\mu > \times \sigma_{\rm spec}/\sqrt{N_s}$, where $< c_\mu >$ is the mean value of the coefficient c_μ determined from each plotted point; this corresponds to a model where the error in recovering b via Q-Q statistics is a constant c_μ times the standard deviation of the mean for the spectroscopic sample. The simulations we have done are sufficient to determine $< c_\mu >$ to $\sim 1\%$.

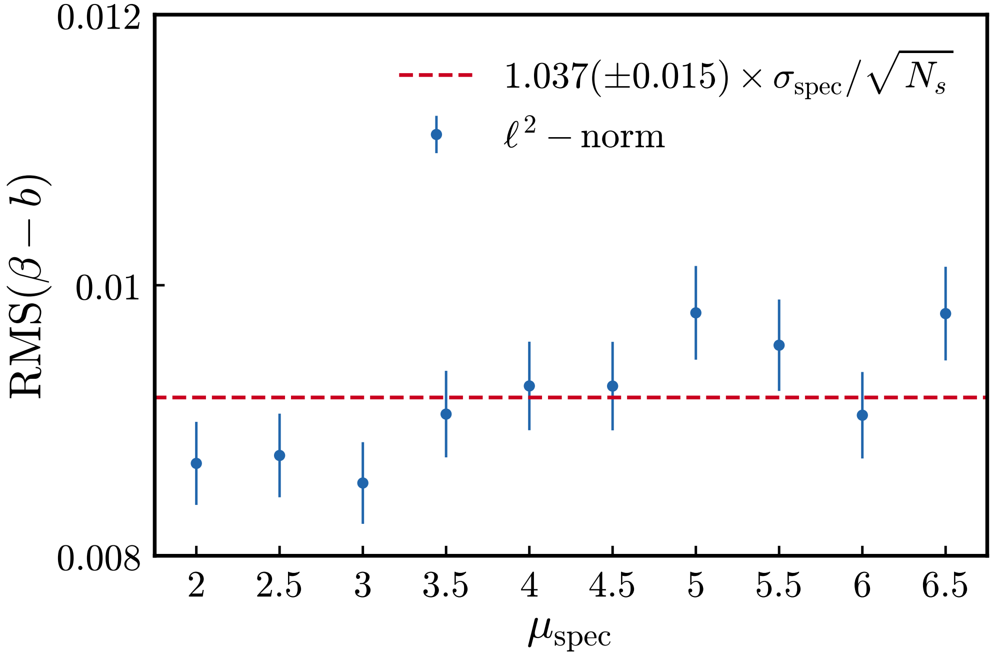


Figure 2.17 RMS values of the offset in the redshift bias, $\beta - b$, as a function of the mean redshift of the spectroscopic sample $\mu_{\rm spec}$ (equivalent to varying b), resulting from minimizing the normalized ℓ^2-norm of the Q-Q curve. Results are based on Monte Carlo simulations of Gaussian redshift distributions with standard deviation $\sigma_{\rm spec} = 0.5$. A total of $N_r = 400$ random realizations of spectroscopic samples of size $N_s = 3200$ were generated to calculate the mean and standard error in the standard deviation for each plotted point. The red dashed line again represents the curve $< c_\mu > \times \sigma_{\rm spec}/\sqrt{N_s}$; this quantity is independent of $\mu_{\rm spec}$ (or equivalently b). All points agree with the expectation that the RMS values are independent of $\mu_{\rm spec}$ at $< 2\sigma$.

redshifts, σ_{spec}, keeping $\mu_{\text{phot}} = \mu_{\text{spec}} = 2$ and sample size $N_s = 3200$, and generating $N_r = 400$ realizations for each value of σ_{spec} considered. In this case we expect a linear increase of the RMS values with σ_{spec} since the standard error is directly proportional to that quantity.

Again, our results are consistent with expectations, as seen in Figure 2.18. Once again blue points with error bars represent the values of $\text{RMS}(\beta - \text{b})$ resulting from our Monte Carlo simulations, while the red dashed line represents the quantity $< c_\mu > \times \sigma_{\text{spec}}/\sqrt{N_s}$. As before, $< c_\mu >$ is the mean value of the coefficient c_μ at each of the ten plotted points, and we indicate its standard error in parentheses.

Based upon the tests we have done varying the sizes of spectroscopic samples, the mean redshift of the sample, and the standard deviation of the redshifts, it appears a model where the errors in recovering the redshift bias b are proportional to the standard error in the ordinary mean redshift is highly effective in capturing our results. Furthermore, the constant of the proportionality is only slightly larger than one; $< c_\mu >$ is less than 1.05 in all cases. This indicates that minimizing the normalized ℓ^2-norm between the Q-Q curve and the unit line recovers a redshift bias almost as well as a statistic which by construction minimizes the RMS about b for the scenario we have implemented in our toy model (i.e., values drawn from identical Gaussian distributions). However, unlike the ordinary mean, the use of Q-Q statistics depends on no assumptions of Gaussianity or of identicality between the PDFs for different objects; each value of Q is calculated using a given object's cumulative distribution function, whatever its form may be. We therefore anticipate that recovering a bias with Q-Q statistics should be a much more generally useful method than simply determining a shift via the ordinary mean.

2.3.2 Identifying errors in the PDF width

In this subsection we perform a similar analysis to the previous one, with the difference being that we explore the utility of Q-Q statistics for determining any inaccuracy in the standard deviation of the distribution (σ_{phot}), instead of in the mean (μ_{spec}). We focus on the determination of the ratio of the estimated standard deviation (from an assumed photo-z PDF) to the true standard deviation, the parameter a described in Section 2.3. We

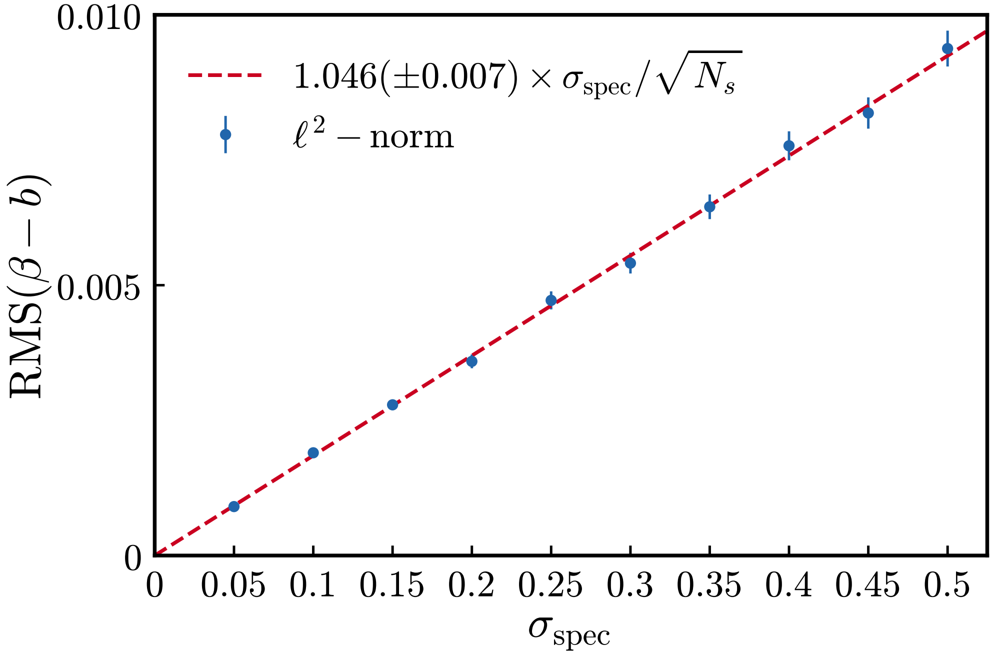


Figure 2.18 RMS values of the offset between the recovered and true redshift bias, $\beta - \mathrm{b}$, as a function of the standard deviation of the redshift distribution, $\sigma_{\rm spec}$. The blue points with error bars are calculated from $N_r = 400$ realizations of spectroscopic samples with mean redshift $\mu_{\rm spec} = 2$ and sample size $N_s = 3200$, while the red dashed line corresponds to the function $< c_\mu > \times \sigma_{\rm spec}/\sqrt{N_s}$. As expected from this function, the RMS errors measured from Monte Carlo simulations increase in a fashion directly proportional to $\sigma_{\rm spec}$.

estimate this quantity by generating PDFs of varying $\alpha = \sigma_{\rm phot}/\sigma_{\rm fiducial}$, where $\sigma_{\rm fiducial}$ is a fiducial value of the standard deviation used to normalize (here, we set it equal to $\sigma_{\rm spec}$ for convenience), and then determining the optimum value of α as the value which minimizes a Q-Q distance metric (e.g., the normalized ℓ^2-norm).

We will again assume that all photo-z PDFs are identical Gaussian distributions for convenience. In this case, the optimum least-squares estimator of the σ parameter is the ordinary sample standard deviation, which has standard error $\text{RMS}(\sigma_{\rm recovered} - \sigma_{\rm spec}) = \sigma_{\rm spec}/\sqrt{2N_s}$, where N_s is the number of spectroscopic redshifts in the sample. As before, we make the assumption that the error in the estimate of the standard deviation obtained by Q-Q statistics will be proportional to the optimum standard error; i.e., $\text{RMS}(\sigma_{\rm recovered} - \sigma_{\rm spec}) = \frac{c_\sigma \times \sigma_{\rm spec}}{\sqrt{2N_s}}$, where c_σ is a factor of order unity (for convenience, we will refer to the recovered standard deviation as $\sigma_{\rm phot}$ below).

We note that changing the standard deviation parameter of a Gaussian distribution by a factor α is equivalent to raising it to a power $1/\alpha^2$ and then renormalizing the distribution to have integral one (as described above). Hence, our analysis of the recovery of an optimum change in the standard deviation parameter by a factor α is equivalent to optimizing the power than a PDF should be taken to so that the quantiles of the observed spectroscopic sample better follow a uniform distribution. As a result, the toy model considered in this section can be used to represent the case of taking arbitrary PDFs to a power, rather than multiplying the σ parameter of a Gaussian by a constant factor.

Following the same procedure as before, we generate sets of Monte Carlo realizations of spectroscopic samples drawn from Gaussian distributions of varying characteristics, determine the optimum value of $\sigma_{\rm phot}$ (or equivalently, α) for each realization, and calculate $\text{RMS}(\sigma_{\rm phot} - \sigma_{\rm spec})$ from the results. Once more, we balance the number of redshifts in the spectroscopic sample, N_s, with the number of realizations, N_r, such that the quantity $N_s \times N_r$ remains constant. This again should ensure that the errors in the RMS remain approximately constant, since $\sigma_{\rm RMS} = \text{RMS}/\sqrt{2N_r} = c_\sigma \times \sigma_{\rm spec}/\sqrt{2N_s}/\sqrt{2N_r} \propto 1/\sqrt{N_s \times N_r}$. Therefore, as before, for sample sizes $N_s = 50$, 100, 200, 400, 800, 1600, 3200 we use $N_r = 25600$, 12800, 6400, 3200, 1600, 800, 400 realizations, respectively, to investigate the accuracy of parameter recovery.

As in the previous subsection, we use blue points with error bars to show the values of $\mathrm{RMS}(\sigma_{\mathrm{phot}} - \sigma_{\mathrm{spec}})$ resulting from each set of realizations, whereas the red dashed line shows the best-fit curve of the form $\mathrm{RMS}(\sigma_{\mathrm{phot}} - \sigma_{\mathrm{spec}}) =< c_\sigma > \times \sigma_{\mathrm{spec}}/\sqrt{2N_s}$, where $< c_\sigma >$ is the mean value of the coefficient c_σ calculated from the set of blue points. We specify the standard error in c_σ in parentheses; the number of independent points is $N_v = 7$ for the case where we vary sample size and $N_v = 10$ for cases where we vary μ_{spec} or σ_{spec}.

We first explore the dependence of the measured RMS error on the sample size N_s. For this, we use a photometric redshift PDF which is a Gaussian distribution with parameters $\mu_{\mathrm{spec}} = 2$ and $\sigma_{\mathrm{spec}} = 0.5$, and investigate how well we recover the standard deviation using values of σ_{phot} from a grid of 101 values in the interval $\sigma_{\mathrm{spec}} \pm 8.5\sigma_{\mathrm{spec}}/\sqrt{2N_s}$. Figure 2.19 shows that the results agree very well with our expectations from a model where the error in recovering σ_{phot} is proportional to the standard error in the standard deviation, corresponding to the red dashed line.

Next, we explore the impact of changing the value of the mean redshift, μ_{spec} while keeping $\sigma_{\mathrm{spec}} = 0.5$ and the sample size $N_s = 3200$. As in the case of a redshift bias, we expect the errors in reconstructing the standard deviation not to depend on the mean, by analogy to the standard error of the standard deviation, which only depends on the true value of the standard deviation and the sample size. This expectation is in good agreement with the results shown in Figure 2.20; all results are consistent with a constant value at $< 2\sigma$.

Finally, we investigate the case of changing σ_{spec}, while keeping the mean spectroscopic redshift as $\mu_{\mathrm{spec}} = 2$ and the sample size as $N_s = 3200$. In this case, we find that the RMS error in reconstructing σ is directly proportional to the true standard deviation, as seen in Figure 2.21. This matches what one would expect if errors resulting from Q-Q analysis are proportional to the standard error in the standard deviation, as we have assumed.

In Table 2.7 we compile the estimated values for $< c_\mu >$ and $< c_\sigma >$ together, along with their respective uncertainties, derived from the results shown in Figures 2.16 – 2.21. We also provide similar summary results resulting from analyses identical to those shown in these figures, except we have minimized the ℓ^1-norm or the D-value between the Q-Q curve and the unit line, rather than the ℓ^2-norm. We also include in this table the inverse-variance-

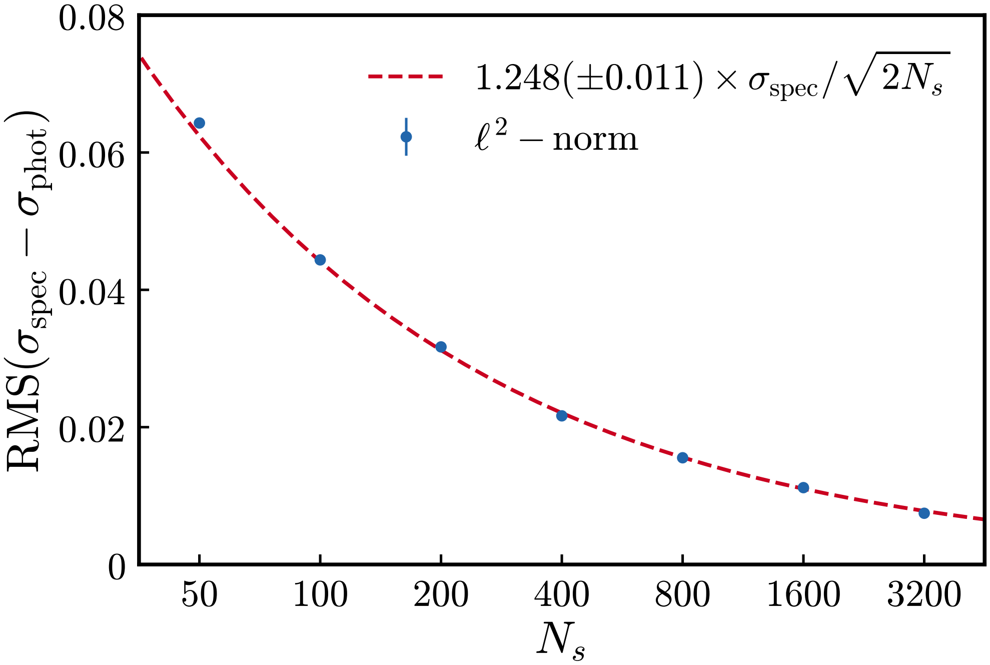


Figure 2.19 RMS error in the difference between the photo-z PDF standard deviation reconstructed from Q-Q analysis, $\sigma_{\rm phot}$, and the true value, $\sigma_{\rm spec}$, as a function of the size of the spectroscopic sample used for the analysis, N_s (blue points with error bars). We assume all objects are drawn from identical Gaussian PDFs with mean $\mu_{\rm spec} = 2$ and that the true distribution has $\sigma_{\rm spec} = 0.5$. As expected, the RMS values decrease with increasing sample size. The red dashed line corresponds to a function of the form $< c_\sigma > \times \sigma_{\rm spec}/\sqrt{2N_s}$, i.e., one differing from the standard error in the standard deviation of the spectroscopic redshifts by a constant factor $< c_\sigma >$.

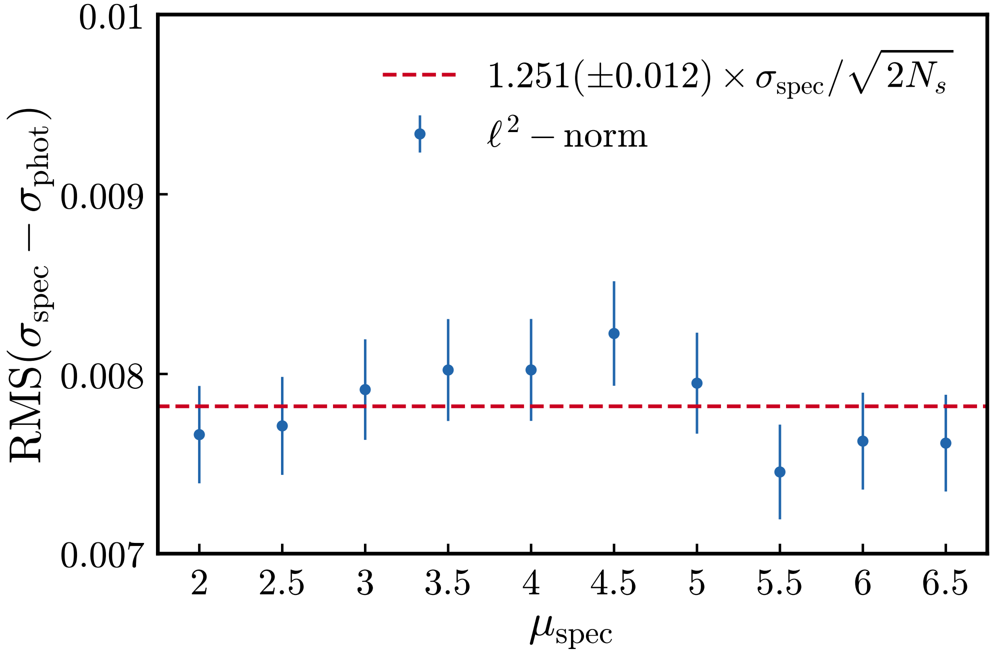


Figure 2.20 RMS of the difference between the photo-z PDF standard deviation reconstructed from Q-Q analysis, $\sigma_{\rm phot}$, and the true value, $\sigma_{\rm spec}$, as a function of the mean redshift of the spectroscopic sample used for the analysis, $\mathrm{mu}_{\rm spec}$ (blue points with error bars). Each point is calculated from $N_r = 400$ realizations of spectroscopic samples of size $N_s = 3200$ drawn from a Gaussian distribution with standard deviation $\sigma_{\rm spec} = 0.5$. The red dashed line corresponds to a function of the form $< c_\sigma > \times \sigma_{\rm spec}/\sqrt{2N_s}$, which is independent of $\mu_{\rm spec}$. As expected, we find no significant dependence of the RMS values on the mean redshift of the spectroscopic sample.

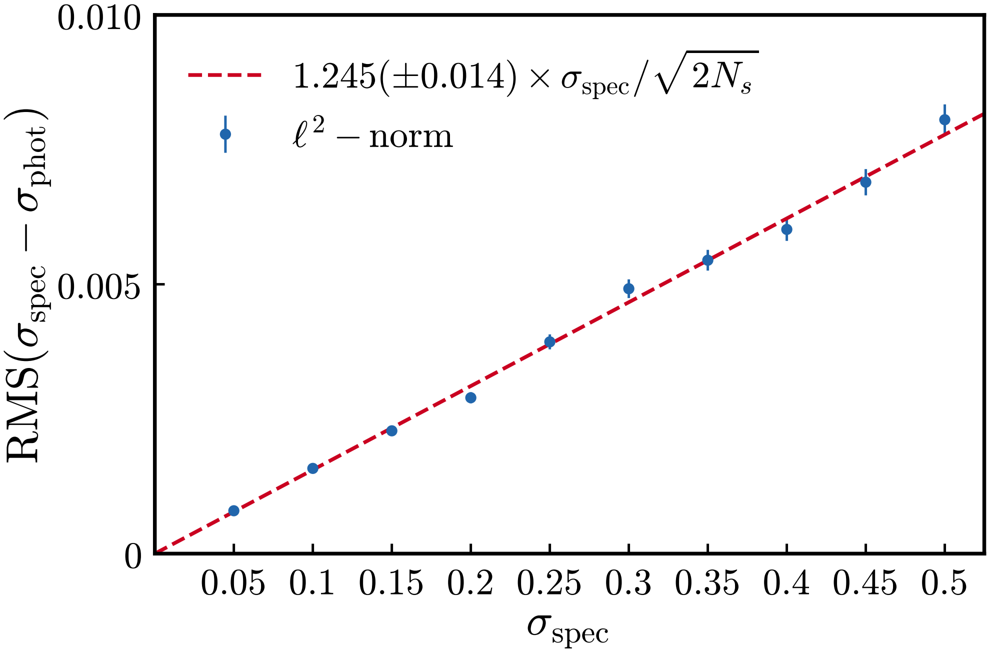


Figure 2.21 RMS of the difference between the photo-z PDF standard deviation reconstructed from Q-Q analysis, $\sigma_{\rm phot}$, and the true value, $\sigma_{\rm spec}$, as a function of the standard deviation of the distribution of redshifts of the spectroscopic sample, σspec (blue points with error bars). The results shown were obtained using $N_r = 400$ realizations of spectroscopic samples of size $N_s = 3200$ with mean redshift $\mu_{\rm spec} = 2$. The RMS values increase linearly with $\sigma_{\rm spec}$, consistent with the red dashed curve which corresponds to the predicted values if the errors resulting from Q-Q analysis are proportional to the standard error in the standard deviation.

weighted mean of all estimates of $< c_\mu >$ or $< c_\sigma >$ using a given estimator, as well as the corresponding uncertainties derived via propagation of errors. These summary statistics allow us to compare the effectiveness of all three metrics at recovering a bias or error in the standard deviation to each other, as well as to assess how close they come to the results of the optimal estimators for data drawn from identical Gaussian distributions, which would correspond to values $< c_\mu > = < c_\sigma > = 1$.

All three distance metrics yield similar results only slightly worse than the ordinary mean (the optimal estimator for our toy model case) for identifying a bias. The $\ell^2-\mathrm{norm}$ gave the lowest weighted mean value for $< \mathrm{c}_\mu >$, while the $\ell^1-\mathrm{norm}$ for $< \mathrm{c}_\sigma >$, but the differences in most cases are not statistically significant. However, for identifying errors in the standard deviation, the $\boldsymbol{D}$-value proved greatly inferior to the other two metrics. So long as the ℓ^1 or ℓ^2-norm is used as a distance metric, the Q-Q statistics considered here proved only modestly inferior to the optimal estimators for our toy model case, with errors $\sim 1.05\times$ larger for recovering a mean shift and $\sim 1.2\times$ larger at recovering an error in the standard deviation parameter.

2.4 SUMMARY AND DISCUSSION

In this paper, we have presented new methods for improving photometric redshift probability density functions such that they better match the statistical definition of a PDF, based upon statistics of the quantile-quantile (Q-Q) curve. We have explored these methods using mock data consisting of sets of photo-z PDFs which are identical probability distributions and corresponding sets of spectroscopic redshifts which are drawn from a single Gaussian (not necessarily matching the assumed photo-z PDF). These scenarios are not intended to accurately represent real-world data, but rather provides simplified testbeds to explore methodologies. The main results of our study are:

- We have found that a variety of ways in which assumed photo-z PDFs and spectroscopic redshift distributions that provide ground truth differ each present distinct signatures in the Q-Q plots. There are clear deviations of the Q-Q curve from the ideal (unit) line

Table 2.7 Values of the scaling coefficients $< c_\mu >$ and $< c_\sigma >$ and their errors, which correspond to the ratio of the RMS error in recovering an offset in the mean redshift or the standard deviation to the standard error for that quantity. We present values resulting from minimizing three different metrics of the distance between the Q-Q curve and the unit line – the normalized ℓ^2-norm, the normalized ℓ^1-norm, and the $\boldsymbol{D}$-value (so labeled because it is analogous to the K-S $\boldsymbol{D}$ statistic) – for the scenarios presented in Figures 2.16 – 2.21 (which show only the results from minimizing the normalized ℓ^2-norm). For the scenarios considered here, the ordinary mean and sample standard deviation are the optimal estimators, and would yield results corresponding to $< c_\mu > = < c_\sigma > = 1$.

quantity	quantity	scaling coefficients		
retrieved	varied	ℓ^2–norm	ℓ^1–norm	D – value
μ	N_s	1.041 ± 0.011	1.055 ± 0.012	1.050 ± 0.010
μ	μ	1.037 ± 0.015	1.045 ± 0.016	1.050 ± 0.015
μ	σ	1.046 ± 0.007	1.062 ± 0.008	1.054 ± 0.007
weighted	mean	1.044 ± 0.005	1.058 ± 0.006	1.052 ± 0.005
σ	N_s	1.248 ± 0.011	1.271 ± 0.017	1.833 ± 0.022
σ	μ	1.251 ± 0.012	1.218 ± 0.011	1.813 ± 0.017
σ	σ	1.245 ± 0.014	1.230 ± 0.009	1.832 ± 0.020
weighted	mean	1.248 ± 0.007	1.232 ± 0.007	1.824 ± 0.011

when the two distributions (spec-z and photo-z) differ in their mean, standard deviation, skewness, and kurtosis. In each case we find that the more the two distributions differ, the larger the deviations of the curves from the diagonal reference line. We note again that because the Q-Q plot is constructed using only the set of cumulative distribution function values evaluated at the observed spectroscopic redshifts, although simplified probability distributions were used for these analyses, the results should extend to arbitrary probability distributions so long as all PDFs are inaccurate in the same way (e.g., a constant redshift bias). For the same reason, although a single distribution was used as a photo-z PDF for all the objects in a sample in most tests in this paper, we expect the same results if each object has its own independent photo-z PDF, as demonstrated in Figure 2.5.

- We find that statistics derived from the Q-Q plot can provide useful information regarding the accuracy of photo-z PDFs. The closer the Q-Q curve is to the unit line, the closer the distribution of CDF values evaluated at the spectroscopic redshifts are to a uniform distribution, as would be expected if the photo-z PDFs are properly calibrated and fulfill the statistical definition of a probability density function. We have used three metrics to measure differences from the unit line: the normalized ℓ^1−norm, the normalized ℓ^2−norm, and the maximum separation between the two (analogous to the K-S test D-value). By minimizing a chosen metric as a function of parameters describing PDF inaccuracies, we can determine ways in which the photo-z probability density functions can be altered to improve them. All three metrics yielded similar results for recovering a redshift bias, with errors in the best-fit shift only $\sim$ 5% worse than those obtained with an optimal estimator, but the $\boldsymbol{D}$-value had comparatively poor performance at recovering errors in PDF width. An additional useful statistic we have identified is the fraction of objects with spec-z's that lie outside the photo-z PDFs ($f_{\rm op}$), which provides a PDF-based analog to the catastrophic outlier rates used to evaluate point photometric redshift statistics.

Unlike the ordinary mean and the sample standard deviation, minimizing the Q-Q statistics we have utilized here provide a much more generally applicable way of constraining common-mode PDF inaccuracies. The construction of these statistics has no dependence on the assumptions of identical Gaussian distributions that underly the derivation of the ordinary mean and sample standard deviation. They should also be more robust to outliers (e.g., incorrect redshifts in a spectroscopic sample, which commonly occur); large changes to the redshifts for a small fraction of the sample will shift quantiles by only a small amount, but will have a larger effect on the mean and especially the sample standard deviation. In cases where photo-z PDFs are available for all objects and spectroscopic redshifts have been measured for a broad subsample, minimizing Q-Q distance metrics should therefore provide a very useful way of characterizing PDF inaccuracies, enabling them to be corrected for.

Of course, the Q-Q plot is far from the only method available for studying the accuracy of photo-z PDFs. An example of an alternative approach is presented in Wittman et al. (2016), which focuses on the distribution of values of the integral of that portion of the photo-z PDF that is larger than the value of the PDF evaluated at the spectroscopic redshift (the "highest probability density confidence interval" or HPD CI), rather than the distribution of CDF values, which integrate the PDF from $z = 0$ to $z = z_{\rm spec}$. The distribution of the HPD CI can provide direct information regarding the standard deviation (i.e., whether PDFs are too narrow or too broad) and the kurtosis (i.e., whether the photo-z PDFs have tails that are too long or too short). On the other hand, due to its symmetric definition, the HPD CI is not sensitive to the presence of a bias or asymmetry in PDFs, nor does it give information on their sign (unlike Q-Q curves). An additional advantage of simple quantile statistics are that they are considerably easier to calculate computationally than the HPD CI.

In this work, we have used simplified scenarios to explore the usefulness of the Q-Q plot for identifying inaccuracies in photometric redshift PDFs. In an upcoming paper (Kodra et al. 2019, in prep.), we demonstrate that these methods can be effective in real-world scenarios as well, using data from the Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey (CANDELS, Grogin et al. (2011); Koekemoer et al. (2011)); we demonstrate there that applying corrections derived by minimizing Q-Q distance metrics for one sample of galaxies with spectroscopic redshifts also improves photo-z PDFs for independent sets of spec-z's. The resulting corrections have been used to derive the final photometric redshift probability density functions for CANDELS. The Q-Q metrics presented here are also being used by the Large Synoptic Survey Telescope Dark Energy Science Collaboration (LSST DESC, LSST Dark Energy Science Collaboration (2012)) to assess a variety of photometric redshift codes (Schmidt et al. 2018, in prep.). Many probes of LSST cosmology depend critically on the availability of accurate photometric redshift PDFs for their calculations; the methods presented here provide an avenue for ensuring that this is achieved.

3 OPTIMIZED PHOTOMETRIC REDSHIFTS FOR THE COSMIC ASSEMBLY NEAR-INFRARED DEEP EXTRAGALACTIC LEGACY SURVEY (CANDELS)

3.1 INTRODUCTION

In order to study how galaxies evolve over cosmic time, measurements of distances and lookback times are needed; we can determine both from a galaxy's redshift. Redshifts can be measured most accurately by acquiring detailed spectra, but obtaining useful signal-to-noise ratios requires infeasible exposure times when studying galaxies at very large distances as they are so faint. As an alternative to spectroscopy, broad-band photometry in multiple filters provides information on all objects within the imaging field of view.

Photometric redshifts (also known as photo-z's) can then be estimated using the measured fluxes of the objects in each filter. Photometry is our primary means for studying the faintest and most distant objects, but the information provided is limited; as a result, photometric redshifts have considerably lower precision than those obtained spectroscopically.

There exist a variety of codes which compute probability density functions (PDFs) for photometric redshifts by comparing the observed fluxes for an object in different filters to templates drawn from various libraries of restframe spectral energy distributions (SEDs) of galaxies. Template-based methods generally derive a posterior probability distribution using a prior on redshift determined from an object's magnitude in some band, which is multiplied by a likelihood proportional to $e^{-\chi^2/2}$, where in this case χ^2 is determined using the observed flux of an object compared to the flux predicted for a given template and redshift. Other "training-based" codes attempt to determine the relationship between galaxy colors and redshift by applying machine learning methods to samples of objects with known spectroscopic redshifts.

Template-fitting codes tend to extrapolate relatively well to regimes where training spectroscopic redshifts are not available. In contrast, empirical, training-based codes, can match or outperform the best template-fitting codes with minimal effort in cases where spectroscopic redshifts are available for a fair sample of the objects for which photo-z's will be derived, but extrapolate very poorly outside the range of spectroscopic coverage.

In this study we present catalogs of optimized photometric redshift PDFs produced by the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) collaboration (Grogin et al. (2011); Koekemoer et al. (2011)). Six participating groups used the CANDELS photometric catalogs and a training set of spectroscopic redshifts in order to train their codes and evaluate PDFs for all objects in the CANDELS photometric catalogs. We test the quality of the redshift PDFs from each group using the Quantile-Quantile plot (or Q-Q plot, for short), which is a graphical method helpful for testing whether a set of data is drawn from a certain distribution or not. In this case, we use Q-Q statistics to assess whether a set of spectroscopic redshifts is consistent with being drawn from the corresponding objects' photo-z PDFs. We also test the performance of each group's PDFs using point estimates (in particular the weighted mean redshift z_{weight}).

In a previous test of photometric redshift methods for CANDELS, Dahlen et al. (2013) found that the codes considered fell well short of the ideal in producing error estimates that match reality (e.g., 68% of spectroscopic redshifts should fall in the 68% credible region of the corresponding object's photo-z PDF). In order to approach this ideal more closely, we optimize the photometric redshift PDFs from each group using Q-Q statistics to derive PDF shifts and exponents to correct for overconfident or underconfident probability distributions, applying the methods described in Kodra & Newman 2019 (in prep.).

Dahlen et al. (2013) also found that when point statistics or PDFs from different codes were combined, they could perform better than any individual input. In this paper, we test this for a broader set of combination methods. In addition to the Hierarchical Bayesian method applied by Dahlen et al. (2013), we introduce a new method of combining photo-z PDFs, the minimum Fréchet distance curve. This method chooses the input PDF which is closest to the other members of the ensemble in aggregate to represent the set; it is the function analog to the median of a set of numbers. This method guarantees that if the input PDFs are well-calibrated to meet the statistical definition of a probability density function, the minimum Fréchet distance curve will be as well.

This paper is organized as follows. In Section 3.2 we describe the datasets employed in this paper. Section 3.3 describes the methods used to improve probability density functions from individual photo-z codes and assesses the results with a variety of summary statistics.

Section 3.4 describes and presents tests of several methods of combining PDF results from multiple photo-z codes. Section 3.5 describes the photometric redshift catalogs being publicly released in conjunction with this paper. Finally, in Section 3.6 we summarize the major findings of this study and provide guidance on the use of these catalogs.

3.2 DATA

Throughout this study we use data from the Cosmic Assembly Near-Infrared Deep Extragalactic Legacy Survey (CANDELS) collaboration, which obtained *Hubble* Space Telescope observations covering five fields spanning a total area of $\sim 800\,\mathrm{arcmin}^2$ on the sky. The CANDELS survey is separated into a Deep and a Wide portion, with areas of $\sim 125\,\mathrm{arcmin}^2$ and $\sim 675\,\mathrm{arcmin}^2$, respectively. The deep portion has a depth of ~ 10 orbits per pointing and includes data in two of the CANDELS fields, GOODS-North and GOODS-South. The wide portion consists of ~ 2 orbits per pointing, and includes buffer regions around the deep fields as well as three additional fields, COSMOS, the Extended Groth Strip (EGS), and the UKIDSS Ultra-Deep Survey field (UDS). Detailed descriptions of the sky coverage and observation strategy of **CANDELS** can be found in Grogin et al. (2011) and Koekemoer et al. (2011). Catalogs based upon matched-model TFIT photometry which incorporates imaging from *Hubble* along with other datasets in each individual CANDELS field are presented in Nayyeri et al. (2017), Stefanon et al. (2017), Guo et al. (2013), Galametz et al. (2013), and Barro et al. 2019 (in prep.).

3.2.1 Photometric Redshifts

Six groups within the CANDELS collaboration have used the TFIT photometric catalogs to estimate probability distribution functions (**PDFs**) for the photometric redshift (photo-z) for each galaxy observed by CANDELS. The codes used are all based upon fitting a set of spectral energy distribution templates to a galaxy as a function of its possible redshift and determining χ^2 between the template prediction and the photometry as a function of template parameters and redshift. Most codes utilize a prior for the redshift based upon the apparent magnitude of an object, combined with a likelihood proportional to $\exp(-\chi^2/2)$, to calculate a posterior PDF for the redshift. Each group within CANDELS has used a different code for calculation, a different template database, and/or different options when using a code (e.g., different priors) to obtain photo-z PDFs.

EAZY (Brammer et al. (2008))[1]

This code fits a linear combination of stellar population templates to the observed U-to-8μm SEDs, minimizing the χ^2 statistic. Fitting is done in flux space, and the template set has been composed to span the range of observed galaxy colors. EAZY allows the use of a template error function that down-weights data points at rest-frame wavelengths $\geq 2\mu$m and a "luminosity-function × volume" prior which assigns a reduced probability to low redshifts and to objects with bright apparent magnitudes at high redshift. In addition, an iterative application of photometric zero point offsets improves the match to the available spectroscopic redshifts. This code was run by two separate groups, led by Steven Finkelstein and Stijn Wuyts, respectively. The two groups differed in their choices of parameters and priors.

zphot

This code was first introduced in Giallongo et al. (1998), more extensively described in Fontana et al. (2000), and then used in a number of papers thereafter. It minimizes χ^2 statistics when fitting template **SEDs**, accepts both fluxes and magnitudes as input, with a rigorous treatment of undetections in the latter case. It allows the user to adopt a wide

[1] http://www.astro.yale.edu/eazy/

variety of templates, including synthetic models taken from BC03, Maraston (2005), and Fioc & Rocca-Volmerange (1997). Experiments with these different models have shown that model that yield more accurate photo-z is the Fioc & Rocca-Volmerange (1997) one, at least with the set of parameters exploited so far. These models are therefore usd in this paper. Additionally, it has the capability to choose from a variety of SFH, as well as observed templates of stars, galaxies, and **AGN**s, while it can also include dust extinction such as Calzetti et al. (2000), and intergalactic medium absorption such as Fan et al. (2006). The code is a full SED analyzer, and returns the best fitting values of redshift (i.e. the photometric one) as well as other parameters of the galaxy template (like age, stellar mass, dust extinction, and so on). For the latter use the redshift can be fixed to a value given by the user, or left free to vary. It can also estimate errors from the probability distribution of the various parameters. The code allows for a minimum photometric error to be set in each photometry band, and optionally applies a cut to avoid unrealistically negative fluxes: this is done in flux space when the flux is below 0 in excess of a given number of times the flux error. Finally, to avoid contamination from non-stellar emission it excludes **IRAC** bands that probe rest frame $> 5\mu m$ (ch3 at $z \leq 0.15$ and ch4 at $z \leq 0.6$). This code was run by the group led by Adriano Fontana.

HyperZ (Bolzonella et al. (2000))[2]

This is another code that minimizes χ^2 statistics, with fitting carried out in flux space, excluding negative fluxes. Shifts to magnitudes can be added manually. In order to avoid unrealistic solutions a prior for the **F160W** band absolute magnitude in the range $-30 < M < -9$ (Vega mag) was applied. Reddening was included in the form of the Calzetti et al. (2000) reddening law varying A_V from 0 to 3 in steps of 0.2. SED templates are based on the Maraston (2005) models with exponentially declining SFHs with e-folding times $\tau = 0.1$, 0.3, 1 and 2 Gyr, and for each of these in four metallicity flavors: 1/5 solar metallicity, 1/2 solar metallicity, solar metallicity, and twice solar metallicity. The allowed age range of the templates was restricted between 0.1 Gyr and the age of the Universe at the given redshift. Finally, the option for a minimum photometric error was set to 0.05 magnitudes. This recipe

[2] http://webast.ast.obs-mip.fr/hyperz/

was introduced in Pforr et al. (2013). This code was run by the group led by Janine Pforr.

LePhare (Arnouts & Ilbert (2011))[3]

This is yet another code that minimizes χ^2 statistics when fitting template SEDs, which can be done both in magnitude and in flux space. It has the capability to use luminosity priors, add extra photometry errors, use a training sample to optimize the template SEDs and to derive zero-point offsets, while contribution for emission can be included (Ilbert et al. (2006, 2009)). The output provides photometric redshifts from χ^2 minimization or marginalization. The median values from the photo-z $PDFs$ are also provided. Here, a prior on the optical absolute magnitude in the range $-24 < M < -8$ is used, while IRAC ch3 and ch4 are excluded. This code was run by the group led by Mara Salvato.

WikZ (Wiklind et al. (2008))

This is again a code that minimizes χ^2 statistics when fitting SED templates to observed photometry, which is done in flux space. The code can be run with two different parametrized star formation histories (SFHs), exponentially declining SFR or so called delayed-τ SFR. In the current application, the code was run using a delayed-τ SFH. Negative fluxes are not completely excluded, but they add up to the χ^2 when the template flux is brighter than the 1σ upper limit, while they are excluded when the flux is lower than the 1σ upper limit. Additionally, it excludes IRAC ch3 and ch4 for $z < 0.5$ and $z < 0.7$ respectively. Finally, it has the capability to add extra smoothing errors to the photometric ones. This code was run by the group led by Tommy Wiklind.

3.2.2 Spectroscopic And 3D-HST Grism Redshifts

Spectroscopic redshift measurements can be used both to train photometric redshift codes – e.g., to identify zero point offsets in photometry that, if removed, will improve fits – as well as for testing results. The training set of spectroscopic redshifts (spec-z's) used by the six groups producing CANDELS photo-z PDFs consists of 5807 high quality spec-z's spanning

[3] http://www.cfht.hawaii.edu/~arnouts/LEPHARE/lephare.html

all five CANDELS fields. We have used the same set of redshifts in order to estimate the parameters used to recalibrate the PDFs from each code, as described in Section 3.3.

Our primary testing set consists of 4089 high quality spectroscopic redshifts, again drawn from a variety of sources, that are completely independent of those included in the training set and with any overlapping objects removed. We use this set in order to assess the performance of point statistics (e.g., the redshift of maximum probability) and to test the quality of photo-z PDFs from each code, both before and after the optimization procedure described below.

In addition to the testing set of spectroscopic redshifts, we also use 3D−HST grism redshifts (grism-z's) to test point estimates and photo-z PDF quality (Momcheva et al. (2016)). This set consists of 3367 highest-quality redshifts spanning all of the CANDELS fields. We include only those objects with redshifts larger than $z_{\text{grism}} > 0.6$. Figure 3.1 and Figure 3.2 show the distribution of objects in each spectroscopic sample in a plot of H-band magnitude versus redshift, as well as their respective histograms, for each of the five CANDELS fields separately. Additionally, Table 1 gives details regarding the construction of these three sets, including references to the original catalogs and the specific cuts applied to select only the highest-quality, most secure redshifts.

3.3 OPTIMIZATION METHODS

The photo-z PDFs produced by the different groups do not behave in detail as true probability distributions that meet the statistical definition of a PDF would be expected to. Given that the set of high-confidence spectroscopic redshifts we employ in this paper should be incorrect only rarely ($< 5\%$ of the time), we can treat them to first order as representing the true redshifts of the objects observed. In that case, we would expect 68% and 95% of them to fall in the 68% and 95% credible intervals of the photo-z PDFs, respectively, if the statistical definition of a PDF is being obeyed.

In an earlier test with CANDELS data, Dahlen et al. (2013) found that while some codes had better results than others, none of them performed well when the coverage of credible intervals was assessed in this way. This indicates that the photo-z PDFs must have

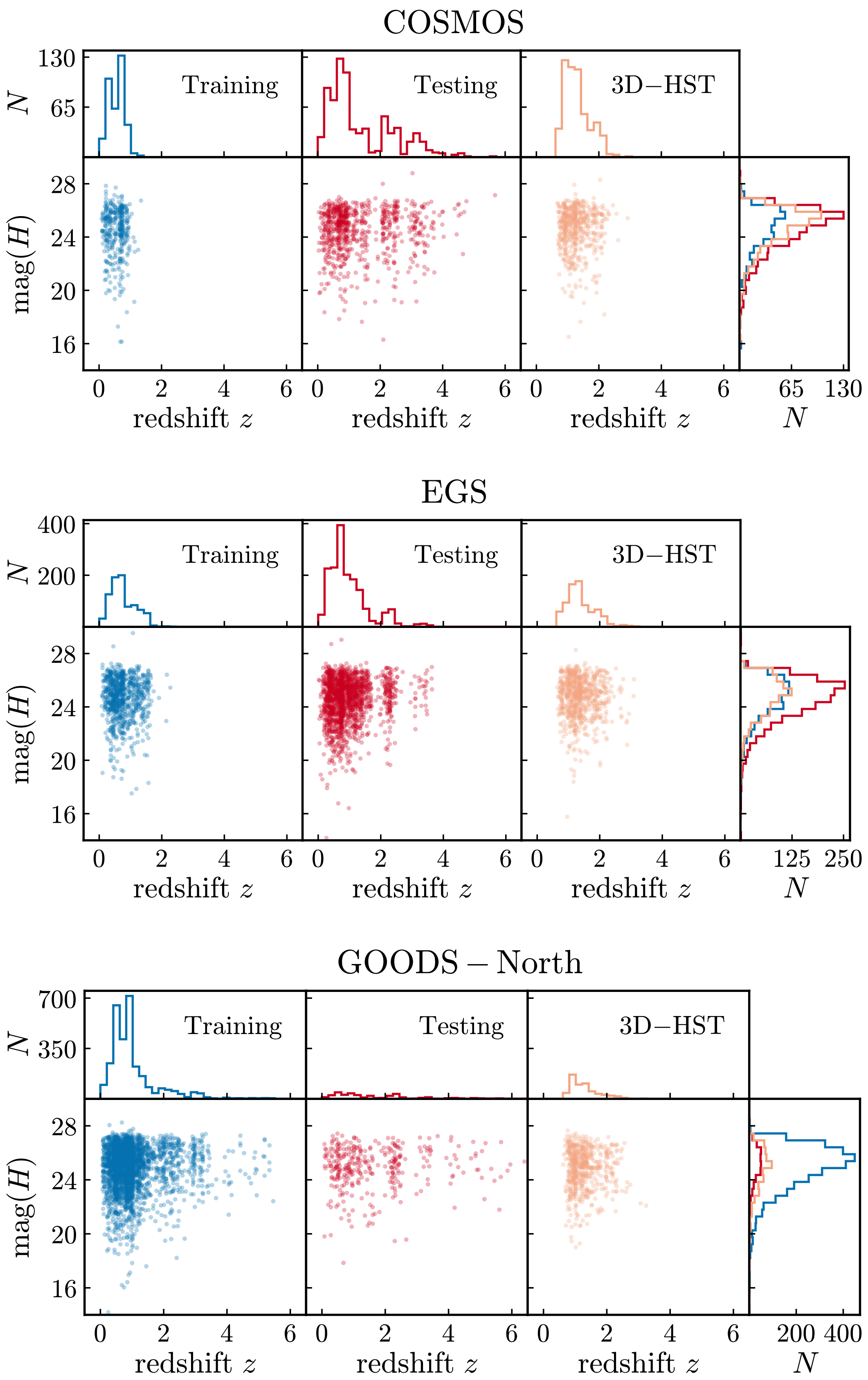


Figure 3.1 Continued in next page

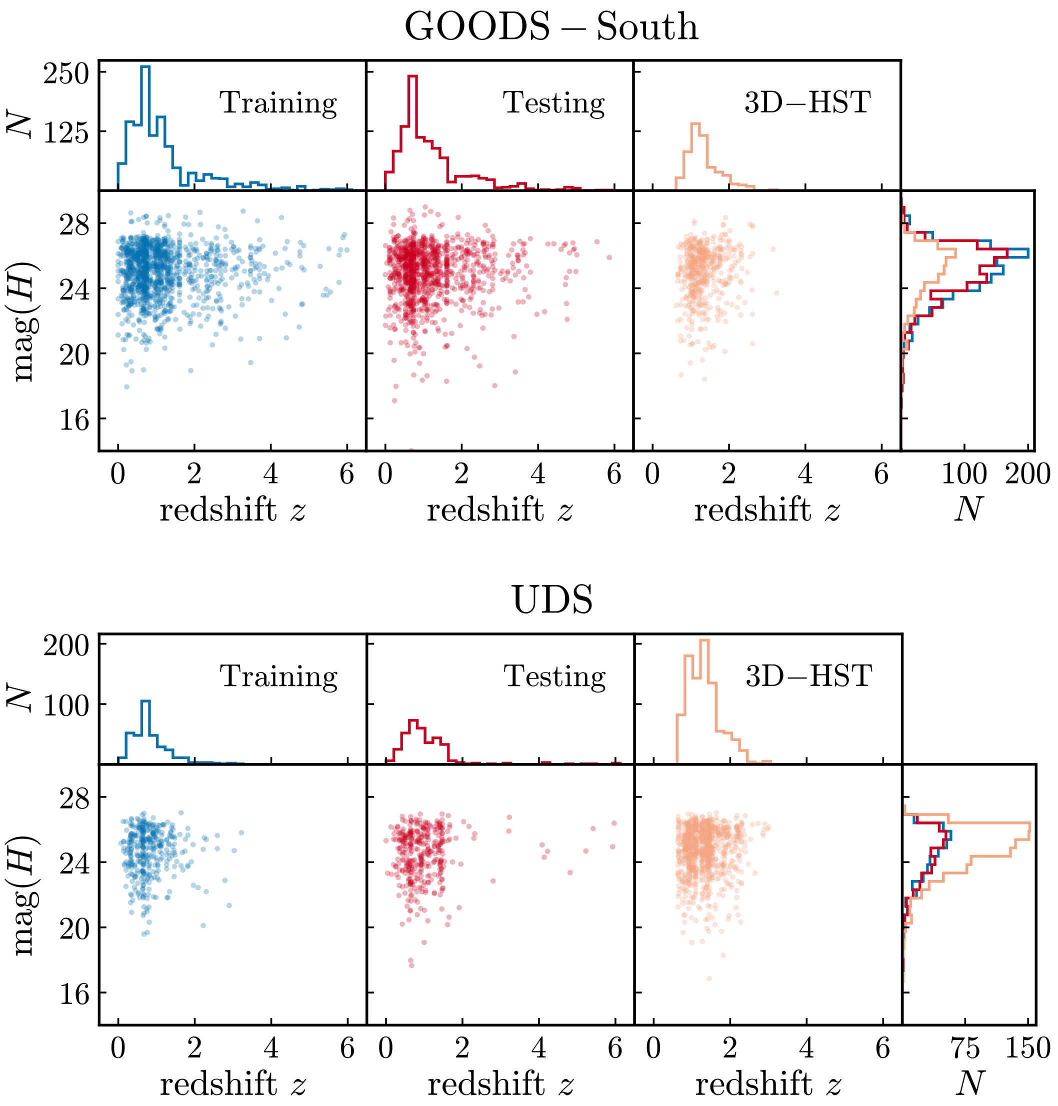


Figure 3.2 Redshift-magnitude plots for the spectroscopic redshifts (divided into separate training and testing sets) as well as the 3D-HST grism redshifts used in this paper, for all five CANDELS fields. The redshift and magnitude ranges of the testing and 3D-HST datasets differ strongly from the training set; as a result, they provide highly independent assessments of the quality of photo-z PDFs which are optimized based upon the training set.

substantial, qualitative problems. Simple examples of issues would be having a bias due to template mismatch, such that the PDFs are shifted to higher or lower redshifts than they should be, or inaccuracies in photometric error models which cause PDFs to be too wide or too narrow.

Following Kodra & Newman 2018 (in prep.), we use statistics based upon the Quantile-Quantile (Q-Q) plot to recalibrate the photo-z PDFs produced by each group; this should improve agreement with the statistical definition while simultaneously yielding more accurate point estimates and error estimates. Specifically, we use the training set of spectroscopic redshifts in order to calculate the distribution of observed quantiles, i.e. the values of the cumulative distribution function (CDF) values corresponding to each redshift. If PDFs meet the statistical definition, the distribution of these quantiles should be uniform between 0 and 1. The Q-Q plot for a particular PDF and set of spec-z's will show, as a function of theoretical quantiles (distributed between 0 and 1) the value that corresponds to that quantile of the set of CDFs evaluated at the spectroscopic redshift. E.g., if the 50th percentile value in the set of CDFs evaluated at the spec-z's is 0.647, (0.5, 0.647) will be a point on the Q-Q plot, as it is shown by the blue dot in Figure 3.3. For the ideal case where the CDF values are uniformly distributed, the Q-Q plot will correspond to the unit line between (0,0) and (1,1).

In Kodra & Newman 2018 (in prep.), we define several statistics based upon the Q-Q plot that can be used to assess how close we are to the ideal case; in this paper, we employ two of them. The first of these is the normalized ℓ^2-norm, which corresponds to the average distance in the y direction between the Q-Q curve for a given dataset and the unit line. This distance will be zero in the ideal case. Additionally, we determine the fraction of objects that have CDF values below 0.0001 or above 0.9999. We exclude such objects from the construction of the Q-Q plot, since their spec-z's fall outside of the region of their corresponding photo-z PDFs which has significant probability; this makes our analysis considerably more robust to the possibility of erroneous spectroscopic redshifts. We label the fraction of objects with PDF values outside this range f_{op} (for "fraction outside of PDF"). As an example we construct the Q-Q plot shown in Figure 3.3 using the training set of spectroscopic redshifts and the photo-z PDFs of Finkelstein. The ℓ^2-norm shown in the figure is normalized by dividing with the square root of the number of objects used for the construction of the Q-Q plot, after removing the objects that qualify as $\mathrm{f_{op}}$. The pink vertical line shows the difference between the observed and expected quantile for the point corresponding to the 50th percentile, or 0.5 quantile.

We note that when we calculate Q-Q statistics for optimization, we exclude objects with redshifts $z_{\text{spec}} \leq 0.3$. This is because the photo-z PDFs from the different groups show divergent behavior at low redshifts, with little correlation to performance at higher z. In order to see this, we separate the objects from each field into magnitude bins of size 0.5mag, and then add the individual redshift PDFs of all objects in each bin. We finally normalize by dividing with the number of objects in each bin. The resulting quantity corresponds to the expectation value for the redshift distribution of objects in a given bin.

As an example, the set of summed PDF curves for EGS objects in the magnitude bin centered at $H = 25$ predicted by each group, both before and after the optimization procedures which we describe below have been applied, are shown in Figure 3.4. Even after optimization, the predicted number of objects at low redshifts varies greatly from group to group; results in this regime may have limited reliability. In Figure 3.5 we show a set of images constructed from curves such as those shown in Figure 3.4. In these images, the $x-$axis position corresponds to the middle of the magnitude bin used to construct summed PDF curves, and the y coordinate corresponds to the redshift, z. The intensity of the color scale indicates the value of the summed PDF at a given magnitude and redshift (corresponding to the y values in Figure 3.5). Although all of the codes used to produce the photo-z PDFs yield good results when evaluated with standard point statistics for objects with spectroscopic redshifts(as described below), and all are using the same photometry as inputs, the predicted redshift distributions as a function of magnitude predicted by each group differ in many details, even after optimization which makes individual photo-z PDFs better obey the statistical definition, as we demonstrate below.

Using the Q-Q statistics defined above, we can identify any aggregate bias in the photo-z PDFs of a given group by applying negative or positive shifts in the redshift direction (dz) to the PDFs. Our sign convention is such that negative dz values correspond to shifts to the left on a plot of PDF(z), such that the PDFs will peak at lower values of redshifts, while positive dz values correspond to shifts to the right, so that PDFs peak at higher redshift values

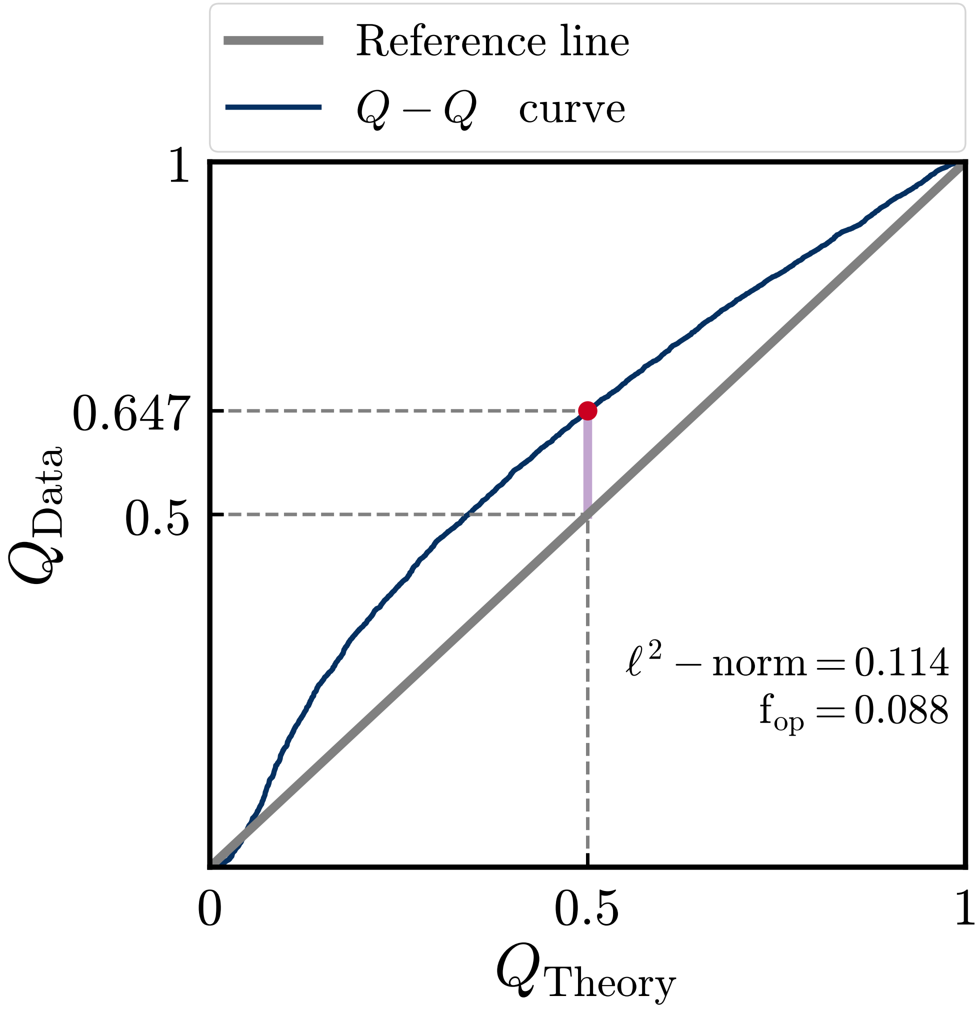


Figure 3.3 Quantile-Quantile (Q-Q) plot constructed using the training set of spec-z's and the photo-z PDFs of Finkelstein. The Q-Q plot shows the CDF value evaluated at the spectroscopic redshift of an object, $\rm Q_{data}$ corresponding to a given quantile of the set of CDF values, $\rm Q_{theory}$. For instance, if the 50th percentile in the set of CDF($z_{\rm spec}$) values is 0.647, then (0.5, 0.647) is a point along the Q-Q curve, as is shown by the blue dot. If photo-z PDFs obey the statistical definition of a probability density function, Q-Q curves will lie along the reference line from (0,0) to (1,1). Their deviation is quantified by the sum of the vertical distance at each point (ℓ^2-norm), such as the one shown by the pink vertical line for the case of the (0.5, 0.647) point. The ℓ^2-norm presented here is normalized by dividing with the square root of the number of points. Additionally, the $\rm f_{op}$ gives the fraction of objects with spec-z's that fall outside of the main region of the corresponding PDFs.

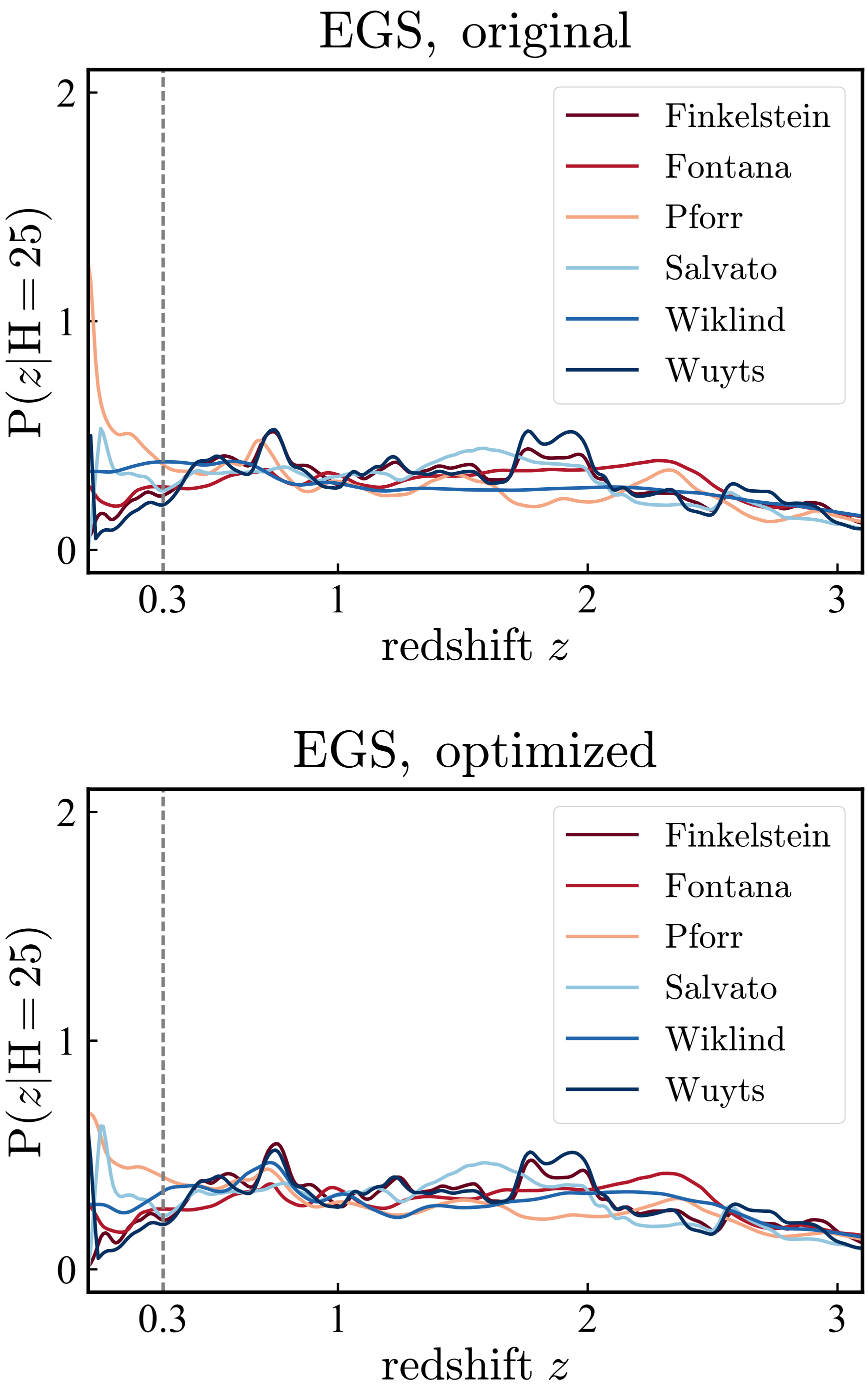
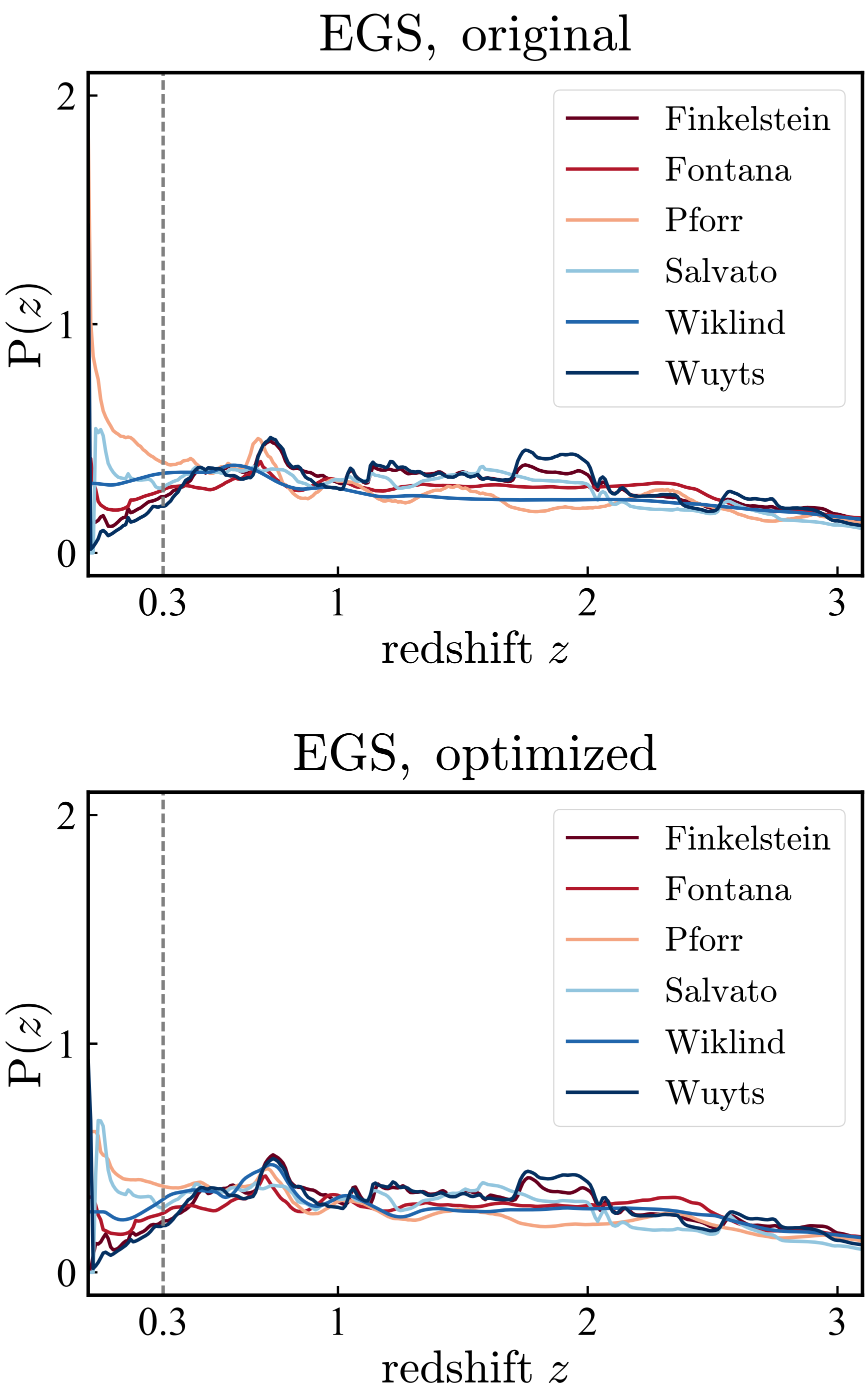


Figure 3.4 Sum of the photo-z PDFs for all objects in a magnitude bin of width 0.5 centered at $\mathrm{mag}(H) = 25$ in the EGS field. The curves correspond to the expectation value for the redshift distribution in this bin. Objects with missing PDFs for any of the groups are excluded from the sums to ensure that all curves are equivalent. The disagreement between groups is greatest at low redshifts (below $z = 0.3$, indicated by the grey dashed line). Because of this divergence, we exclude objects with $z \leq 0.3$ from the calculation of Q-Q statistics. Although all codes deliver good performance when PDF peak redshifts are compared to spectroscopic zs, the aggregate predictions for broad galaxy samples differ significantly from code to code.


EGS, original

EGS, optimized

Figure 3.5 Summed PDFs as a function of magnitude and redshift, before and after optimization procedures are applied, for the EGS field. The color scale of the image indicates the summed and renormalized probability that an object in a particular magnitude bin is at the selected redshift, equivalent to the y axis in Figure 3.4. Objects with missing PDFs for any of the groups are excluded from the sums. The shape and locations of the bright regions (which should correspond to redshifts where there is an excess of objects due to sample/cosmic variance), as well as other detailed features, differ significantly from group to group, even though all groups used identical photometry for the same set of objects to calculate the PDFs used here.

than they originally did. We apply shifts corresponding to an array of 151 equally-spaced values over the interval $[-0.5, 1.]$, and construct the Q-Q curve in each case after removing objects with spec-z's outside their PDFs. We then identify the shift value that minimizes the normalized ℓ^2-norm; that is, the value which yields a Q-Q curve as close as possible to the ideal unit line. To do this, we interpolate the $\ell^2-\mathrm{norm}$ values using a quadratic univariate spline (using the `scipy.interpolate.UnivariateSpline` routine with parameters k = 2 and s = 1.) The goal of this interpolation is for greater accuracy and smoother results.

We present the dependence of the ℓ^2-norm on dz for each group and CANDELS field, separately in Figure 3.6. Results from some groups exhibit larger biases than others, with the bias varying from field to field; this may be due to the differences in the imaging bands included in each field. With only two exceptions (Pforr results for the $\mathrm{GOODS-North}$ and GOODS-South fields) the shifts are all positive. That indicates that for almost all fields and groups, the photo-z PDFs peak at lower values than they should, which in turn results in point estimates (such as the redshift of the PDF peak) which are biased low. Table 3.1 gives the optimal shift values that we find should be applied to the photo-z PDFs from each group and each field to yield better performance in the Q-Q plot. We will assess the level of improvement achieved in both Q-Q statistics and point estimates in sections to follow, using the independent testing set of spectroscopic redshifts as well as 3D-HST.

We continue the recalibration procedure by improving the accuracy of PDF widths. We do this by raising a PDF to a power of $\alpha = 1/\gamma$ and then renormalizing to have integral one. For a Gaussian PDF, this is equivalent to multiplying the σ parameter by $\gamma^{1/2}$, and so provides a way to correct for either overestimates or underestimates of errors; however, the procedure we use can be applied to any PDF whether it is Gaussian or not.

We determine optimal values of γ after we apply the optimized shifts dz that were determined by the preceding analysis. We consider values in the range $0 < \gamma < 1$ as well as in the range $\gamma \geq 1$, corresponding to $\alpha > 1$ and $0 < \alpha \leq 1$, respectively. The former range corresponds to making photo-z PDFs sharper, since high peaks of the PDF become higher while low PDF values in the tails or valleys between peaks become smaller. Conversely, $\gamma \geq 1$ leads to broader photo-z PDFs, since the contrast between the highest peaks and low values of the PDF is reduced.

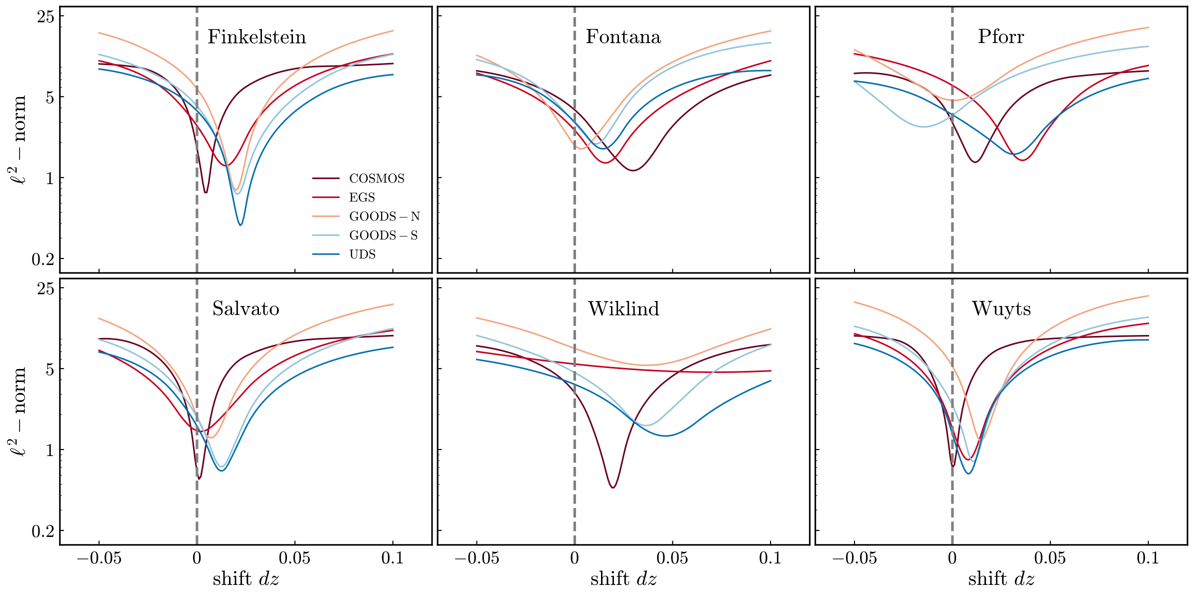


Figure 3.6 Plots of the normalized ℓ^2-norm as a function of the shift dz applied to the photo-z **PDFs** for each of the five **CANDELS** fields, calculated using the training set of spectroscopic redshifts. The normalized ℓ^2-norm will have a smaller value the more closely PDFs obey the statistical definition of a probability distribution; as a result, we wish to minimize this quantity. The vertical grey dashed line indicates a zero shift, indicating that the calibration of the PDFs from a given group were accurate. A different shift seems to be needed for each group and each field, without any clear pattern.

Table 3.1 The optimal values of shifts dz for the PDFs from each group for each field, determined by identifying the minimima in the curves shown in Figure 3.6. Most of the shifts are positive, indicating that PDF(z) curves need to be shifted to the right. Results from some groups exhibit larger biases than others, with Salvato and Wuyts needing the smallest shifts and Wiklind the largest, especially in the EGS field.

Field	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts
COSMOS	0.005	0.030	0.012	0.001	0.020	0.000
EGS	0.015	0.016	0.036	0.002	0.071	0.008
GOODS – N	0.020	0.003	0.001	0.007	0.036	0.015
GOODS – S	0.021	0.011	-0.015	0.013	0.036	0.011
UDS	0.022	0.014	0.031	0.013	0.046	0.008

For each group's PDFs in a particular field we search a coarse array of values for the parameter γ covering the interval $[0.05, 7]$, and after constructing the Q-Q plot in each case, we evaluate the normalized ℓ^2-norm for each value of γ. We then use a finer grid of values of γ focused around the coarse-grid value that minimizes the ℓ^2-norm to obtain improved precision in the optimal value. We present plots of the dependence of the ℓ^2-norm on γ for each group and CANDELS field separately in Figure 3.7.

Optimal values of the parameter γ that are close to unity imply that little change is needed in the width of the PDFs; this is the case for some groups in some fields, but is not universal. Conversely, when the optimal values differ significantly from unity, then large differences in the widths of the PDFs are implied (corresponding to significantly overconfident or underconfident error models). Table 3.2 lists the optimal γ values that we find should be applied to the raw photo-z PDFs from each group for each field to bring Q-Q curves closer to the ideal, and hence making the estimated photo-z PDFs hew more closely to the statistical definition.

3.3.1 Evaluation using Q-Q Statistics

After identifying the optimal values of the dz and γ parameters using the training set of spectroscopic redshifts and their corresponding photo-z PDFs, we can apply the corresponding transformations to all the PDFs from a particular group and field. We can then construct the Q-Q plots and evaluate the overall normalized ℓ^2-norm and f_{op} values for each group evaluated with the training data (which should have results that are biased low), as well as with the independent testing set of spectroscopic redshifts and with the 3D-HST grism redshifts. As for some fields relatively few testing redshifts are available, we combine the objects from all five CANDELS fields to make the Q-Q plots. We present the Q-Q plots for each group in Figure 3.8. In each panel, dashed curves correspond to the Q-Q plots for each spectroscopic dataset before applying shift and power transformations, and solid curves show the Q-Q plots after transformations. It is clear that all curves follow the diagonal reference line more closely after optimization. We plot the normalized ℓ^2-norm and f_{op} values before and after optimization of the PDFs in Figure 3.9. Although in a few cases the f_{op} value becomes slightly larger after optimization, gains in the normalized ℓ^2-norm are substantial and ubiquitous. As should be expected, the optimized ℓ^2-norm values are lowest for the training set, as by construction our optimization chooses the transformation parameter values which make that quantity as small as feasible. However, it is very encouraging that substantial improvements are also found for the testing and 3D-HST datasets, which have very different coverage in magnitude-redshift space from the training set.

3.3.2 Evaluation Using Point Statistics

Although our primary focus is on producing accurate photometric redshift probability distributions, we also can evaluate the performance of point (or summary) statistics for the photometric redshifts of CANDELS objects before and after optimization. We focus on two different point statistics which are constructed from the photo-z PDFs: the most probable redshift (which we label z_{peak}, as it corresponds to the redshift where the highest peak of the photo-z PDF occurs), and the probability-weighted expectation value of the redshift (z_{weight}). Following Dahlen et al. (2013), we compute this quantity using only the region

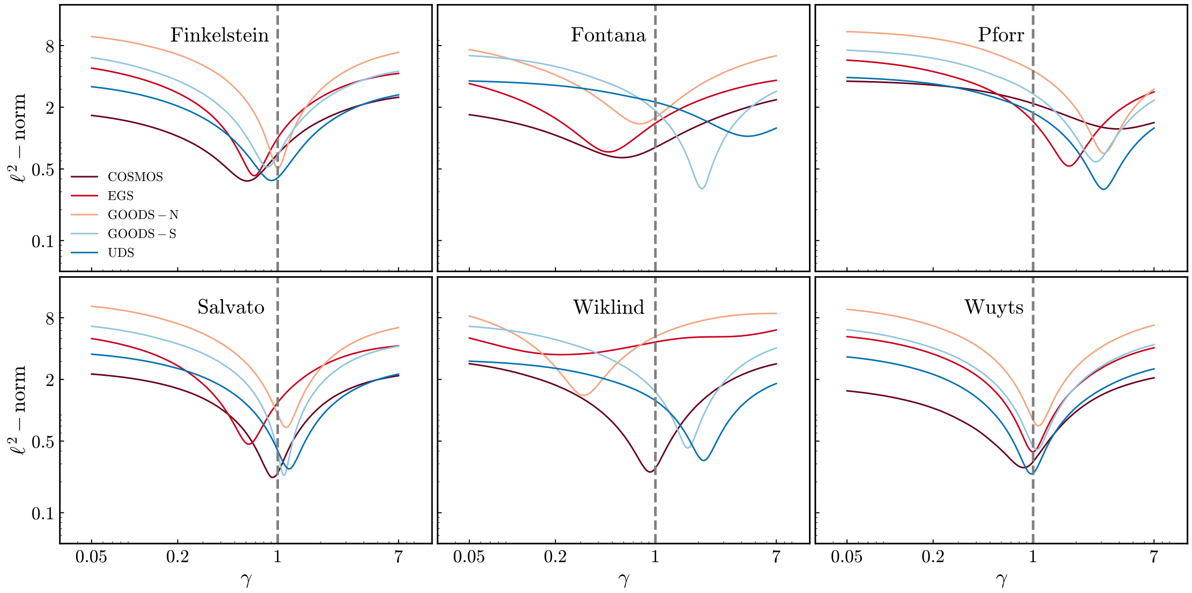


Figure 3.7 Plots of the normalized ℓ^2-norm as a function of the parameter γ which we use to improve the widths of the photo-z PDFs. PDFs are raised to the power $1/\gamma$ in this analysis, corresponding to altering the standard deviation parameter of a Gaussian distribution by a factor of $\sqrt{\gamma}$. These plots were constructed using only the training set of spectroscopic redshifts in each of the five **CANDELS** fields, for each group's PDFs independently, after applying the shifts in redshift listed in Table 3.1. The vertical grey dashed line indicates where $\gamma = 1$, corresponding to the case where no changes are applied to PDFs. Different values of the parameter γ are needed for each group's PDFs in each field, without any clear pattern.

Table 3.2 The optimal values of the PDF rescaling parameter γ (used to improve the widths of the PDFs) for each group and each field. While some groups' results yield γ values that are close to unity, others have significantly lower or significantly larger values, implying substantial changes to the widths of their photo-z PDFs.

Field	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts
COSMOS	0.611	0.582	3.959	0.922	0.916	0.859
EGS	0.686	0.470	1.784	0.628	0.220	0.996
GOODS – N	1.003	0.777	3.118	1.145	0.315	1.090
GOODS – S	0.848	2.115	2.750	1.106	1.683	1.060
UDS	0.905	4.418	3.114	1.203	2.174	0.973

surrounding the highest peak of the PDF (specifically, the redshift range within that peak where the PDF value remains above $0.05 \times \mathrm{PDF}(z_{\mathrm{peak}})$). To determine this range, we first linearly interpolate the PDF onto a grid of 64001 redshifts in place of the original 1001, then use cubic spline interpolation on this finer grid to find the redshifts around z_{peak} where the PDF is equal to $0.05 \times \mathrm{PDF}(z_{\mathrm{peak}})$. If the PDF values are never lower than $0.05 \times \mathrm{PDF}(z_{\mathrm{peak}})$, then the bounds used to calculate z_{weight} are the same as the bounds of the grid on which the PDFs are defined. In general the Dahlen et al. (2013) definition of z_{weight} yields better results than the overall expectation value of the redshift evaluated over the full PDF, as in cases where PDFs have multiple peaks the overall expectation value will often lie between peaks in regions of negligible probability.

We use two quantities to evaluate the accuracy of these point statistics before and after optimization. The first is the normalized median absolute deviation (σ_{NMAD}) of the differences between photo-z point estimates and spectroscopic redshifts for the corresponding objects, defined by:

$$\sigma_{\mathrm{NMAD}} = 1.48 \times \mathit{median}\left(\frac{|\Delta z|}{1 + z_{\mathrm{spec}}}\right), \tag{3.1}$$

where $\Delta z = z_{\mathrm{phot}} - z_{\mathrm{spec}}$ is the difference between a point estimate of the photometric redshift

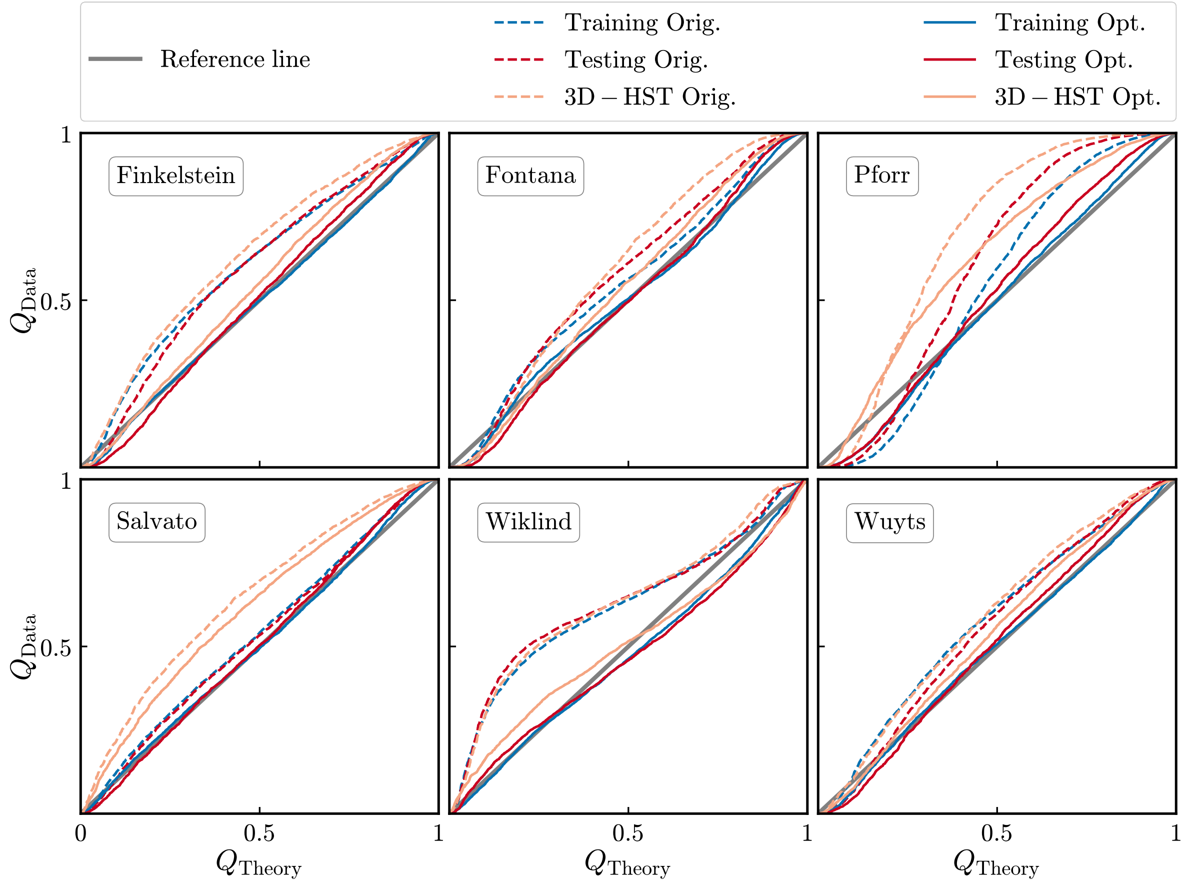


Figure 3.8 Quantile-Quantile (Q-Q) plots for photo-z PDFs both before (“original”, dashed curves) and after (“optimized”, solid curves) optimized transformations have been applied. The Q-Q plot shows the CDF value evaluated at the spectroscopic redshift of an object, Q_{data} corresponding to a given quantile of the set of CDF values, Q_{theory}. For instance, if the 30th percentile in the set of CDF(z_{spec}) values is 0.21, (0.3, 0.21) would be a point along the Q-Q curve. If photo-z PDFs obey the statistical definition of a probability density function, Q-Q curves will lie along the unit line from (0,0) to (1,1). We evaluate the results from each group separately using the independent training and testing sets of spectroscopic redshifts as well as the 3D-HST grism redshifts; results from all five CANDELS fields are combined together here. It is clear that our optimization methods improve the results for each group and for every set of redshifts. The normalized ℓ^2-norm used to evaluate PDF accuracy corresponds to the RMS deviation between a Q-Q curve and the diagonal in the y direction in these plots.

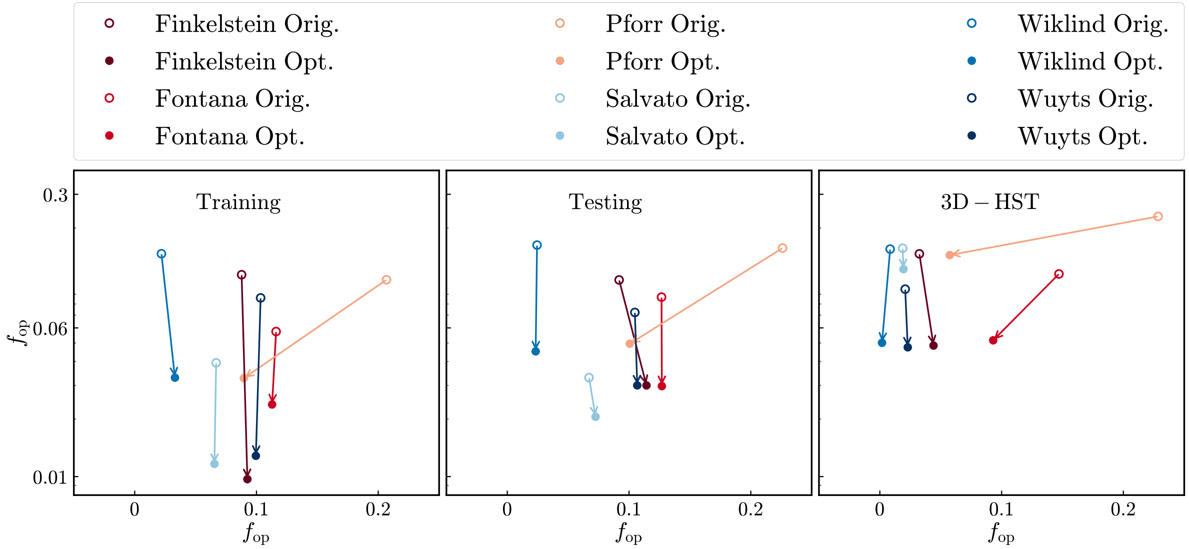


Figure 3.9 Plots of the normalized ℓ^2-norm versus the out-of-PDF fraction $f_{\rm op}$ for the PDFs from each group both before (squares at arrow bases) and after (circles at arrow heads) optimization. The three panels show these statistics evaluated using the training set of spectroscopic redshift (left); the testing set (middle); and 3D-HST grism redshifts (right). The ℓ^2-norms are reduced in all cases; in most cases, the $f_{\rm op}$ values are lower as well, and where they increase it is only by a small amount. No single group produced superior results in all cases.

for an object and its spectroscopic redshift. The normalizing factor of 1.48 in the definition of $\sigma_{\rm NMAD}$ causes the expectation value of the NMAD to equal the standard deviation for data that is drawn from a normal distribution.

In addition to the NMAD, we track the fraction of catastrophic outliers in a particular group's results, $f_{\rm co}$, as another useful statistic for characterizing photo-z quality from each group. We define catastrophic outliers as those objects which fulfill the condition:

$$\frac{|\Delta z|}{1 + z_{\rm spec}} > 0.15. \tag{3.2}$$

Here the limit 0.15 is somewhat arbitrarily chosen; we will show below that typical photometric redshift errors for CANDELS objects with spectroscopic redshifts are $\sim 0.03(1+z)$, so the threshold of 0.15 approximately corresponds to 5σ outliers.

We find that the NMAD is smaller when computed using $z_{\rm weight}$ rather than $z_{\rm peak}$; that is, $z_{\rm weight}$ provides a superior estimate of the redshift of an object. Therefore, we only show results for statistics computed using $z_{\rm weight}$ in the remainder of this paper, although we still include $z_{\rm peak}$ in our final photo-z catalogs.

Figure 3.10 shows the scatter plots of $z_{\rm weight}$ vs $z_{\rm spec}$ for all six participants, using only objects belonging to the training set of spec-z's. The $z_{\rm weight}$ values are calculated using the optimized version of the photo-z PDFs. In the same plots we present the $\sigma_{\rm NMAD}$, the percentages (and number of objects in parentheses) of catastrophic outliers, as well as the percentage (and number of objects in parentheses) for the objects that have missing photo-z PDFs and therefore missing $z_{\rm weight}$ values. Below each scatter plot, we also present a $z_{\rm spec}$ dependence of $\Delta z/1 + z_{\rm spec}$, where $\Delta z = z_{\rm weight} - z_{\rm spec}$. In the same plot we also present the redshift dependence of $\sigma_{\rm NMAD} = 1.48 \times \Delta z/1 + z_{\rm spec}$ (dashed gray line), and the Hodges-Lehmann mean of $\Delta z/1 + z_{\rm spec}$ (solid gray line), where bins of 0.5 have been used along $z_{\rm spec}$ for their construction. They generally do not seem to vary significantly with redshift, apart from the spikes observed at $z_{\rm spec} \sim 5.5$, where the number of objects becomes very low to yield satisfying results.

Figure 3.11 plots the $\sigma_{\rm NMAD}$ values and catastrophic outlier fractions for the photometric redshifts from each group, calculated using the PDFs both before and after optimization has

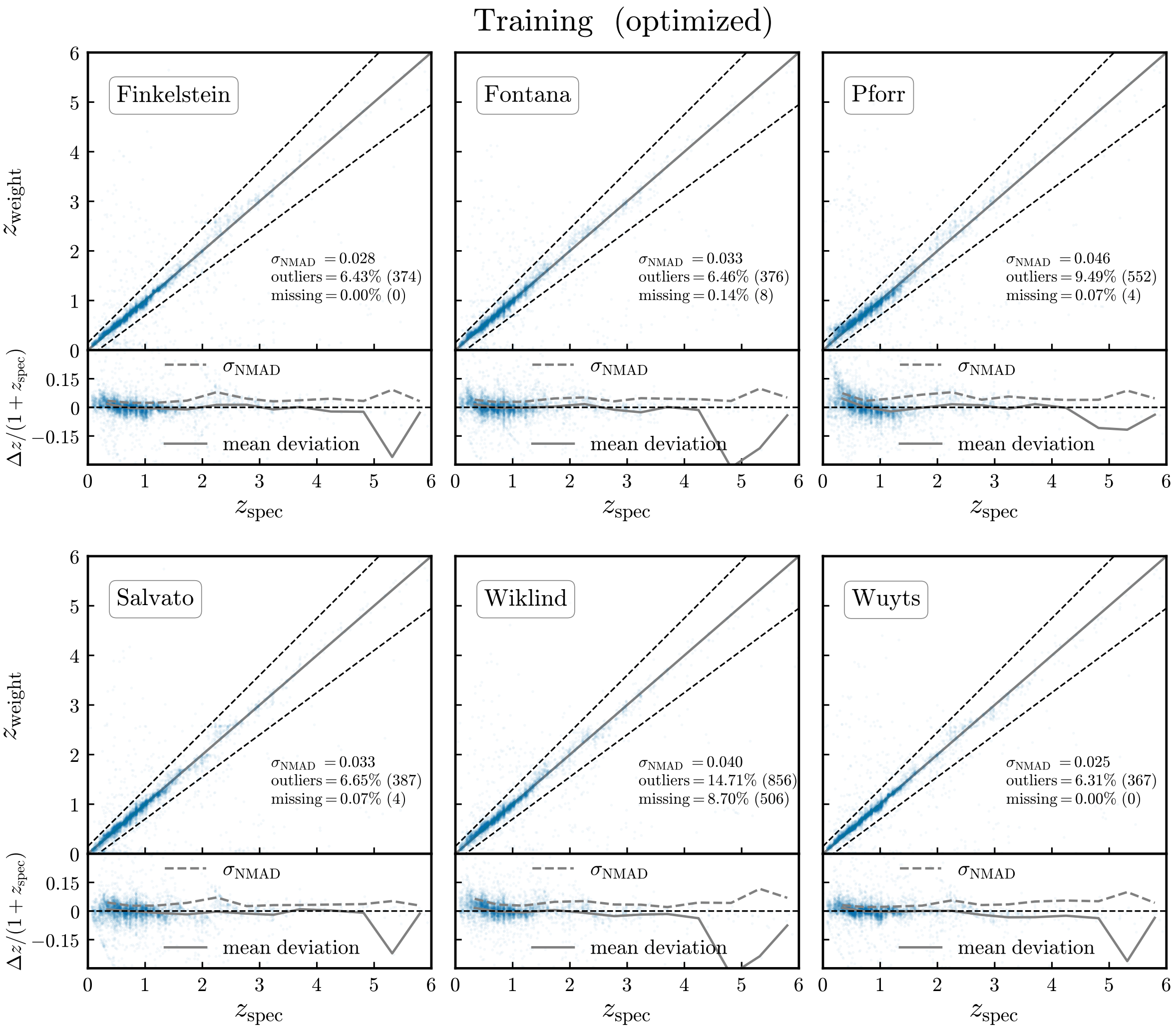


Figure 3.10 Scatter plot of z_{weight} vs z_{spec} (upper panels), using the optimized version of photo-z PDFs for the calculation of z_{weight} from the six participating groups. The σ_{NMAD}, the percentage of objects that are characterized as outliers (in parentheses the actual number of objects), as well as the percentage of objects that have missing photo-z PDFs (therefore missing z_{weight}, in parentheses the actual number of objects) is presented in the scatter plots as well for all six cases. Additionally, $\Delta z/1 + z_{\text{spec}}$ vs z_{spec} (lower panels) is presented below each scatter plot. The gray dashed and solid lines correspond to σ_{NMAD} and Hodges-Lehmann mean of $\Delta z/1 + z_{\text{spec}}$, as a function of z_{spec}. Bins of 0.5 in z_{spec} have been used for the calculation of the two curves.

been applied; results for each spectroscopic sample are shown in a separate panel. Recalibrating the PDFs generally significantly improves the scatter between photo-z's and spectroscopic redshifts, while improving or only negligibly degrading the outlier fraction.

For the objects belonging to the training or testing sets, the outlier fraction is similar from all the different groups (roughly 8%), while it ranges from 2 to 5% for objects with 3D-HST grism redshifts. There are two possible explanations for this. One is that a significant ($\sim 5\%$) fraction of the objects in the training and testing samples have incorrect redshifts assigned to them (or, alternatively, incorrect coordinates such that they are matched with the wrong CANDELS galaxy). The other possibility is that the smaller outlier fraction for 3D-HST is artificial, reflecting the fact that photometric redshifts calculated with the EAZY code (used here by Finkelstein and Wuyts) are incorporated in the grism redshift determination, so by construction the grism-z may match the photo-z.

3.4 METHODS OF COMBINING PROBABILITY DENSITY FUNCTIONS FROM MULTIPLE CODES

Previous works have found that combining results from multiple photometric redshift codes can give superior performance in tests against spec-z's than individual codes yield. Dahlen et al. (2013) found that the median of the photo-z's from all the groups contributing measurements yielded a lower scatter than any single code. Since we are not only interested in point values (such as z_{weight} or z_{peak}) but also desire accurate photo-z PDFs (or, equivalently, accurate error estimates), we require methods which can combine multiple probability density functions. In this paper we utilize two different methods for combining PDFs: the Hierarchical Bayesian approach (HB) first presented in Dahlen et al. (2013), which gives an entirely new PDF constructed based upon the input PDFs, and the minimum Fréchet Distance method (mFD), which selects one of the input PDFs as being the most representative for a given object (much as the median of a set is one element of a set which approximates its central value).

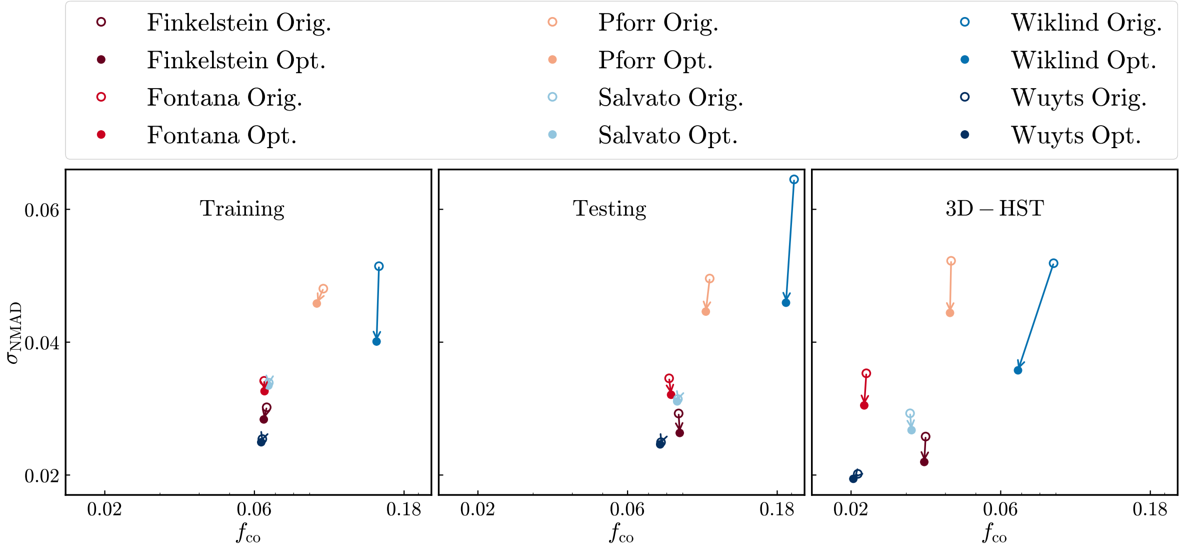


Figure 3.11 Plot of the normalized median absolute deviation between photometric and spectroscopic redshifts, σ_{NMAD}, as a function of the catastrophic outlier fraction, f_{co}. Results shown were calculated using the weighted mean photo-z peak redshifts (z_{weight}) from each group. The three panels show these statistics evaluated using the training set of spectroscopic redshifts (left); the testing set (middle); and 3D-HST grism redshifts (right). Results from Wuyts outperform the other groups in all cases, especially when evaluated using the 3D-HST data set. We caution that the 3D-HST redshift determination incorporates results from the EAZY code used by Finkelstein and Wuyts as part of the grism redshift determination process, so the level of agreement for them may be artificial.

3.4.1 Hierarchical Bayesian Combination of Photometric Redshift PDFs

The Hierarchical Bayesian approach introduced by Dahlen et al. (2013) combines PDF results from multiple codes based upon the assumptions that estimated PDFs for a given object are not entirely statistically independent (since they are based upon the same photometry) and that some may be inaccurate. This basic framework has a number of antecedents in the literature (Press (1997); Newman et al. (1999); Lang & Hogg (2012))

If we assume that the PDF from a given group is either informative about the true redshift of an object ("good") or uninformative ("bad"), we can write the posterior probability for the redshift of an object from the i'th code as:

$$P_i(z) = P_i(z|\text{bad})P(\text{bad}) + P_i(z|\text{good})[1 - P(\text{bad})], \tag{3.3}$$

where $P_i(z)$ is the posterior PDF for the redshift of the object from the i'th code when incorporating the possibility of inaccurate PDFs; $P_i(z|\text{bad})$ is the PDF resulting from an uninformative measurement (which we take here to be a uniform probability distribution from $z = 0$ to $z = 10$); $P(\text{bad})$ is the probability a randomly selected object has an uninformative measurement; and $P_i(z|\text{good})$ is the redshift PDF for a given object predicted by the i'th photo-z code. We assume that $P(\text{bad})$ is the same for all objects, and hence it will be equal to the fraction of all PDF measurements which are uninformative, which we label as f_{bad}. In that case, given a value of f_{bad} the posterior PDF will be:

$$P_i(z, f_{\text{bad}}) = P_i(z|\text{bad})f_{\text{bad}} + P_i(z|\text{good})(1 - f_{\text{bad}}). \tag{3.4}$$

The Bayesian posterior for the redshift of a given object combining the information from each PDF will then be given by:

$$P(z, f_{\text{bad}}) = \prod_i P_i(z, f_{\text{bad}})^{1/\alpha}, \tag{3.5}$$

where we introduce a parameter α which provides a correction for the covariance between the different PDFs. If the PDFs are all statistically independent, the probabilities from each one would multiply, so $\alpha = 1$. In contrast, if they were completely covariant, we would need α to match the number of PDFs being combined, n_p, so that after multiplication and exponentiation $P(z, f_{\text{bad}})$ would match the result from a single PDF.

Based upon tests with the CANDELS data, we find that in our case the optimal value of α value depends in a simple way on the number of PDFs being combined, following the equation:

$$\alpha = 1 + (n_p - 1) \times 1.1/4, \tag{3.6}$$

which yields $\alpha = 1$ for $n_p = 1$ and $\alpha = 2.1$ for $n_p = 5$ (matching the results from Dahlen et al. (2013) when $n_p = 5$). Hence, it appears that the covariance between results from different participating groups is non-negligible (since $\alpha > 1$ is optimal in general), but also far from complete (since $\alpha < n_p$).

We can then marginalize over the fraction of measurements which are bad to obtain a PDF for redshift alone:

$$P(z) = \int_0^1 P(z, f_{\text{bad}})\, df_{\text{bad}}, \tag{3.7}$$

assuming a prior distribution for f_{bad} which is flat over the interval $[0, 1]$.

The result of this procedure is a PDF that matches the PDFs from each code when they agree, but when they disagree will include extra probability at the redshifts predicted by each individual PDF. Hence, HB photo-z constraints are appropriately degraded when the PDFs disagree, in a manner which is agnostic about which PDFs are most accurate.

3.4.2 Minimum Fréchet Distance Approach

The Fréchet Distance is a commonly-used measure of the difference (or similarity) between curves (Eiter & Mannila (1994); Alt & Godau (1995)). For two given curves $P_1(z)$ and $P_2(z)$ which are defined at the same set of discrete points, the Fréchet Distance is found by combining the difference between the values of the curves at each ordinate point, as illustrated in Figure 3.12.

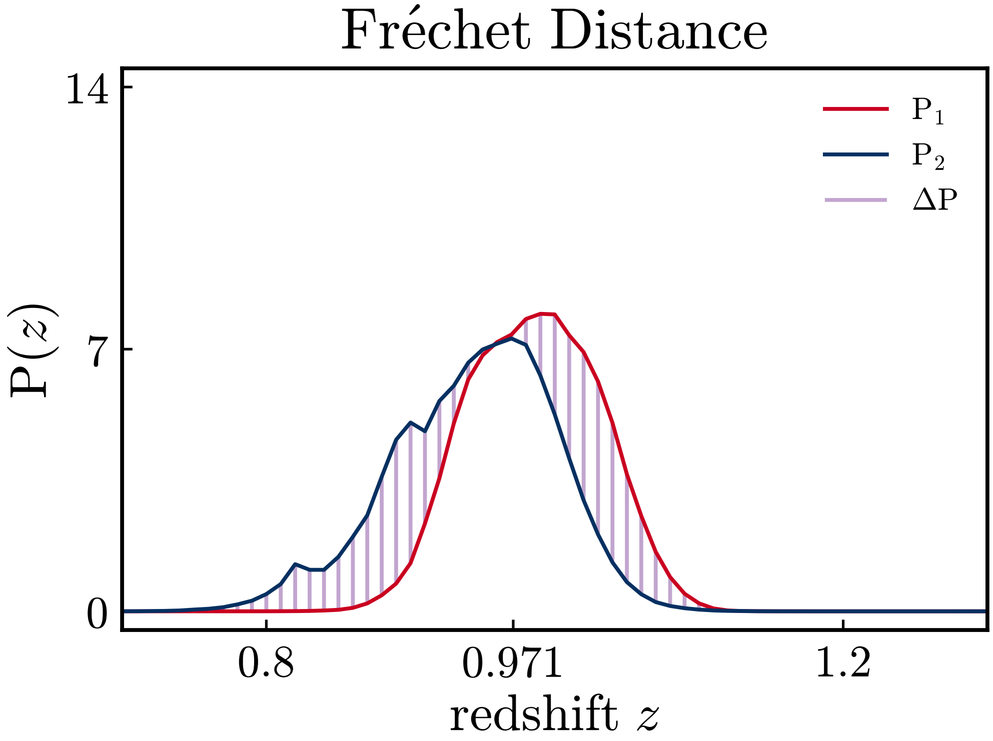


Figure 3.12 Plots of two example photometric redshift PDFs, $P_1(z)$ and $P_2(z)$, as well as the vertical difference between them at each point where they are defined, $\Delta P(z_i) = P_1(z_i) - P_2(z_i)$. The sum of the differences from each point gives the Fréchet Distance between the two curves.

In this paper, we consider Fréchet distances calculated either by summing the absolute value of the difference (an ℓ^1 distance) or by taking the square root of the sum of the squares of the differences (an ℓ^2 distance):

$$FD_{\mathrm{abs}} = \sum_i |P_1(z_i) - P_2(z_i)| \quad \text{and} \tag{3.8a}$$

$$FD_{\mathrm{sqr}} = \left\{ \sum_i [P_1(z_i) - P_2(z_i)]^2 \right\}^{1/2}, \tag{3.8b}$$

where z_i indicates the i'th redshift point at which the distances are evaluated and we use FD_{abs} and FD_{sqr} to label the ℓ^1 and ℓ^2 versions of the Fréchet distance, respectively.

In our case, six groups have provided photo-z PDFs for each object in the CANDELS photometric catalogs. We can then calculate the sum of the Fréchet distances between each PDF and each of the other five results, yielding the total distance of that curve from the rest (in the case where the ℓ^2 distance is used we sum in quadrature). We repeat this procedure for each PDF for a given object, and identify the one for which the total distance to the other PDFs is lowest; that is, the one with the minimum Fréchet Distance (mFD) from the other PDFs. This curve will be the most similar to all the rest, and therefore provides a reasonable way to summarize the ensemble of PDFs. This construction is analogous to the use of the median value of an array as a summary statistic; the median minimizes the sum of the absolute values of the deviations from all points in an array.

In the ℓ^1 distance (FD_{abs}) case, the total sum will be less affected by the largest excursions between two curves than the ℓ^2 distance (FD_{sqr}), as in the latter case differences are squared before summation, providing strong weight to large deviations. As a result the minimum Fréchet distance curve selected using an ℓ^1 distance metric, which we label as "mFDa" as it is based upon the sum of absolute values, is less sensitive to "outlier" curves than when we use an ℓ^2 metric, which we label as "mFDs" as it is based upon the sum of squared differences. The situation is analogous to the reasons why the median statistic is more robust to outliers in a data array than the mean is.

3.4.3 Comparing Combination Methods

Figure 3.13 illustrates the results of the Hierarchical Bayesian and Minimum Fréchet Distance approaches for a galaxy in the GOODS–North field. In this figure we present the PDFs from six different groups, as well as the results of each combination method. Whereas the Hierarchical Bayesian method provides a completely new photo-z PDF that is distinct from all the input curves, the minimum Fréchet Distance method selects one of the original PDFs as being the best representative for a given object. For this object, the mFDa and mFDs algorithms select different PDFs, but it is more common that they agree with each other than not.

For each of the three methods of combining photometric redshift PDFs applied here, we also explore how the quality of results changes when only a subset of the groups' PDFs are used as inputs. We note that some groups' PDFs performed significantly better than others when evaluated using the ℓ^2-norm, $f_{\rm op}$, $\sigma_{\rm NMAD}$, or $f_{\rm co}$, for both the testing set of spectroscopic redshifts and the 3D-HST grism redshifts. Given the possibility that some groups' PDFs are more useful for computing combined distributions than others, we have computed PDFs using the HB and mFD methods using only subsets of groups which yielded the lowest $\sigma_{\rm NMAD}$ and $f_{\rm co}$ values. We use the label 6 for when all six groups are used, the label 5 for when the five best groups are used, the label 4 for when the four best groups are used, and finally the label 3 for when only the three best groups are used; hence, results labeled HB4 correspond to a Hierarchical Bayesian combination of PDFs from the four best-performing codes.

In Figure 3.14 we plot results for the ℓ^2-norm and $f_{\rm op}$ from each of these subsets both before and after PDF optimization, averaging the results for these statistics from the testing set of spectroscopic redshifts and the 3D-HST set of grism redshifts. After identifying the best-performing subset for each combination method, we plot these best cases on the same set of axes to allow identification of the best method for producing combined PDFs for the CANDELS dataset. Figure 3.15 makes clear that the best choice overall is mFDa4, i.e., the case where the minimum Fréchet distance curve is chosen using only the PDFs from the four groups with the lowest individual scatters in spectroscopic redshift comparisons. This

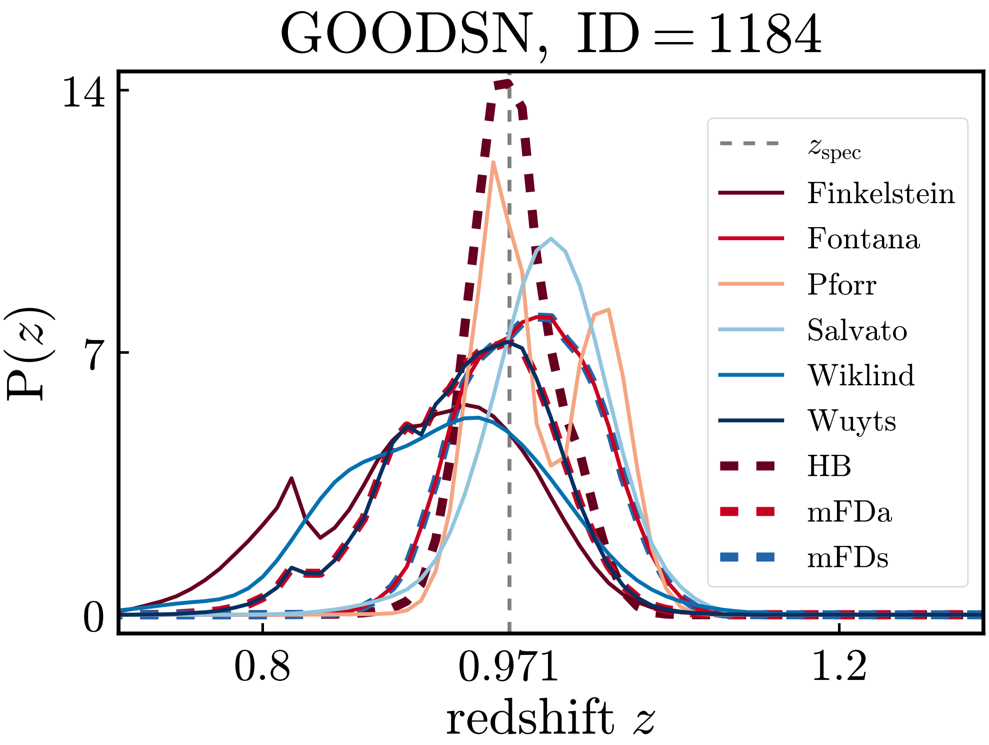


Figure 3.13 Optimized PDFs from all six groups providing photo-z's, as well as the results from three PDF combination methods, for CANDELS galaxy 1184 in the GOODS−North field. The Hierarchical Bayesian (HB, dark brown dashed curve) method produces a new PDF by combining information from each input probability distribution and accounting for the possibility that some are inaccurate. The minimum Fréchet Distance method identifies the PDF that has the smallest total Fréchet Distance from the other curves. When the sum of absolute values of differences is used as a distance metric, the minimum Fréchet distance curve corresponds to the Salvato PDF for this object (mFDa, red dashed curve), whereas when the square root of the sum of the squares of separations is used to compute distances the Fontana curve is selected (mFDs, blue dashed curve). The vertical gray dashed line indicates the galaxy's spectroscopic redshift, $z_{\rm spec}$ = 0.971, which is consistent with all PDFs shown.

method yields a lower Q-Q ℓ^2-norm than either alternative, while having an out-of-PDF fraction ($f_{\rm op}$) that is only marginally larger than mFDs4 ($f_{\rm op}$ is zero for the HB combined PDFs by construction, as the uniform probability distribution used to model uninformative measurements yields a small amount of probability at all redshifts in the final HB combination). We conclude that the minimum Fréchet distance PDF selected from the four highest-quality results comes closest to meeting the statistical definition of a probability density function for the actual redshift of a galaxy, and therefore is the preferred combination technique for CANDELS when a consensus PDF is desired.

Although mFDa4 provides the most accurate PDFs of any combination method considered, there are cases when the accuracy of point measures of the redshift, not the accuracy of the PDF (and hence e.g. error estimates), is what is desired. We therefore evaluate the accuracy of $z_{\rm weight}$ estimates from the same set of combination methods and subsets of the input PDFs considered above, using the $\sigma_{\rm NMAD}$ and $f_{\rm co}$ statistics to evaluate them. We present the averaged results from the testing set of spectroscopic redshifts and the 3D-HST grism redshifts in Figure 3.16, combining objects from all CANDELS fields. Amongst the Hierarchical Bayesian combinations considered, HB4 has the lowest $\sigma_{\rm NMAD}$ value; mFDa4 and mFDs3 yield the lowest scatters amongst the minimum Fréchet distance combinations. We compare the results for these three best cases against each other using each set of redshifts separately in Figure 3.17. In every case, the Hierarchical Bayesian combination of the best four PDFs, HB4, yields the lowest $\sigma_{\rm NMAD}$; it is therefore the combination method which provides the most accurate point values of photo-z's, with $\sigma_z \sim 0.02(1 + z)$ or better for all three spectroscopic samples. In every case it outperforms the best individual group results (compare to Fig. 3.11).

Table 5 summarizes the performance of the optimized photo-z PDFs from each group, as well as the PDFs produced by the best combination methods of each type, applying the statistics defined above to all three spectroscopic datasets across the five CANDELS fields. The method yielding the best performance for a given statistic is highlighted in blue, and the combination method with the best performance on average is indicated in red. We include one additional quantity in this table which was not defined above, $f_{\rm missing}$, which we define as the fraction of objects for which a PDF file was not provided by a given group (generally

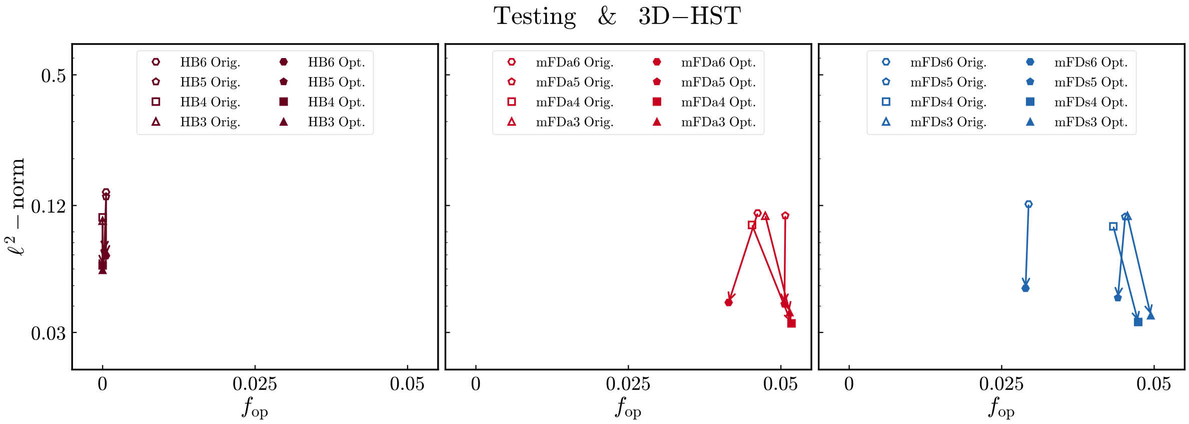


Figure 3.14 Average values of the normalized ℓ^2-norm and out-of-PDF fraction $f_{\rm op}$ for each PDF combination method considered here, combining results from the testing and 3D-HST redshifts across all CANDELS fields. We show values both without (squares at arrow tails) and with (circles at arrow heads) optimization of the PDFs from each group in advance of combination. The panels differ in the method of combination considered; here "HB" indicates Hierarchical Bayesian combination, "mFDa" indicates the minimum Fréchet distance curve selected using an ℓ^1 (sum of absolute values) metric; and "mFDs" indicates the minimum Fréchet distance curve selected using an ℓ^2 (sum of squares) metric. The number following the combination type indicates the number of PDFs combined; e.g., in the HB3 case we combine the PDFs from the best 3 groups (ranked according to their performance at point value metrics). Within each type of combination, the lowest values for the ℓ^2-norm are obtained using the optimized PDFs for **HB3**, **mFDa4**, and **mFDs4**, respectively.

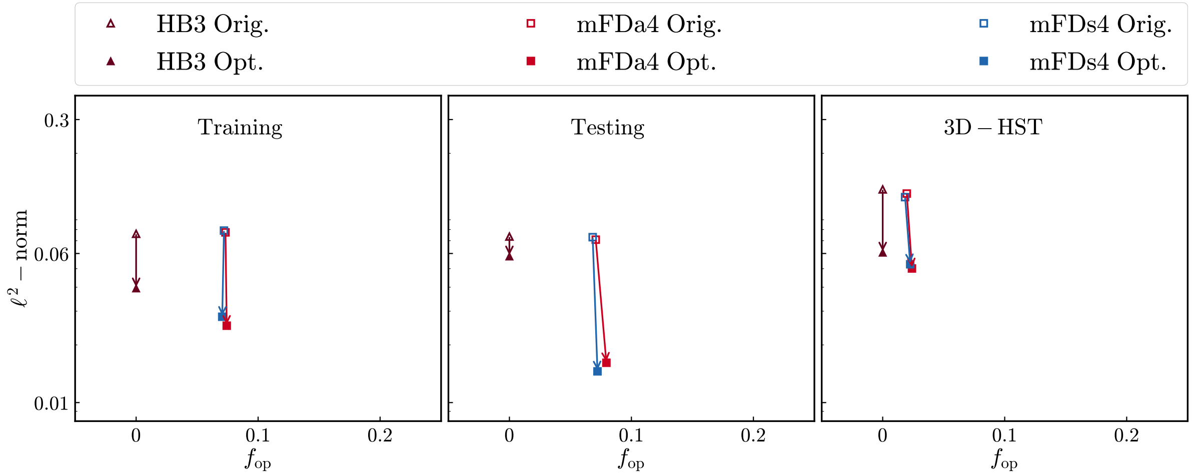


Figure 3.15 Comparison of normalized ℓ^2-norm and $f_{\rm op}$ for the lowest-norm PDF combinations of each type: HB3, mFDa4, and mFDs4. In this case we show results for the training and testing sets of spectroscopic redshifts and the 3D-HST grism redshifts separately in each panel, but combine objects from all CANDELS fields together in each. mFDa4 gives the lowest ℓ^2-norm values in every case, indicating that the minimum Fréchet distance curve constructed from the four highest-quality PDFs comes closest to meeting the statistical definition of a probability density function for the actual redshift of a galaxy.

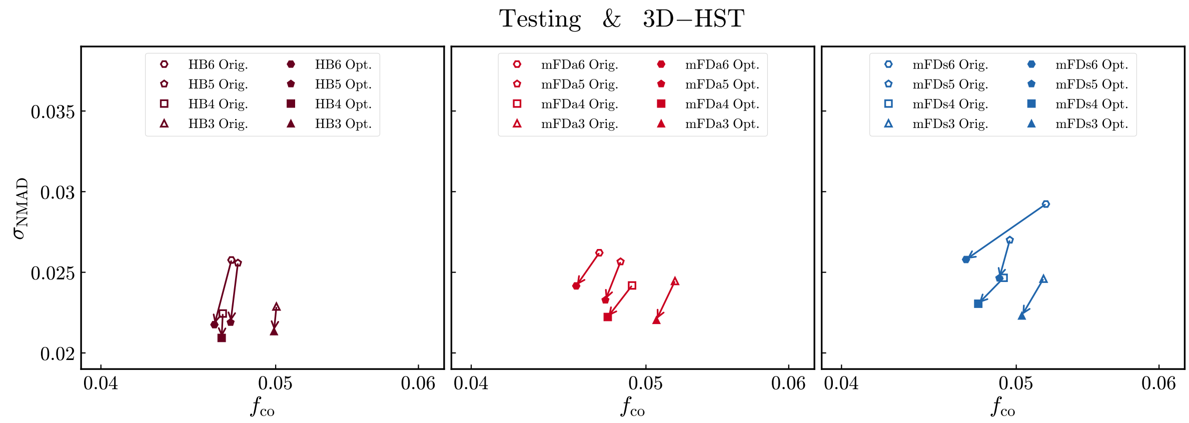


Figure 3.16 Comparison of performance of point statistics (z_{weight}) from different combination methods using the NMAD (σ_{NMAD}) and outlier fraction (f_{O}) statistics. Plotted points are the average of the results from the testing and 3D-HST redshifts, combining all five CANDELS fields. Methods are labeled in the same way as in Figure 3.14. The HB4, mFDa4, and mFDs3 combination methods yielded the lowest NMAD values, and hence the most accurate point estimates when evaluated with these samples.

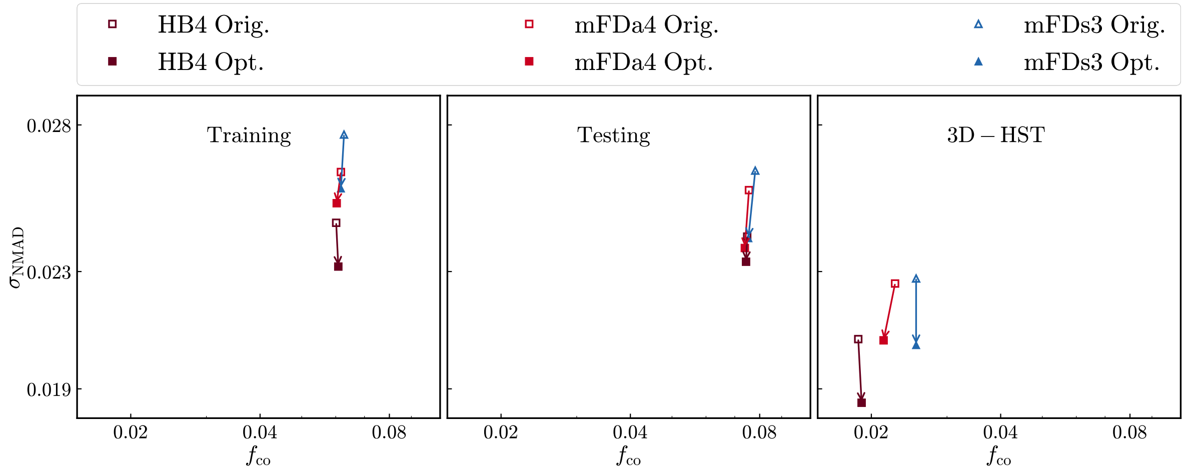


Figure 3.17 Comparison of the best point statistic results from each PDF combination method to each other on the plane of σ_{NMAD} versus f_{O}. Results for the training and testing sets of spectroscopic redshifts and for the 3D-HST set of grism redshifts are shown in separate panels. In every case, the Hierarchical Bayesian combination of the best four PDFs, **HB4**, gives the lowest σ_{NMAD} values for each set and outlier fractions comparable to or lower than other methods, making this the combination which provides the most accurate point estimates, significantly outperforming any single code.

because a given photo-z code failed to yield results for an object).

The Fontana group's PDFs yielded the smallest average normalized ℓ^2-norm, while Wiklind had the smallest out-of-PDF fraction $f_{\rm op}$. However, the mFDa4 combination method yielded only marginally larger ℓ^2-norm than Fontana while having a much smaller $f_{\rm op}$; we therefore recommend the use of this combination if the most accurate PDFs are desired.

Whereas individual groups' results yielded the best performance for some PDF quality statistics, the best point statistics were obtained using combinations of multiple photo-z PDFs. The HB4 (Hierarchical Bayesian combination of the four best PDFs) combination yielded both the smallest average NMAD (0.022) and the lowest average outlier rate (4.6%). However, it is worth noting that the results from the Wuyts group taken on their own were almost as accurate.

In conclusion, Figure 3.18 shows the diagram of the procedure from the initial photo-z PDFs obtained from the six different groups, to the final form of photo-z PDFs that can be used by the scientific community.

3.5 CANDELS PHOTOMETRIC REDSHIFT CATALOGS

Incorporating all the insights from the analyses described above, we have tabulated photometric redshift PDFs and point estimates for all objects in the five CANDELS fields.

First, we provide photo-z PDFs in a separate file for every CANDELS object (with the object identifier specified in the filename, e.g. `ALL_OPTIMIZED_PDFS_GOODSN_ID00001.pzd`). Each file has columns specifying the redshift; the PDF provided by each group after the optimization procedure from Section 3.3 has been applied; and the PDFs resulting from the two best combination methods, **HB4** and **mFDa4**, (hence the term `ALL` in the filename). The PDFs cover the redshift interval$[0, 10]$ with a step size of $\Delta z = 0.01$. The details of the columns included in these files are presented in Table 2. We additionally provide files with the original photo-z PDFs, as they are provided by the six different groups, before applying the optimization method described above. The format of this files is exactly the same as the one with the optimized PDFs, while the file name in this case is

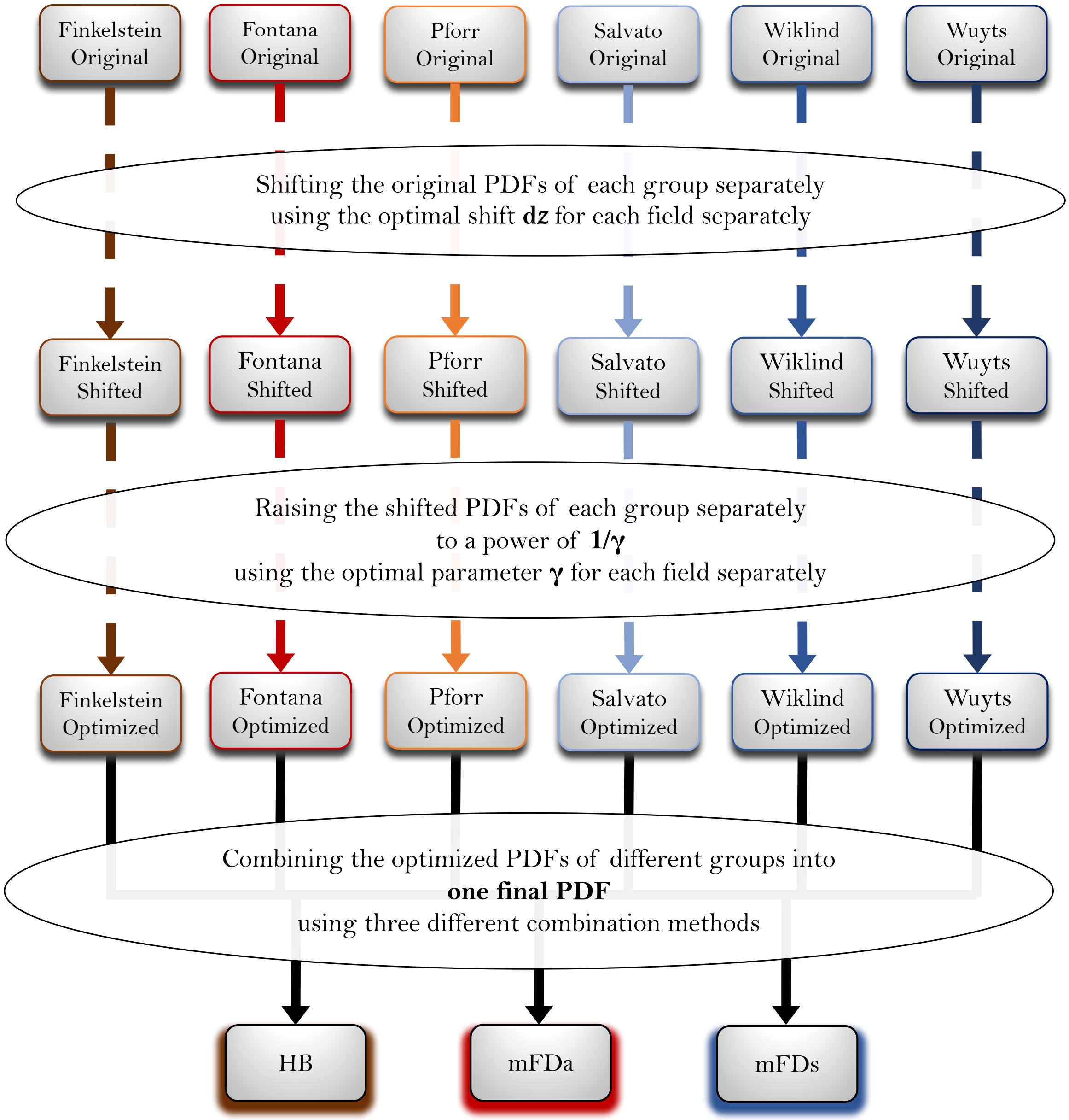


Figure 3.18 Diagram of the optimization procedure for obtaining the final products of photo-z PDFs, starting from the initially provided PDFs. First the PDFs are shifted, then raised to a power, resulting in optimized PDFs. Then the PDFs from different groups are combined into one final PDF that can be used for science; the combination methods are three.

ALL_ORIGINAL_PDFS_GOODSN_ID00001.pzd.

Additionally, we provide summary catalogs for objects in each CANDELS field. These catalogs contain photo-z point statistics constructed from the optimized PDFs and the best combinations, as well as spectroscopic and/or grism redshifts where available. These catalogs include estimates of the 1σ and 2σ credible intervals for the photometric redshift constructed from the PDFs. The columns of these files are described in detail in Table 3.

3.6 SUMMARY AND DISCUSSION

In this paper, we have described the construction of the primary photometric redshift catalogs produced by the CANDELS collaboration, providing information on more than 150,000 objects in the five fields covered by the survey. We begin with probability density functions measured by six groups within the collaboration by applying a variety of template-based methods to the same photometric catalogs. We have determined the optimal shift and exponentiation parameters for the PDFs from each group using statistics based upon the Q-Q plot, which we measure using the same training set of spectroscopic redshifts provided to each group to tune their photo-z algorithms.

In tests with both the training set of redshifts and with independent sets of spectroscopic and grism redshifts, the optimized PDFs much more closely match the statistical definition of a probability density function than those originally provided by each group, with the normalized ℓ^2-norm statistic (a measure of the accuracy of the photo-z cumulative distribution functions) improving by more than a factor of two in some cases. Point estimates of the redshift (e.g., z_{weight}) derived from the optimized photo-z PDFs also exhibit significantly smaller scatter (as measured by the normalized median absolute deviation) and smaller or negligibly worse catastrophic outlier rates, in the best cases yielding photo-z errors of $\sim 0.02(1+z)$.

After optimizing the results from individual groups, we have explored the gains from three different methods of combining the six PDFs available for each object: the Hierarchical Bayesian (HB) method described in Dahlen et al. (2013) as well as two techniques introduced here that identify the PDF with the minimum Fréchet Distance labelled (**mFDa** and **mFDs**).

We construct new PDFs by applying each method to subsets of the six results for each object. Comparing them against each other with the same statistics used to assess individual groups' results shows that combining the PDFs from the four best-performing groups produced the best results. The Hierarchical Bayesian method yielded the lowest scatter in point statistics, while the minimum Fréchet distance curve computed with an ℓ^1 metric (mFDa) had the lowest ℓ^2-norm values, indicating that it provides the most accurate PDFs, and hence the most accurate credible intervals as well.

Based upon these results, we have constructed publicly-available catalogs of optimized PDFs and photometric redshift summary statistics for all objects from the CANDELS photometric catalogs used to calculate the photo-z's. Basic advice on how best to use these catalogs is as follows:

- In general, results from different photometric redshift codes are sufficiently different from each other (as can be seen most clearly in Figure 3.5) that we recommend performing an analysis multiple times using photo-z point estimates or PDFs from different groups each time to ensure that conclusions are robust to these variations.
- Different columns of the summary catalog are better for different purposes. For instance, if one wants the best estimate for the redshift of an individual object (where uniformity does not matter) the `z_best` value from the catalog (which is determined from the combined dataset of spectroscopic redshifts, 3D-HST grism redshifts, and mFDa4 photometric redshifts) would be most appropriate. If instead the smallest-scatter estimator of redshift for a uniform sample is needed, `HB4_z_weight` (the z_{weight} value computed from the Hierarchical Bayesian combination of the four best PDFs for each object) would be most appropriate. This photo-z point estimate yielded $\sigma_{\text{NMAD}} = 0.0227/0.0189$ and $|\Delta z/(1+z)| > 0.15$ outlier fraction = 0.067/0.019 for the testing and 3D-HST redshifts, respectively.
- The mFDa4 (minimum Fréchet distance curve constructed from the best four PDFs) yielded probability density functions that best meet the statistical definition of a PDF. As a consequence, this is the preferred set of PDFs to use when the accuracy of credible intervals on redshifts is desired. Correspondingly, the `mFDa4_z_weight` column of the summary table will have the best-characterized error estimates associated with it.

The photometric redshift catalogs presented here represent the culmination of a considerable amount of effort by the CANDELS collaboration to obtain a broad range of imaging data, measure uniform photometry with TFIT, and calculate photometric redshifts. They represent a public legacy of the survey which should contribute to a wide variety of science in the future, such as the estimation of stellar masses of galaxies.

4 MORPHOLOGY-DENSITY RELATION WITH CANDELS PHOTO-Z'S

4.1 INTRODUCTION

One of the most important questions that still remains unanswered in extragalactic astronomy is the formation and evolution of galaxies. It is clear that their formation and evolution is guided by gravity and it is the gravitational instabilities that lead to the eventual collapse of slightly overdense regions and subsequently the formation of galaxies, but the exact mechanisms with which this happens still elude us. Since the fraction of matter from the dark sector dominates over baryonic matter, it is understood that its initial clumping together into the formation of dark matter halos, will help ordinary matter to start collapsing. Of course the largest halos will attract more matter and more baryonic matter will accumulate in their vicinity, leading to the formation of larger number of galaxies of greater sizes and masses. This shows that galaxies are perfect tracers of dark matter halos and their study could shed light into the properties of this yet unknown constituent of out universe.

It has been shown that different types of galaxies reside in different environments. There have been numerous studies showing that density-morphology relation is strong in the local universe. Dense regions are mostly populated by elliptical galaxies with very massive red galaxies in their centers, while spirals tend to be found in less dense regions and are rarely found to be central galaxies. This finding has been verified repeatedly even using the large volume of data from SDSS. Additionally, studies have shown that galaxies that cluster more strongly and are more dominant in dense regions are red galaxies. On the contrary, bluer galaxies are found to cluster less strongly, and reside in less dense environments. This is known as the density-color relation which has also been verified by many studies. These findings clearly suggest that the environment plays a crucial role in the formation and evolution of galaxies.

It is of great interest to see how the density-morphology relation holds at higher redshifts and see how it compares to the results mentioned so far, which have involved the relatively low redshift, therefore the local universe. In order to do so, we need observations of the high redshift universe, obtained from deep surveys. Our study does just that using data from the CANDELS collaboration, which is one of the largest collaborations using HST. Their observations go beyond redshift of $z \sim 3$, though with decreasing number of objects since galaxies become very faint at such great distances. Additionally, the collaboration has estimates of photometric redshift probability distributions for the observed galaxies which surpass the 150000 in five different areas of the sky. Making use of these photo-z PDFs, as well as the 3D-HST grism-z PDFs and spec-z's where they are available, in combination with morphology catalogs provided by CANDELS, we estimate projected correlation functions for different morphological types of galaxies, separated in different redshift bins and see how they evolve.

This chapter is constructed as follows: In Section 4.2 we describe in detail the data that we use in our study. We follow in Section 4.3 with the description and application of the method used to evaluate the density morphology relation at different redshifts. We then proceed with Section 4.4 where we present our results and conclude with a discussion about the implications of our findings as well as the future work.

4.2 DATA

First, in this work we use the data from the **CANDELS** collaborations that we have described in Chapter 3, that is the photo-z PDFs that we estimated by the team's participants and improved using the method described in Chapter 2, as well as the 3D-HST grism-z PDFs and spec-z's. In a sequence of decreasing importance the redshift information used is: spec-z wherever available, grism-z PDFs wherever available and no spec-z information, and finally photo-z PDFs if no spec-z or grism-z information is available. The photo-z PDFs used are the ones from the minimum Fréchet Distance absolute values method using only the 4 best performing participants (mFDa4), while the 3D-HST grism-z PDFs are only used if they are found to be of very high quality after applying very conservative cuts.

Apart from the redshift data, we use the morphological catalogs provided by the team, which is compiled together by Jeyhan Kartaltepe (Kartaltepe et al. (2015)). The classifications were performed by team members by visual inspection of thousands of galaxies, where a large volume of information was provided for each galaxy by more than one inspectors. The detailed description of these catalogs is presented in Table 4. We select 3 main morphological types to use in our analysis: spheroids, disks, and irregulars/peculiars, and we consider a galaxy to be a member of one of these types only if all classifiers have classified that particular galaxy with the same morphological type, therefore the fractional vote for that galaxy should be 1 for one of these three types.

4.2.1 Generated data

The number of objects in the three aforementioned morphological classes are unfortunately not enough to produce strong signals in the calculation of correlation functions. For this reason, we generate more data using the information we have from the existing data. From each object, we generate 20 more objects that have the same morphology as the initial object, and the same RA and DEC coordinates, but different redshifts. The redshifts of the generated objects follow the corresponding PDF of the original galaxy, i.e., the grism-z PDF if that is available, or if that is not available, the photo-z PDF. If the high quality spec-z is provided for the original object, then their distribution is very narrow (almost a delta function), therefore the same value of redshift is assigned to all generated data. After we perform this for all galaxies in our initial sample, we end up with a new sample with 20 times more objects, rendering it satisfying for our analysis. This approach is justified by what the redshift PDFs represent. If the PDF is representative of an objects redshift, then if we were to observe more objects of the same characteristics, then their true redshifts would follow the distribution of the original one. Thus, this new larger sample is going to be the data sample that we can use in our correlation function estimations.

Additionally, from each original object 200 more objects are generated that have the same morphological type as the original one, their redshifts follow the same distribution as the previous set of generated data, but their coordinates are not the same as the ones of the original objects, but are chosen to be randomly distributed in an area that covers the entire field where the original object belongs. This new larger sample of generated data will be the random sample which is used as a tracers sample. This is needed since by definition the correlation function gives the excess probability to find pairs at some separation, compared to what one would expect from a random distribution of objects. Furthermore, we introduce some additional noise in the redshift distribution of the random samples, by adding values from a normal distribution of mean $\mu = 0$, and standard deviation $\sigma = 0.4$, which is equivalent to convolving with the same normal distribution. This is done to avoid the randoms follow the same redshift distribution as the data.

4.2.2 Data cuts and separations

The CANDELS fields are somewhat peculiar in shapes, with many edges and angles. Given that we want the new generated samples (both data and random) to cover the exact same area in the RA-DEC plane, we apply cuts of straight lines to secure this same coverage. Additionally, in order to assign uncertainties to any of the estimated quantities, we use jackknife resampling, therefore we separate our data into subsamples in the RA-DEC plane. Due to the data not being uniform in all five CANDELS fields, we use only three of the areas: COSMOS, GOODS-South, and UDS, and each of these are divided into three subfields, therefore making it 9 subsamples in total, as shown in Figure 4.1. Furthermore, we want to investigate how the density-morphology relation changes with redshift, for which we separate our data into redshift bins: $(0.3, 0.5]$, $(0.5, 1.0]$, $(1.0, 1.5]$, $(1.5, 2.0]$, $(2.0, 2.5]$, $(2.5, 3.0]$. Unfortunately, the first and last redshift bins are very scarcely populated, therefore the results for these two redshift bins are to be taken with greater skepticism.

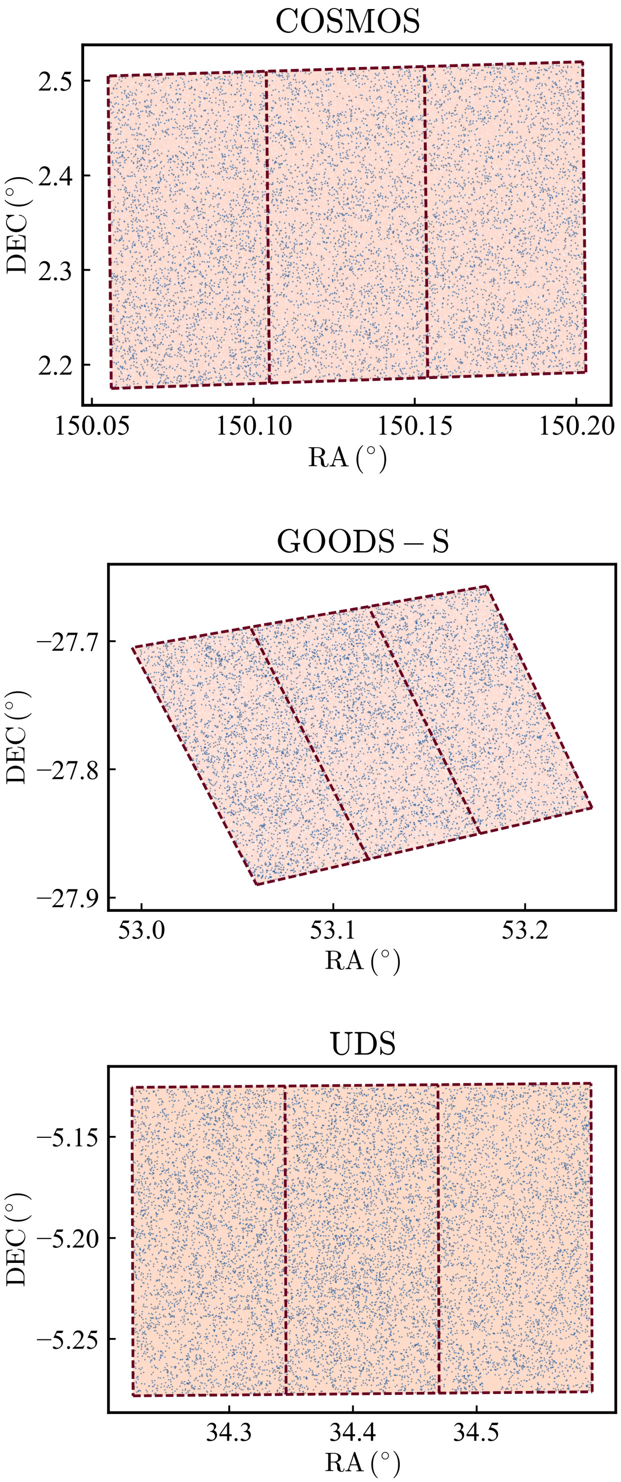


Figure 4.1 The data and random samples for the 3 fields, each divided into 3 subfields of equal area which are used for jackknife resampling.

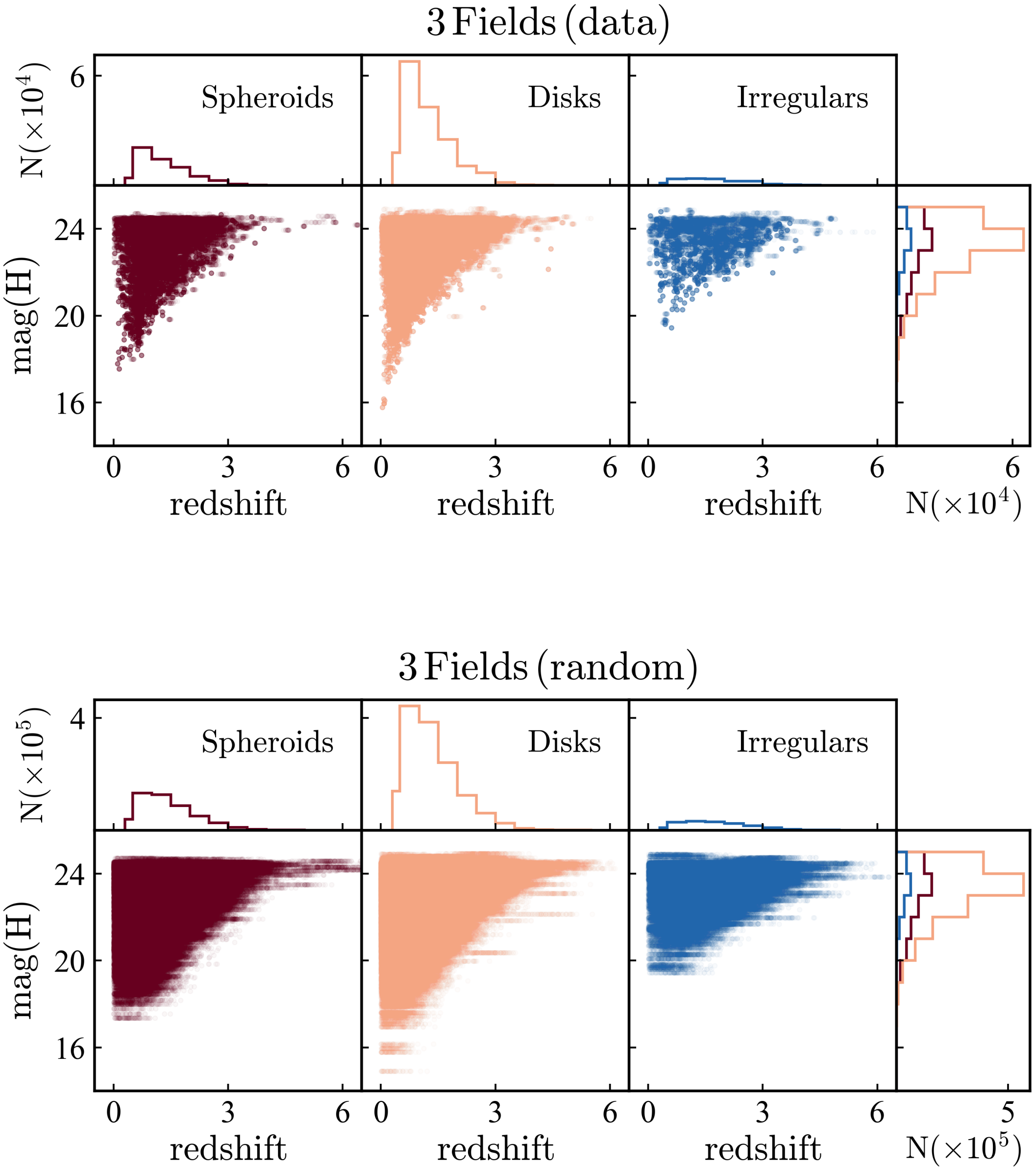


Figure 4.2 Redshift-magnitude plot of the data and random samples of the 3 morphology types for all 3 fields. The number of objects is substantially low beyond the redshift of $z \sim 3$, making the use of such objects impractical. Additionally, the sample size of irregular galaxies is fairly small to produce interesting results.

4.3 CORRELATION FUNCTION

In order to see how different morphological types of galaxies cluster together, we use the cross correlation function, i.e., we cross correlate each morphological type with the full sample of galaxies. The two point correlation function is a measure of the excess probability of finding a pair of galaxies at some separation r, compared to probability when the population is randomly distributed. Its mathematical form is defined by:

$$dP = n[1 + \xi(r)]dV \tag{4.1}$$

where n is the number density of galaxies, dV is the volume element, and dP is the excess probability (Peebles (1980)). It is a function of the 3 dimensional distance between pairs of objects r, and is usually modeled as a power law, given by:

$$\xi(r) = (r_0/r)^\gamma \tag{4.2}$$

where r_0 is the cluster length and γ is the slope with which the correlation function drops in log space. These two parameters are not completely independent, such that the value of γ will affect the value of r_0 as well. Instead of the real space separation r of two distant objects, we can use the separation along the line of sight π and the one perpendicular to it, r_p, therefore making the correlation function a function of two variables $\xi(\pi, r_p)$. In order to avoid distortions due to peculiar velocities along the line of sight, we can estimate the projected correlation function, which is found by integrating the over the line of sight, leading to (Davis & Peebles (1983)):

$$w_p(r_p) = 2\int_0^\infty \xi(\pi, r_p)d\pi \tag{4.3}$$

which is usually integrated up to a maximum value π_{max}, since for very large values of π, $\xi(\pi, r_p)$ becomes very noisy. For slowly varying If $\xi(r)$ is modeled by Equation 4.2, then the projected correlation function is given by:

$$w_p(r_p) = r_p \left(\frac{r_0}{r_p}\right)^{\gamma} \frac{\Gamma(1/2)\,\Gamma[(\gamma-1)/2]}{\Gamma(\gamma/2)} \tag{4.4}$$

with Γ being the known gamma function (Coil (2013)).

For our data, we want to cross correlate galaxies of a particular morphological type with the full sample of galaxies, which we accomplish by using the python module is (Jarvis et al. (2004)). Since we have data and random samples for each individual morphological type and for the full sample, as was described in Subsection 4.2.1, we calculate the modified correlation function using the generalized Landy-Szalay estimator (Landy & Szalay (1993)):

$$\xi_{LS} = \frac{D_1D_2 - D_1R_2 - R_1D_2 + R_1R_2}{R_1R_2} \tag{4.5}$$

where D_1D_2 are the data-data pairs between one particular morphological type and the full sample of galaxies, D_1R_2 are the data-random pairs between that morphological type and the full sample, R_1D_2 are the random-data pairs, and R_1R_2 are the random-random pairs. If one assumes that R_1R_2 varies slowly with π, then it can be shown that:

$$w_p(r_p) = 2\int_0^{\pi_{max}} \xi(\pi, r_p)d\pi \approx 2\pi_{max}\,\xi(\pi, r_p) \tag{4.6}$$

We can therefore use our data to estimate $\xi(\pi, r_p)$ and subsequently the projected correlation function $w_p(r_p)$. Then we can fit for the parameters r_0 and γ using Equation 4.4, finding this way the clustering length of different populations of galaxies.

4.4 RESULTS AND DISCUSSION

We use the data from the CANDELS collaboration for three main morphological types: spheroids, disks, and irregulars, and estimate the projected correlation functions as described in the previous section. In order to achieve this, we separate the data into 9 smaller regions, so we can estimate jackknife errors, by omitting one region at a time and performing our estimations. Additionally, following the estimation of $w_p(r_p)$, we fit for the parameters r_0 and γ. Since the correlation function deviates from a power low at large separations, we also fit for a third parameter, a constant value c, which is added to the original correlation function, in order to slightly increase the very small values at large projected separations r_p. Figure 4.3 shows the projected correlation functions of the spheroids and disks for the 6 redshift bins. The plots for the irregular type of galaxies is shown separately since this type has a small sample size and therefore not a very clear correlation signal, as is seen in Figure 4.4. The results of the fits of the parameters are presented in Table 4.1

Additionally, we show the redshift dependence of the parameters r_0 and γ in Figure 4.5, for the spheroid and disk galaxies, and we don't see any statistically significant difference between the values of the two populations, given their error bars. Nevertheless, the picture we observe is that different types of galaxies cluster together in different ways. More specifically spheroid galaxies cluster more strongly than disk galaxies, and this is obvious in the bins of lower redshift as well as the larger redshift ones. The clustering signal is stronger for spheroids when the projected separation is smaller, which indicates that these types of galaxies reside in the central regions of dense environments of clusters, whereas disk galaxies reside mainly in the edges of these structures. At higher redshift bins, the clustering signal of disk galaxies starts to become stronger and it is comparable, or even stronger than the signal for spheroids. In order to see such galaxy types at these early times, they have to form closer to regions of the universe where there is enough material to fuel such a rapid galaxy evolution. That of course can happen in the dense environments of galaxy clusters, therefore the clustering signal of these galaxies becomes stronger at lower separations for higher redshift, where the matter density is greater. This can also be seen when calculating the relative bias between the spheroid and the disk galaxies, which is given by the ratio of the projected correlation

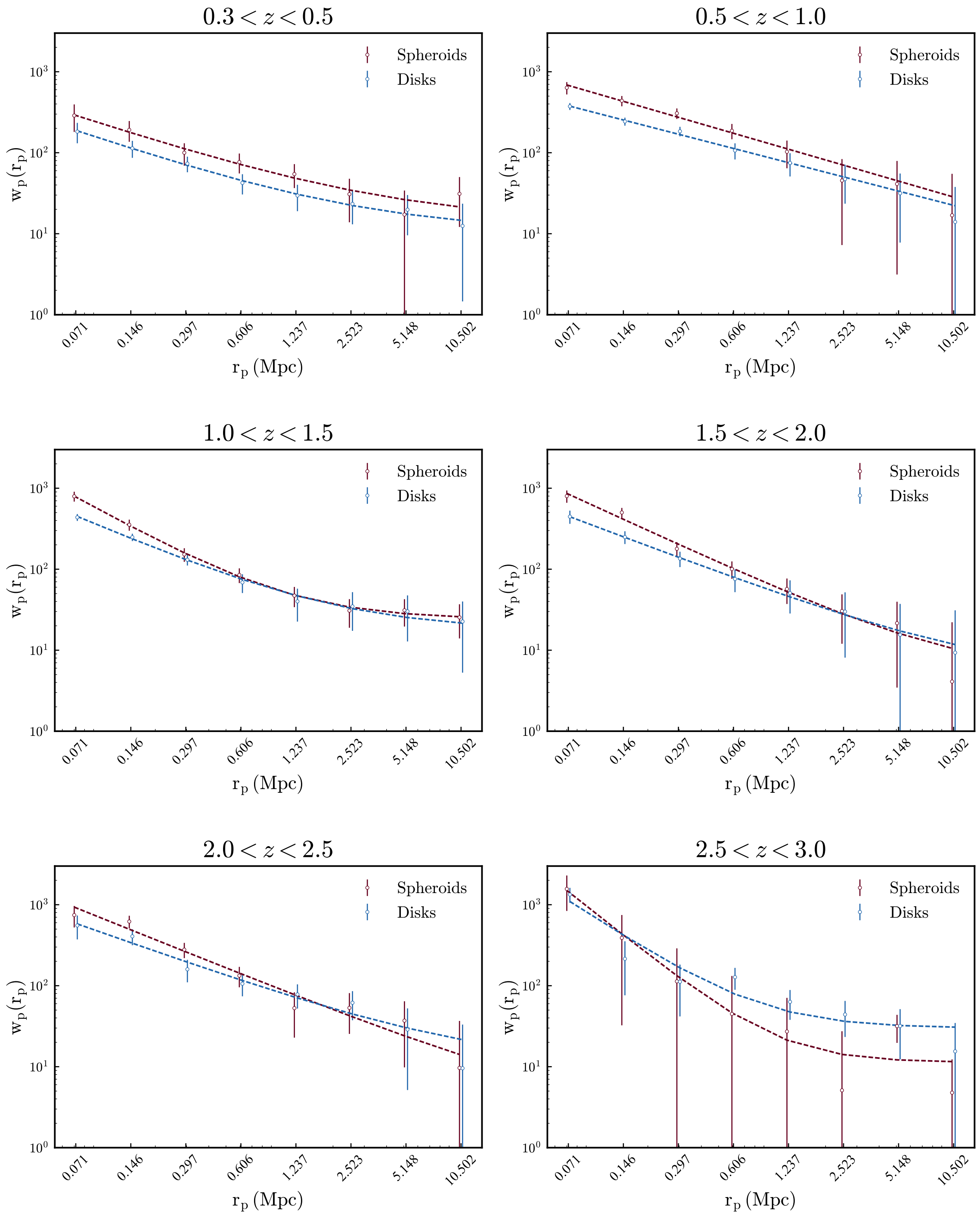


Figure 4.3 Projected correlation functions for the two morphological types, in the 6 different redshift bins. The larger values for the spheroids show that this type of galaxies clusters more strongly than disks for the lower projected separations (r_p). The picture start to change for larger separations and higher redshifts.

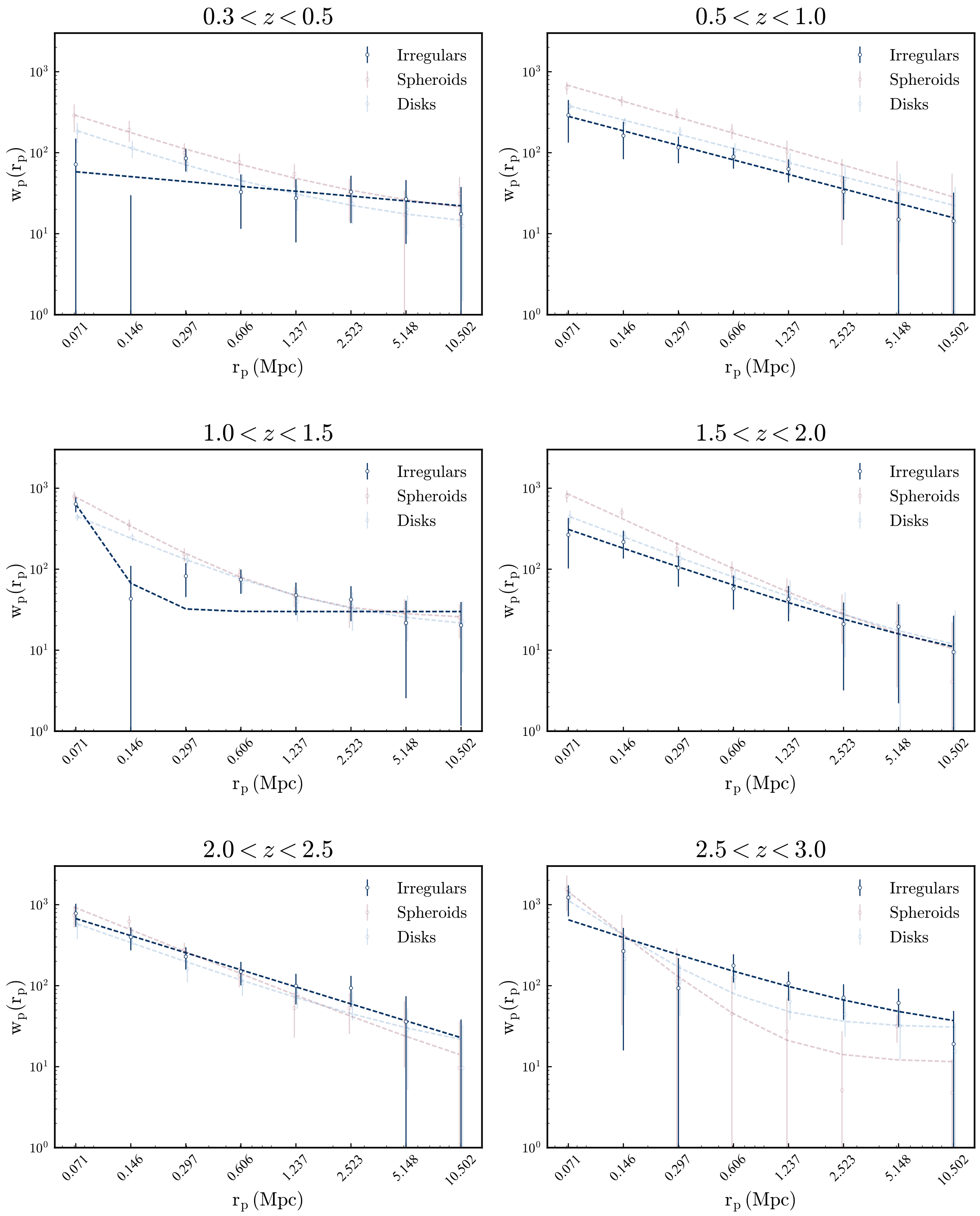


Figure 4.4 Projected correlation functions for the irregular galaxy type, in the 6 different redshift bins. The fits for these galaxies are not very good, probably due to the low sample size.

Table 4.1 Values of the fitting parameters of $w_p(r_p)$ for the different redshift bins and different morphologies. There appears to be no statistically significant difference between the populations, in any of the redshift bins, with the exception of γ in $1 < z \leq 1.5$, between spheroids and disks.

redshift bin	morphology	r_0	γ	c
0.3 – 0.5	spheroids	3.905 ± 1.501	1.725 ± 0.161	13 ± 10.634
	disks	2.738 ± 1.08	1.787 ± 0.279	11.542 ± 12.645
	irregulars	2.239 ± 2.798	1.194 ± 0.29	3.333 ± 26.667
0.5 – 1.0	spheroids	7.896 ± 1.617	1.637 ± 0.077	0
	disks	6.229 ± 1.365	1.571 ± 0.105	0
	irregulars	5.095 ± 1.507	1.584 ± 0.195	0
1.0 – 1.5	spheroids	2.973 ± 0.408	2.222 ± 0.126	23.901 ± 14.955
	disks	3.415 ± 0.457	1.946 ± 0.082	17.515 ± 14.136
	irregulars	0.737 ± 2.194	5.403 ± 6.519	30 ± 0
1.5 – 2.0	spheroids	4.3 ± 1.154	2.021 ± 0.169	5.137 ± 10.012
	disks	4.198 ± 1.743	1.839 ± 0.314	4.277 ± 12.186
	irregulars	3.797 ± 1.418	1.781 ± 0.29	4.385 ± 7.488
2.0 – 2.5	spheroids	5.591 ± 1.063	1.894 ± 0.119	4.964 ± 19.024
	disks	5.2 ± 1.413	1.793 ± 0.184	10.824 ± 21.562
	irregulars	6.646 ± 2.856	1.725 ± 0.265	6.706 ± 21.439
2.5 – 3.0	spheroids	2.011 ± 0.881	2.761 ± 0.443	11.103 ± 16.501
	disks	2.676 ± 1.557	2.419 ± 0.538	28.005 ± 8.143
	irregulars	6.07 ± 3.459	1.757 ± 0.498	19.034 ± 34.268

functions:

$$b_{s,d} = \frac{b_s}{b_d} = \frac{w_{p,s}(r_p)}{w_{p,d}(r_p)} \tag{4.7}$$

with $w_{p,s}(r_p)$ the projected correlation function of spheroids, and $w_{p,d}(r_p)$ the one for disks. This indicates that when this ratio is greater than unity, then the spheroids have a stronger clustering signal than the disks. In Figure 4.6 we see that at the lowest redshift bins, the relative bias is greater than one up to large separations, while it decreases and becomes less than unity for smaller separations at higher redshift bins. For the redshift bin $0.5 < z \leq 1$ in particular, which is also the bin with the largest sample size, the relative bias favors spheroid galaxies up to relatively large separations. We want to test the hypothesis that the relative bias is consistent with unity by calculating the reduced χ^2 with 8 data points and no degrees of freedom, using the uncertainties estimated with jackknife resampling. We then evaluate the p-value for each redshift bin in order to try to reject the null hypothesis. We perform the same test not against unity but against the relative bias values from the $0.5 < z \leq 1$ reference bin of redshifts, and present the results in Table 4.2. The $\mathrm{p-values} < 0.05$ for the redshift bins $0.5 - 1.0$ and $1.0 - 1.5$, reject the null hypothesis that the relative bias is consistent with the value of 1. Two of the other redshift bins, $0.3 - 0.5$ and $1.5 - 2.0$, do not completely reject the null hypothesis, but the low $\mathrm{p-values}$ of these bins suggest that there is tension with it. For the last two bins, $2.0 - 2.5$ and $2.5 - 3.0$, the $\mathrm{p-values}$ are even larger, certainly not rejecting the null hypothesis, though the fact that they're much lower than unity, there is still some tension with the null hypothesis. These results reinforce the idea that the two different populations of galaxies cluster differently. More specifically, spheroid galaxies tend to pair together more strongly at smaller separations, and therefore occupy the central regions of galaxy clusters, whereas disk galaxies are found to pair at larger separations therefore occupying the outskirts of clusters. Finally, the χ^2 and $\mathrm{p-values}$ for the comparison with the reference redshift bin $0.5-1.0$, show that there is difference between the redshift bins, but not a very large one. The very low $\mathrm{p-value}$ for the last redshift bin is possibly due to that bin being very noisy and the results for it are probably not to be trusted.

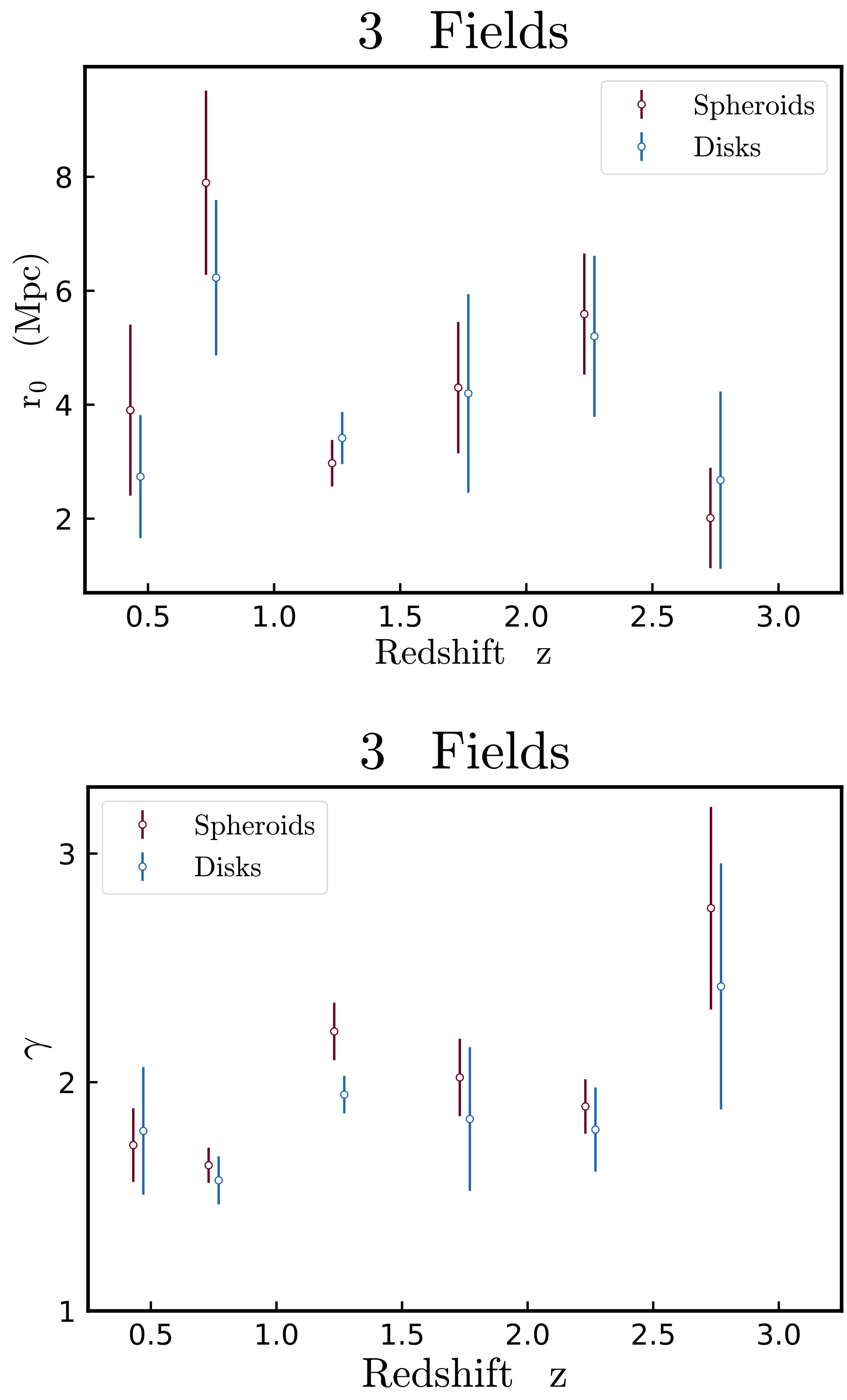


Figure 4.5 The main clustering parameters r_0 and γ as a function of redshift. No statistically significant difference is observed between the two populations, with the exception of γ in $1 < z \leq 1.5$.

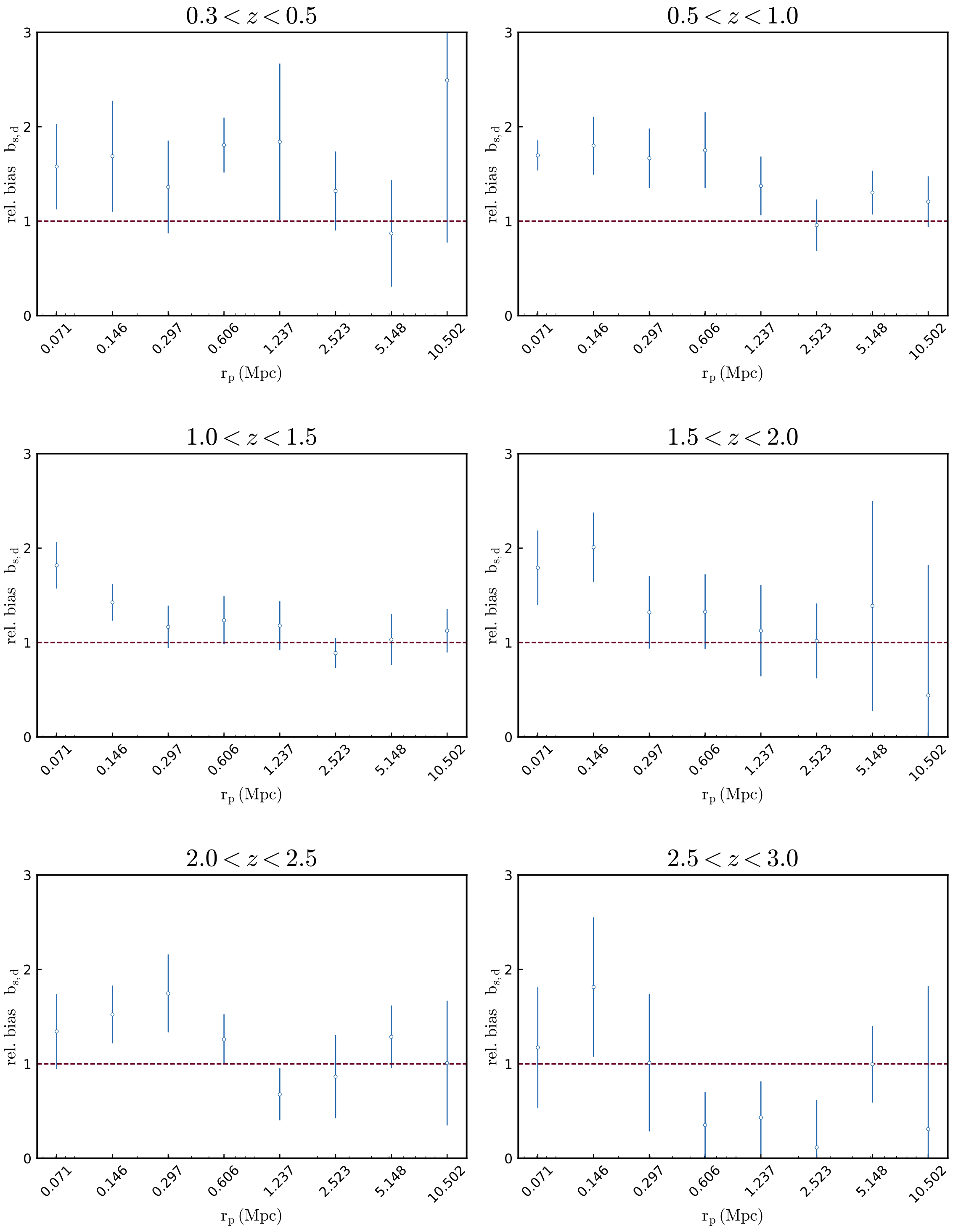


Figure 4.6 Relative bias $\mathsf{b_{s,d} = b_s/b_d}$ as a function of separation r_p, for the 6 redshift bins. The horizontal dashed line represents a relative bias of unity, which represents the two types of galaxies having the same values of projected correlation functions. There is a trend for the relative bias to be above the dashed line for almost all redshift bins, except the last one.

Table 4.2 χ^2 and p-values for relative bias compared to value of 1, and values of $0.5 < z \leq 1$.

redshift	Comparison with values of 1		Comparison with values of $0.5 < z \leq 1$	
bin	χ^2	p – value	χ^2	p – value
0.3 – 0.5	13.69	0.09	2.229	0.973
0.5 – 1.0	37.559	0	–	–
1.0 – 1.5	18.62	0.017	5.016	0.756
1.5 – 2.0	13.323	0.101	1.807	0.986
2.0 – 2.5	10.109	0.257	5.095	0.747
2.5 – 3.0	10.341	0.242	14.909	0.061

Our study produces results that our in agreement with previous studies finding that spheroid galaxies cluster more strongly than disk galaxies. This has been observed numerous times in the local universe and has been one of the main topics of interest in the field of Astronomy, known as the morphology-density relation. Our findings continue to show a similar picture at higher redshifts also, meaning that the mechanisms that drive this observation must have been in place since redshift of $z \sim 2$. While this study needs to be expanded in order to have a more definitive answer regarding this phenomenon, this study provides a first interesting result that sheds some light in the topic of galaxy formation and evolution from early times until today. At what exact time the mechanisms responsible for the observed picture in the local and distant universe started to affect galaxy evolution, remains yet to be seen. Nevertheless, our results suggest that these processes were already effective at earlier times than previously thought by the scientific community.

Another interesting conclusion that can be derived from our findings is related to previous analyses performed regarding other characteristics of galaxies at high redshift, such as their colors. Previous studies have shown that the color-density relation ceases to exit at redshifts higher than $z \sim 1.3$, whereas in the local universe, observations show that the color-density relation and the morphology-density relation go hand in hand. This seems to not be the case for redshifts higher than $z \sim 1.3$, which suggests that the mechanisms driving the color-density relation must have started to produce an effect on galaxies at later times.

Nevertheless, as we have already mentioned, for a more conclusive answer regarding our findings, this exercise needs to be expanded for larger datasets, covering larger volumes, something that will become possible with new surveys such as LSST (Ivezic et al. (2009)), EUCLID (Laureijs et al. (2012)), WFIRST (Spergel et al. (2015)) and many others, where the number of galaxies observed is vastly larger than that of CANDELS. With such large samples of data, a more thorough study can be accomplished, with finer grids in redshift, as well as to investigate the color-density relation by dividing galaxies into separate groups according to their color.

5 DISSERTATION CONCLUSION

The study of galaxies and their evolution is of the utmost importance for understanding our universe and its unknown constituents such as dark matter and dark energy. In this study we present the steps we have taken to study the morphology-density relation at high redshifts, using photometric data from the CANDELS collaboration (Grogin et al. (2011); Koekemoer et al. (2011)). This relation and its evolution with time provides useful information regarding some of the biggest questions of the formation and evolution of galaxies. Our current picture of the universe is that ordinary matter cools and collapses gravitationally as the universe expands, forming large structures of stars such as galaxies, that in turn group together to form clusters. This picture is not yet complete since it does not fully explain the variety of galaxies of different properties such as mass, size, shape, color, etc., that we observe in the universe. This diverse population is observed both locally and at high redshift, though there is clear evolution of galaxy properties with time. The low redshift regime has been extensively investigated using data from surveys such as Lahav et al. (2002), and York et al. (2000), regarding the dependence of galaxy properties on environment. There have also been studies of these dependencies at higher redshift using data from the Lilly et al. (2007) and Davis et al. (2003) surveys, showing similar trends up to $z \sim 1$. In this study, using the photometric data of CANDELS we investigate how morphology depends on environment up to a redshift of $z \sim 3$, something that has never been studied before. This would provide very useful information about how the picture that we see more locally differs from that at such high redshifts. We present below a detailed summary of the previous chapters of our work.

5.1 SUMMARY

In Chapter 2 we described statistical methods that can be used to improve the quality of photometric redshift probability distributions. This work makes use of the Q-Q plot, which is a very useful tool to assess issues with any photo-z PDFs moments, including the mean, standard deviation, skewness and kurtosis. We show that if photo-z PDFs do not match the statistical definition of a PDF, we can detect this using a sample of objetcs with spectroscopic redshifts. Any issues are quite visible in the Q-Q plot, and this information can be used to recalibrate photo-z PDFs and make significant improvements. We show that we can detect deviations of the photo-z PDFs from the spectroscopic distribution in the mean, standard deviation, skewness, and kurtosis, by using mock generated data. Furthermore, we investigate the possibility of identifying biases in the PDFs compared to the true redshift distributions, using perfect Gaussian distributions for simplicity. We perform this inquiry for cases of different sample sizes, as well as cases of spectroscopic distributions with different means or standard deviations. We continue our work by trying to identify errors in the widths of the PDFs, i.e., their standard deviations compared to the spec-z distribution, again for cases of different sizes, means, and standard deviations. Our results show clearly that we can retrieve these quantities to great accuracy, with the advantage that the method presented here of minimizing the Q-Q plot does not require any assumptions about Gaussianity and works with any type of distributions. Other methods that provide informations regarding the photo-z PDFs quality have been proposed by other scientific groups, with one of them being presented in Wittman et al. (2016). This method can provide information regarding the standard deviation and kurtosis but fails in the bias and skewness of the PDFs, due to the symmetric nature of the statistics used.

In Chapter 3 we implemented this new technique using data from the CANDELS Collaboration, which has obtained photo-z PDFs from six different groups. We demonstrate that the photo-z's after recalibration using the methods from Chapter 2 show significant improvements when compared to the ”true”, independent (from the training set) spectroscopic redshifts of subsamples of our data. Additionally, we show even more improved results after using methods that combine the photo-z distributions of multiple participants compared to

PDFs from each participant separately. In order to construct the Q-Q plots and apply the optimization methods a set of spectroscopic redshifts was used, labelled the training set. We then test the optimized photo-z PDFs both with the Q-Q plot statistics and point statistics by calculating the scatter, using an independent sample of spec-z's. We also use 3D-HST grism redshifts and perform additional tests with the same methods (Q-Q plot and NMAD), in order to reinforce our belief of improved PDFs, or to point out any inconsistencies of the optimization method. After the production of the new and improved photometric redshift PDFs, we generate catalogs of summary statistics such as z_{peak}, z_{weight}, as well as 68.3% and 95.4% credible intervals using the data provided by the six participating groups, and the two best combination methods (HB4, and mFDa4). These catalogs along with the optimized PDFs, are provided to the CANDELS Collaboration and will soon be published and provided to the scientific community to be used in plenty of studies regarding the data from the five fields of CANDELS.

Finally, in Chapter 4 we use the improved photo-z PDFs from CANDELS developed in Chapter 3, as well as 3D-HST grism-z PDFs and spec-z's from other sources, in order to study the morphology-density relation, by calculating two-point cross-correlation functions for different morphological types of galaxies with the overall CANDELS galaxy sample. We also make use of the collaboration's morphology catalogs, which contain visual classifications of galaxies as spheroids, disks, and irregulars. Therefore we investigate which morphological types of galaxies dominate at different projected separations r_p for different redshift bins, to see if the general picture evolves with time. We find that spheroid galaxies cluster more strongly, especially at small separations, compared to disk and irregular galaxies. This is particularly true for the redshift bin $0.5 < z \leq 1$, which has the largest sample sizes and the strongest correlation signal. We find a similar behavior for bins at larger redshift, though the signal becomes too noisy to make definitive determinations for the highest redshift bins, with results below the threshold for statistical significance. The same can be said for the lowest redshift bin also, for which the sample sizes are also small, giving statistically insignificant results, although for small separations the clustering for spheroids continues to appear stronger than for disk galaxies.

5.2 DISCUSSION

Our results are in general agreement with previous findings (such as those discussed in Section 1.4, which similarly suggest that elliptical and S0 galaxies are found to dominate the higher density regions such as the centers of clusters at low and moderate redshifts, whereas spiral galaxies tend to be found in the field or on the outskirts of clusters, an observation that is commonly known as the morphology-density relation. The tendency of spheroids to dominate in cluster cores causes many to be found near each other in such regions, boosting their clustering signal compared to disks at small projected separations (r_p).

The fact that we observe an enhanced clustering of spheroids relative to disks on cluster-like scales ($<$ 1 Mpc) at higher redshifts suggests that the mechanisms responsible for the formation of elliptical galaxies were already efficient at such early times. The excess is statistically significant up to redshift 1.5, and highly consistent with having a constant strength at redshifts as high as 2.5. This would suggest that environmental processes that produce spheroids were likely in place only a few billion years after the Big Bang.

If we consider this finding in light of the mechanisms described in Section 1.3, we conclude that major mergers would be one of the main processes that would convert disk and irregular galaxies into spheroids in clusters. That this process could be effective at early times is not tremendously surprising since at these early epochs, groups and early clusters are expected to have appropriate conditions for mergers to take place in abundance. This is because the number density of galaxies is expected to have been relatively high, while galaxies themselves are expected to be more gathered together in groups rather than clusters at these early stages. Therefore the galaxies' peculiar velocities are not too high for the merging process to be efficient. Other mechanisms such as tidal effects from other galaxies and/or the cluster potential or galaxy strangulation, as well as hydrodynamical effects such as ram pressure or viscous stripping and thermal evaporation can significantly alter the shape and size of disks, while also quenching star formation and therefore greatly affecting the colors of galaxies. Nevertheless they cannot generally cause disk or irregular galaxies to transform into spheroids. Therefore our results favor a scenario where major mergers of galaxies occurred efficiently at high redshifts, up to at least $z \sim 2$.

In contrast to our findings on the relationship between morphology and environment, previous studies of the color-density relation at high z (e.g. Cooper et al. (2007)) show that the fraction of red galaxies in dense environments at high redshifts ($z \sim 1.3$) was smaller, favoring a scenario where in the past the most massive blue galaxies could be found in denser regions, in contrast to what is observed today. Locally, central regions of clusters are usually occupied by massive red early type (usually elliptical and lenticular) galaxies, whereas blue disk and/or irregular galaxies are found primarily on the outskirts of clusters or in the field. This observation of the low-z universe suggests that today the morphology-density and color-density relations go hand in hand, something that does not seem to have been the case at higher redshifts. If this finding holds, it suggests that the mechanisms driving the morphology-density relation were in place at earlier times ($z \sim 2$) than the ones responsible for the color-density relation. A consequence of this would be major mergers being the leading process responsible for galaxy evolution at such high redshifts.

5.3 FUTURE WORK

This work provides an initial study of the morphology-density relation at relatively high redshifts. In order to get a clearer picture and a better understanding of galaxy evolution, samples covering larger volumes are needed. This could be provided by obtaining morphological classifications for more galaxies from the five CANDELS fields, since only a small fraction of them were definitively classified by the team, even after very exhaustive efforts incorporating many of the collaboration's members. At the time when this study was done, only a few thousand galaxies had been classified for their shape, and only in three out of the five CANDELS field. Recently new catalogs with a larger number of galaxies that have been morphologically classified were distributed to the collaboration; therefore it would be interesting to repeat this exercise and see if our findings still hold and if their statistical significance might increase.

Additionally, there exists another set of morphological classifications of galaxies from the CANDELS fields via an effort led by Brooke Simmons with the Galaxy Zoo project Simmons et al. (2017), similar to what was done by with galaxies from the SDSS survey (Lintott et al. (2008)). This work classifies galaxies into different types such as smooth or featured galaxies. The smooth category would generally correspond to early-type galaxies (E, S0) whereas the featured category would correspond to late-type ones (S & Irr). Our present work could easily be further extended using these morphological catalogs as well, an effort that was not able to be completed to date due to lack of time. This might enable studies with enlarged samples that again might provide greater statistical significance.

An obvious addition to our work would also be to perform the same investigation in samples that are not only redshift-limited, but also mass-limited. We have already mentioned that various galaxy properties and their evolution with time (and therefore redshift) depend not only on environment but also on stellar mass. Even though previous studies have found similar trends to this work when using mass-limited samples, it would be interesting to see how our results would differ by applying the same types of constraints. This again requires a larger sample of classified objects, as was mentioned before, but also requires stellar mass estimates. The CANDELS collaboration is currently working to measure stellar masses for galaxies based on the improved photometric redshifts we have produced and described in this thesis. The results will be provided to the rest of the collaboration in the near future.

Another exercise we can perform is to test how the color-density relation evolves with redshift, and see how it compares with the morphology-density relation that we have studied here. As we have already mentioned, color is closely related to the morphology of a galaxy in the local universe, but it correlates more weakly with environment at higher redshifts. Therefore investigating the evolution of the color-density relation using CANDELS data and compare the results with our findings would be a very interesting addition to this work. In order to do so, again our new and improved photometric redshifts presented in Chapter 3 can be used in order to estimate the rest-frame colors of galaxies and determine which are red. This is an effort that is also being done currently by a number of groups of the CANDELS collaboration, which are calculating fits for the galaxies' Spectral Energy Distributions (SEDs), i.e., the distributions of energy over the full range of wavelengths for each galaxy, from which one can easily estimate the color of a galaxy. We emphasize that this is being done using our photo-z catalogs and the final results will also be provided to the collaboration in the near future.

Finally, a much larger volume of measurements of galaxy morphologies at higher redshifts will be available in the next decade from a new generation of large surveys such as WFIRST (Spergel et al. (2015)) and EUCLID (Laureijs et al. (2012)); the number of galaxies they will observe is vastly larger than that of CANDELS. With such large samples of data, a more thorough study can be accomplished with finer grids in redshift or in galaxy properties, as well as to investigate the aforementioned color-density relation by dividing galaxies into separate groups according to their color. For such large surveys a reliable classification scheme for the various types of galaxies that can be applied to such large samples is required. This can only be done via computer measurements, given the vast number of objects involved. One such effort is being led by B. Robertson (private communication) using a Neural Network scheme for galaxy morphological classifications. Exciting times are on the way for observational and theoretical Astronomy, following the new discoveries that lie ahead.

Table 1: Details of the origins of the spectroscopic redshifts and grism redshifts used in this study. For each dataset used we provide the name of the survey or instrument used, where applicable; a reference for the source catalog; the number of redshifts provided in each CANDELS field; and any cuts applied in order to restrict to the most robust redshifts. A large portion of the spectroscopic redshifts were taken from catalogs compiled and provided by Nimish Hathi (private communication).

Survey / Instrument (Reference)	Number Of Redshifts in each field					Cuts applied/
	COSMOS	EGS	GOODS−N	GOODS−S	UDS	Flags
Training (Private Communication: mobasher@ucr.edu)	370	–	–	–	–	$z > 0$, flag = 1
	–	840[1]	–	–	–	$z > 0$, flag = none[2]
	–	–	2994	–	–	$z > 0$, flag ≥ 3
	–	–	–	1249	–	$z > 0$, GRISM_FLAG = 0[3]
	–	–	–	–	354	$z > 0$, flag = 1
3D−HST (Momcheva et al. (2016))	566	771	579	523	928	z_max_grism > 0.6 use_zgrism = 1, use_phot = 1 flag1 = 0, flag2 = 0 z_best_s ≠ 0, z_spec ≤ 0 z_phot_u68 − z_phot_l68 > 0 z_grism_u68 − z_grism_l68 < 0.01 $\frac{\text{z_grism_u68} - \text{z_grism_l68}}{\text{z_phot_u68} - \text{z_phot_l68}} < 0.1$ z_max_grism > z_phot_l95 z_max_grism < z_phot_u95

[1] DEEP3, Cooper et al. (2011, 2012b)

[2] High quality redshifts only

[3] Flag to avoid grism-z's and only keep spec-z's

MULTIPLE FIELDS						
DEIMOS (Faber et al. (2003); Private Communication: mobasher@ucr.edu)	172	–	70	39	179	$z > 0$, flag = 1
MOSDEF (Kriek et al. (2015))	189	268	73	10	26	$z > 0$, flag > 0
MOSFIRE (Trump et al. (2013); Wirth et al. (2015))	–	–	22	75	–	$z > 0$, flag ≥ 3
VUDS (Le Fèvre et al. (2015))	101	–	–	77	–	$z > 0$, flag ≥ 3
COSMOS FIELD ONLY						
PRIMUS (Coil et al. (2011))	232	–	–	–	–	$z > 0$, flag > 3
WFC3 (Krogager et al. (2014))	12	–	–	–	–	$z > 0$, flag > 3)
zCOSMOS (Lilly et al. (2007))	7	–	–	–	–	$z > 0$, flag = 1.5, 2.5, 9.3, 9.5 3.x, 4.x, 13.x, 14.x secondary targets
zBRIGHT (Lilly et al. (2009))	2	–	–	–	–	$z > 0$, flag = 1.5, 2.5, 9.3, 9.5 3.x, 4.x, 13.x, 14.x secondary targets
EGS FIELD ONLY						
DEEP2 (Davis et al. (2003); Davis et al. (2007); Newman et al. (2013))	–	1432	–	–	–	$z > 0$, flag ≥ 3
GOODS – N FIELD ONLY						
DEEP3 (Cooper et al. (2011))	–	–	3	–	–	$z > 0$, flag ≥ 3
DEIMOS (Wirth et al. (2004))	–	–	3	–	–	$z > 0$, flag ≥ 3
PEARS/ACT (Pirzkal et al. (2013))	–	–	81	–	–	$z > 0$, flag ≥ 3

GOODS − S FIELD ONLY						
COMBO−17 (Wolf et al. (2004))	–	–	–	6	–	$z > 0$, flag = A
CXO−CDFS (Szokoly et al. (2004))	–	–	–	84	–	$z > 0$, flag = A
K20 (Mignoli et al. (2005))	–	–	–	101	–	$z > 0$, flag = A
VLT_IMAG (Ravikumar et al. (2007))	–	–	–	7	–	$z > 0$, flag = A
VLT_2008 (Vanzella et al. (2008))	–	–	–	147	–	$z > 0$, flag = A
VLT_LBGs (Vanzella et al. (2009))	–	–	–	1	–	$z > 0$, flag = A
VIMOS_2009 (Popesso et al. (2009))	–	–	–	3	–	$z > 0$, flag = 4
VIMOS_2010 (Balestra et al. (2010))	–	–	–	198	–	$z > 0$, flag = A
ACES (Cooper et al. (2012a))	–	–	–	103	–	$z > 0$, flag ≥ 3
VVDS (Le Fèvre et al. (2013))	–	–	–	116	–	$z > 0$, flag > 3
GMASS (Kurk, J. et al. (2013))	–	–	–	49	–	$z > 0$, flag = A
WFC3 (Morris et al. (2015))	–	–	–	54	–	$z > 0$, flag ≥ 3
UDS FIELD ONLY						
UDSz (Bradshaw et al. (2013); McLure et al. (2013))	–	–	–	–	63	$z > 0$, flag = 4 or A
MAGELLAN/IMACS (Santini et al. (2015))	–	–	–	–	61	$z > 0$, flag > 3
SXDS (Akiyama et al. (2015))	–	–	–	–	23	$z > 0$, flag = A

Table 2 Detailed description of the files containing the PDFs from each participant as well as the two best combination methods. The number of models used in the combination methods is reported, as well as the value of the parameter α from the Hierarchical Bayesian method. Note that while the four best participants are included in the evaluation of the combination methods, one or more participants might have missing PDFs for a given object, therefore the total number of PDFs used is not always 4. Both the original and optimized versions of PDFs are provided in separate files.

Column	Description
#1 z	Redshift values for the grid on which PDFs are tabulated
#2 Finkelstein	Probability Density Function (PDF) from Finkelstein
#3 Fontana	PDF from Fontana
#4 Pforr	PDF from Pforr
#5 Salvato	PDF from Salvato
#6 Wiklind	PDF from Wiklind
#7 Wuyts	PDF from Wuyts
#8 HB4	PDF from Hierarchical Bayesian combination, constructed using the PDFs from the best-performing four groups
#9 mFDa4	PDF from the minimum Fréchet Distance combination (computed with ℓ^1 distance metric), constructed using the PDFs from the best-performing four groups

Table 3: Detailed description of the columns of the CANDELS photometric redshift catalogs, which provide point statistics constructed from the optimized photometric PDFs, as well as spectroscopic and/or grism redshifts where available, for all objects in each **CANDELS** field. Each CANDELS object corresponds to one row in the catalog. Statistics based upon the optimized PDFs from all six groups, as well as the two best combination methods, **mFDa4** and **HB4**, are provided within the catalog. The full set of statistics tabulated for the minimum Fréchet distance (mFDa4) PDF are detailed. Corresponding statistics are provided for each groups' results are included in the catalog, with the column identifier only differing in its prefix (i.e., HB4, Finkelstein, Fontana, Pforr, Salvato, Wiklind, or Wuyts) from the column identifier for mFDa4.

Column	Description
# 1 file	Name of the PDF file used to estimate photometric point values.
# 2 ID	CANDELS ID of the object as used in the photometric catalogs.
# 3 RA	Right Ascension of object (from photometric catalog).
# 4 DEC	Declination of object (from photometric catalog).
# 5 z_best	Best redshift value which can be spectroscopic, grism, or photometric.
# 6 z_best_type	Type of photometric redshift: s = spec-z, g = grism-z, p = photo-z.
# 7 z_spec	Spectroscopic redshift if available.
# 8 z_spec_ref	Reference of catalog from which the spectroscopic redshift is obtained.
# 9 z_grism	3D-HST grism redshift of object if available
# 10 mFDa4_z_peak	Peak value of mFDa4 PDF
# 11 mFDa4_z_weight	Weighted average value of mFDa4 PDF

# 12 mFDa4_z683_low	Lower boundary of 68.3% (1σ) credible interval of mFDa4 PDF
# 13 mFDa4_z683_high	Higher boundary of 68.3% (1σ) credible interval of mFDa4 PDF
# 14 mFDa4_z954_low	Lower boundary of 95.4% (2σ) credible interval of mFDa4 PDF
# 15 mFDa4_z954_high	Lower boundary of 95.4% (2σ) credible interval of mFDa4 PDF
# 16 HB4_z_peak	Peak value of HB4 PDF
⋮	
# 22 Finkelstein_z_peak	Peak value of Finkelstein PDF
⋮	
# 28 Fontana_z_peak	Peak value of Fontana PDF
⋮	
# 34 Pforr_z_peak	Peak value of Pforr PDF
⋮	
# 40 Salvato_z_peak	Peak value of Salvato PDF
⋮	
# 46 Wiklind_z_peak	Peak value of Wiklind PDF
⋮	
# 52 Wuyts_z_peak	Peak value of Wuyts PDF
⋮	

Table 4: Detailed description of the columns of the CANDELS morphology catalogs, which provide visual classifications of galaxies as fractions of number of classifiers choosing a particular morphological type divided by the total number of classifiers. These catalogs were compiled by Jeyhan Kartaltepe. Here we show the columns that are of interest to our work only.

Column	Description
Column 1	ID number
Column 2	Right Ascension
Column 3	Declination
⋮	
Column 7	f_Spheroid: Fraction of classifiers that checked Spheroid
Column 8	f_Disk: Fraction of classifiers that checked Disk
Column 9	f_Irr: Fraction of classifiers that checked Irregular
⋮	

Table 5: Table of quantities used to assess the quality of the optimized photometric redshift PDFs and their point statistics. We separately list statistics determined using the training, testing, or 3D-HST spectroscopic redshifts, as well as the average of the results from the testing and 3D-HST sets (statistics computed from the training set can be biased low). Results are provided for the PDFs from all six groups, as well as for the optimal combination method of each type (Hierarchical Bayesian (HB), minimum Fréchet distance computed with an ℓ^1 metric (mFDa), and minimum Fréchet distance computed with an ℓ^2 metric (mFDs). The lowest value in a given row is shown in boldfaced blue font; the lowest average result amongst combination methods is shown in boldface red. The quantities used to evaluate the quality of photo-z PDFs are the normalized ℓ^2-norm between the Q-Q curve for a given set of PDFs and the unity line (ℓ^2-norm) and the fraction of spectroscopic redshifts that lie outside the photo-z PDF for their corresponding object ($f_{\rm op}$). To test the performance of each group and combination method for the weighted-mean peak redshift, $z_{\rm weight}$, we use the normalized Median Absolute Deviation ($\sigma_{\rm NMAD}$) and the $\Delta z/(1+z) > 0.15$ catastrophic outlier fraction $f_{\rm co}$. Finally, we list the fraction of objects for which a PDF file was not provided by a given group as $f_{\rm missing}$ (fraction of missing files). Note that combinations of different numbers of PDFs (either the best three or the best four) yielded best results for PDF statistics (where HB3, mFDa4, and mFDs4 proved superior) than for point statistics (where HB4, mFDa4, and mFDs3 were preferred). Note that $f_{\rm op}$ values for the Hierarchical Bayesian combination method are zero by construction.

Quantity	Set	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts	HB3	mFDa4	mFDs4
ℓ^2 − norm	Training	**1.67×10^{-4}**	4.12×10^{-4}	2.05×10^{-4}	1.82×10^{-4}	5.09×10^{-4}	2.55×10^{-4}	4.91×10^{-4}	4.38×10^{-4}	5.09×10^{-4}
	Testing	5.15×10^{-4}	4.65×10^{-4}	4.72×10^{-4}	3.33×10^{-4}	6.71×10^{-4}	4.73×10^{-4}	9.03×10^{-4}	**2.63×10^{-4}**	2.90×10^{-4}
	3D−HST	8.48×10^{-4}	**5.31×10^{-4}**	2.24×10^{-3}	2.05×10^{-3}	8.68×10^{-4}	7.95×10^{-4}	1.00×10^{-3}	8.65×10^{-4}	9.50×10^{-4}
	Average	6.81×10^{-4}	**4.98×10^{-4}**	1.35×10^{-3}	1.19×10^{-3}	7.69×10^{-4}	6.34×10^{-4}	9.53×10^{-4}	**5.64×10^{-4}**	6.20×10^{-4}

f_{op}	Training	8.89×10^{-2}	9.77×10^{-2}	8.11×10^{-2}	6.38×10^{-2}	$\mathbf{3.24 \times 10^{-2}}$	9.42×10^{-2}	0	6.98×10^{-2}	6.69×10^{-2}
	Testing	1.11×10^{-1}	1.12×10^{-1}	8.91×10^{-2}	7.14×10^{-2}	$\mathbf{2.50 \times 10^{-2}}$	9.98×10^{-2}	0	7.70×10^{-2}	7.08×10^{-2}
	3D–HST	4.02×10^{-2}	7.23×10^{-2}	5.02×10^{-2}	1.87×10^{-2}	1.87×10^{-3}	$\mathbf{1.81 \times 10^{-2}}$	0	2.15×10^{-2}	2.02×10^{-2}
	Average	7.55×10^{-2}	9.22×10^{-2}	6.96×10^{-2}	4.51×10^{-2}	$\mathbf{1.34 \times 10^{-2}}$	5.89×10^{-2}	0	$\mathbf{4.93 \times 10^{-2}}$	4.55×10^{-2}

Quantity	Set	Finkelstein	Fontana	Pforr	Salvato	Wiklind	Wuyts	HB4	mFDa4	mFDs3
σ_{NMAD}	Training	2.77×10^{-2}	3.06×10^{-2}	4.40×10^{-2}	3.31×10^{-2}	3.30×10^{-2}	2.47×10^{-2}	$\mathbf{2.24 \times 10^{-2}}$	2.48×10^{-2}	2.52×10^{-2}
	Testing	2.54×10^{-2}	3.00×10^{-2}	4.08×10^{-2}	2.98×10^{-2}	3.66×10^{-2}	2.41×10^{-2}	$\mathbf{2.22 \times 10^{-2}}$	2.30×10^{-2}	2.36×10^{-2}
	3D–HST	2.24×10^{-2}	2.93×10^{-2}	4.37×10^{-2}	2.74×10^{-2}	3.32×10^{-2}	1.99×10^{-2}	$\mathbf{1.87 \times 10^{-2}}$	2.07×10^{-2}	2.07×10^{-2}
	Average	2.75×10^{-2}	3.42×10^{-2}	4.96×10^{-2}	2.96×10^{-2}	5.33×10^{-2}	2.25×10^{-2}	$\mathbf{2.23 \times 10^{-2}}$	2.43×10^{-2}	2.47×10^{-2}
f_{co}	Training	6.94×10^{-2}	6.65×10^{-2}	8.25×10^{-2}	6.75×10^{-2}	6.71×10^{-2}	6.57×10^{-2}	6.40×10^{-2}	$\mathbf{6.36 \times 10^{-2}}$	6.42×10^{-2}
	Testing	8.72×10^{-2}	7.62×10^{-2}	9.31×10^{-2}	8.18×10^{-2}	9.58×10^{-2}	7.28×10^{-2}	7.17×10^{-2}	$\mathbf{7.11 \times 10^{-2}}$	7.34×10^{-2}
	3D–HST	3.33×10^{-2}	2.18×10^{-2}	3.61×10^{-2}	2.90×10^{-2}	3.15×10^{-2}	2.056×10^{-2}	$\mathbf{1.90 \times 10^{-2}}$	2.09×10^{-2}	2.46×10^{-2}
	Average	6.06×10^{-2}	4.83×10^{-2}	6.69×10^{-2}	5.57×10^{-2}	8.01×10^{-2}	4.73×10^{-2}	$\mathbf{4.56 \times 10^{-2}}$	4.82×10^{-2}	5.07×10^{-2}
f_{missing}	Training	0	1.14×10^{-3}	5.69×10^{-4}	5.69×10^{-4}	7.51×10^{-2}	0	–	–	–
	Testing	0	1.78×10^{-3}	3.30×10^{-3}	7.61×10^{-4}	9.59×10^{-2}	0	–	–	–
	3D–HST	0	2.97×10^{-4}	1.19×10^{-3}	2.97×10^{-4}	4.51×10^{-2}	0	–	–	–

BIBLIOGRAPHY

Abadi, M. G., Moore, B., & Bower, R. G. 1999, MNRAS, 308, 947

Akiyama, M., Ueda, Y., Watson, M. G., et al. 2015, PASJ, 67, 82

Alpaslan, M., Driver, S., Robotham, A. S. G., et al. 2015, MNRAS, 451, 3249

Alt, H., & Godau, M. 1995, Int. J. Comput. Geometry Appl., 5, 75

Arnouts, S., & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate, Astrophysics Source Code Library, ascl:1108.009

Balestra, I., Mainieri, V., Popesso, P., et al. 2010, A&A, 512, A12

Balogh, M. L., Navarro, J. F., & Morris, S. L. 2000, ApJ, 540, 113

Balsara, D., Livio, M., & O'Dea, C. P. 1994, ApJ, 437, 83

Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. 2009, MNRAS, 393, 1324

Bekki, K., & Couch, W. J. 2003, The Astrophysical Journal Letters, 596, L13

Bekki, K., Couch, W. J., Drinkwater, M. J., & Shioya, Y. 2003, MNRAS, 344, 399

Bekki, K., Couch, W. J., & Shioya, Y. 2002, ApJ, 577, 651

Benson, A. J., Bower, R. G., Frenk, C. S., et al. 2003, ApJ, 599, 38

Blakeslee, J. P., Holden, B. P., Franx, M., et al. 2006, ApJ, 644, 30

Blanton, M. R., Eisenstein, D., Hogg, D. W., Schlegel, D. J., & Brinkmann, J. 2005, ApJ, 629, 143

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476

Boselli, A., Boissier, S., Cortese, L., & Gavazzi, G. 2008, The Astrophysical Journal, 674, 742

Boselli, A., & Gavazzi, G. 2006, PASP, 118, 517

Bradshaw, E. J., Almaini, O., Hartley, W. G., et al. 2013, MNRAS, 433, 194

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503

Byrd, G., & Valtonen, M. 1990, ApJ, 350, 89

Calvi, R., Poggianti, B. M., Fasano, G., & Vulcani, B. 2012, MNRAS, 419, L14

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682

Capak, P., Aussel, H., Ajiki, M., et al. 2007, ApJS, 172, 99

Coil, A. L. 2013, The Large-Scale Structure of the Universe, ed. T. D. Oswalt & W. C. Keel, 387

Coil, A. L., Newman, J. A., Croton, D., et al. 2008, ApJ, 672, 153

Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, ApJ, 741, 8

Colless, M., & Dunn, A. M. 1996, ApJ, 458, 435

Cooper, M. C., Newman, J. A., Coil, A. L., et al. 2007, MNRAS, 376, 1445

Cooper, M. C., Aird, J. A., Coil, A. L., et al. 2011, ApJS, 193, 14

Cooper, M. C., Yan, R., Dickinson, M., et al. 2012a, MNRAS, 425, 2116

Cooper, M. C., Griffith, R. L., Newman, J. A., et al. 2012b, MNRAS, 419, 3018

Cowie, L. L., & McKee, C. F. 1977, ApJ, 211, 135

Cowie, L. L., & Songaila, A. 1977, Nature, 266, 501

Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93

Dalgarno, A., & McCray, R. A. 1972, ARA&A, 10, 375

Davis, M., & Peebles, P. J. E. 1983, ApJ, 267, 465

Davis, M., Faber, S. M., Newman, J., et al. 2003, in Proc. SPIE, Vol. 4834, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II, ed. P. Guhathakurta, 161–172

Davis, M., Guhathakurta, P., Konidaris, N. P., et al. 2007, ApJ, 660, L1

de Vaucouleurs, G. 1958, ApJ, 128, 465

—. 1961, ApJS, 5, 233

Diaferio, A., Kauffmann, G., Balogh, M. L., et al. 2001, MNRAS, 323, 999

Dressler, A. 1980, ApJ, 236, 351

Eiter, T., & Mannila, H. 1994

Faber, S. M., Phillips, A. C., Kibrick, R. I., et al. 2003, in Proc. SPIE, Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes, ed. M. Iye & A. F. M. Moorwood, 1657–1669

Fan, X., Strauss, M. A., Becker, R. H., et al. 2006, AJ, 132, 117

Fernández-Soto, A., Lanzetta, K. M., Chen, H.-W., Levine, B., & Yahata, N. 2002, MNRAS, 330, 889

Fioc, M., & Rocca-Volmerange, B. 1997, A&A, 326, 950

Firmani, C., & Avila-Reese, V. 1999, in Observational Cosmology: The Development of Galaxy Systems, ed. G. Giuricin, M. Mezzetti, & P. Salucci, Vol. 176, 406

Firmani, C., & Avila-Reese, V. 2003, in Revista Mexicana de Astronomia y Astrofisica Conference Series, ed. V. Avila-Reese, C. Firmani, C. S. Frenk, & C. Allen, Vol. 17, 107–120

Fontana, A., D'Odorico, S., Poli, F., et al. 2000, AJ, 120, 2206

Freeman, P. E., Izbicki, R., & Lee, A. B. 2017, MNRAS, 468, 4556

Frenk, C. S., & White, S. D. M. 2012, Annalen der Physik, 524, 507

Galametz, A., Grazian, A., Fontana, A., et al. 2013, The Astrophysical Journal Supplement Series, 206, 10

Giallongo, E., D'Odorico, S., Fontana, A., et al. 1998, AJ, 115, 2169

Gnedin, N. Y. 2000, ApJ, 542, 535

Goto, T., Yamauchi, C., Fujita, Y., et al. 2003, MNRAS, 346, 601

Greif, T. H., Johnson, J. L., Klessen, R. S., & Bromm, V. 2008, MNRAS, 387, 1021

Grogin, N. A., Kocevski, D. D., Faber, S. M., et al. 2011, ApJS, 197, 35

Gunn, J. E., & Gott, III, J. R. 1972, ApJ, 176, 1

Guo, Y., Ferguson, H. C., Giavalisco, M., et al. 2013, ApJS, 207, 24

Henriksen, M., & Byrd, G. 1996, ApJ, 459, 82

Hernquist, L., Heyl, J. S., & Spergel, D. N. 1993, ApJ, 416, L9

Hildebrandt, H., Wolf, C., & Benítez, N. 2008, A&A, 480, 703

Hogg, D. W. 1999, ArXiv Astrophysics e-prints, astro-ph/9905116

Holmberg, E. 1958, Meddelanden fran Lunds Astronomiska Observatorium Serie II, 136, 1

Hubble, E. P. 1926, ApJ, 64, doi:10.1086/143018

—. 1936, Realm of the Nebulae

Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841

Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236

Ivezic, Z., Tyson, J. A., Axelrod, T., et al. 2009, in Bulletin of the American Astronomical Society, Vol. 41, American Astronomical Society Meeting Abstracts #213, 366

Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338

Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, ApJS, 221, 11

Kawinwanichakij, L., Papovich, C., Quadri, R. F., et al. 2017, ApJ, 847, 134

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, ApJS, 197, 36

Kraljic, K. B. 2014, Theses, Université Paris Sud - Paris XI

Kriek, M., Shapley, A. E., Reddy, N. A., et al. 2015, ApJS, 218, 15

Krogager, J.-K., Zirm, A. W., Toft, S., Man, A., & Brammer, G. 2014, ApJ, 797, 17

Kurk, J., Cimatti, A., Daddi, E., et al. 2013, A&A, 549, A63

Lahav, O., Bridle, S. L., Percival, W. J., et al. 2002, MNRAS, 333, 961

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64

Lang, D., & Hogg, D. W. 2012, AJ, 144, 46

Larson, R. B., Tinsley, B. M., & Caldwell, C. N. 1980, ApJ, 237, 692

Laureijs, R., Gondoin, P., Duvet, L., et al. 2012, in Proc. SPIE, Vol. 8442, Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave, 84420T

Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, A&A, 559, A14

Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. 2015, A&A, 576, A79

Licquia, T. C., Newman, J. A., & Brinchmann, J. 2015, The Astrophysical Journal, 809, 96

Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007, ApJS, 172, 70

Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179

Livio, M., Regev, O., & Shaviv, G. 1980, ApJ, 240, L83

LSST Dark Energy Science Collaboration. 2012, ArXiv e-prints, arXiv:1211.0310

Maller, A. H., & Dekel, A. 2002, MNRAS, 335, 487

Maraston, C. 2005, MNRAS, 362, 799

Mayer, L. 2010, Advances in Astronomy, 2010, 278434

McLure, R. J., Pearce, H. J., Dunlop, J. S., et al. 2013, MNRAS, 428, 1088

Meier, D. L. 1999, ApJ, 522, 753

—. 2001, ApJ, 548, L9

Merritt, D. 1984, ApJ, 276, 26

Mignoli, M., Cimatti, A., Zamorani, G., et al. 2005, A&A, 437, 883

Miller, R. H. 1986, A&A, 167, 41

Mo, H., van den Bosch, F. C., & White, S. 2010, Galaxy Formation and Evolution

Momcheva, I. G., Brammer, G. B., van Dokkum, P. G., et al. 2016, ApJS, 225, 27

Moore, B., Katz, N., Lake, G., Dressler, A., & Oemler, A. 1996, Nature, 379, 613

Moore, B., Lake, G., & Katz, N. 1998, ApJ, 495, 139

Moore, B., Lake, G., Quinn, T., & Stadel, J. 1999, MNRAS, 304, 465

Moran, S. M., Ellis, R. S., Treu, T., et al. 2007, The Astrophysical Journal, 671, 1503

Morris, A. M., Kocevski, D. D., Trump, J. R., et al. 2015, AJ, 149, 178

Mulchaey, J. S., Dressler, A., Oemler, A., & Willis, J. 2005, The Observatory, 125, 282

Nayyeri, H., Hemmati, S., Mobasher, B., et al. 2017, The Astrophysical Journal Supplement Series, 228, 7

Nepveu, M. 1981, A&A, 101, 362

Newman, J. A., Zepf, S. E., Davis, M., et al. 1999, The Astrophysical Journal, 523, 506

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5

Nipoti, C. 2017, MNRAS, 467, 661

Nulsen, P. E. J. 1982, MNRAS, 198, 1007

Okamoto, T., & Nagashima, M. 2001, ApJ, 547, 109

Okamoto, T., & Nagashima, M. 2003, in Revista Mexicana de Astronomia y Astrofisica, vol. 27, Vol. 17, Revista Mexicana de Astronomia y Astrofisica Conference Series, ed. V. Avila-Reese, C. Firmani, C. S. Frenk, & C. Allen, 98–99

Peebles, P. J. E. 1980, The large-scale structure of the universe

Pforr, J., Maraston, C., & Tonini, C. 2013, MNRAS, 435, 1389

Pirzkal, N., Rothberg, B., Ly, C., et al. 2013, ApJ, 772, 48

Poggianti, B. M., Desai, V., Finn, R., et al. 2008, ApJ, 684, 888

Popesso, P., Dickinson, M., Nonino, M., et al. 2009, A&A, 494, 443

Postman, M., & Geller, M. J. 1984, ApJ, 281, 95

Press, W. H. 1997, in Unsolved Problems in Astrophysics, ed. J. N. Bahcall & J. P. Ostriker, 49–60

Quilis, V., Moore, B., & Bower, R. 2000, Science, 288, 1617

Quinn, T., Katz, N., & Efstathiou, G. 1996, Monthly Notices of the Royal Astronomical Society, 278, L49

Ravikumar, C. D., Puech, M., Flores, H., et al. 2007, A&A, 465, 1099

Rines, K., Geller, M. J., Kurtz, M. J., & Diaferio, A. 2003, AJ, 126, 2152

Santini, P., Ferguson, H. C., Fontana, A., et al. 2015, ApJ, 801, 97

Sarazin, C. L. 2009, X-Ray Emission from Clusters of Galaxies

Schawinski, K., Thomas, D., Sarzi, M., et al. 2007, MNRAS, 382, 1415

Schawinski, K., Urry, C. M., Simmons, B. D., et al. 2014, MNRAS, 440, 889

Schneider, P. 2006, in Extragalctic Astronomy and Cosmology, An Introduction (Springer Berlin Heidelberg New York)

Schulz, S., & Struck, C. 2001, MNRAS, 328, 185

Sheth, R. K., Jimenez, R., Panter, B., & Heavens, A. F. 2006, ApJ, 650, L25

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, MNRAS, 464, 4420

Skibba, R. A., Sheth, R. K., Croton, D. J., et al. 2013, MNRAS, 429, 458

Spergel, D., Gehrels, N., Baltay, C., et al. 2015, ArXiv e-prints, arXiv:1503.03757

Springel, V., White, S. D. M., Tormen, G., & Kauffmann, G. 2001, MNRAS, 328, 726

Stefanon, M., Yan, H., Mobasher, B., et al. 2017, ApJS, 229, 32

Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861

Szokoly, G. P., Bergeron, J., Hasinger, G., et al. 2004, ApJS, 155, 271

Tasca, L. A. M., Kneib, J.-P., Iovino, A., et al. 2009, A&A, 503, 379

Treu, T., Ellis, R. S., Kneib, J.-P., et al. 2003, ApJ, 591, 53

Trump, J. R., Konidaris, N. P., Barro, G., et al. 2013, ApJ, 763, L6

van den Bergh, S. 1960, ApJ, 131, 215

van der Wel, A. 2008, ApJ, 675, L13

Vanzella, E., Cristiani, S., Dickinson, M., et al. 2008, A&A, 478, 83

Vanzella, E., Giavalisco, M., Dickinson, M., et al. 2009, ApJ, 695, 1163

Vollmer, B., Braine, J., Balkowski, C., Cayatte, V., & Duschl, W. J. 2001, A&A, 374, 824

Vollmer, B., Cayatte, V., Balkowski, C., & Duschl, W. 2000, in Astronomical Society of the Pacific Conference Series, Vol. 221, Stars, Gas and Dust in Galaxies: Exploring the Links, ed. D. Alloin, K. Olsen, & G. Galaz, 265

Weiner, B. J., Phillips, A. C., Faber, S. M., et al. 2005, ApJ, 620, 595

White, S. D. M. 1978, MNRAS, 184, 185

Wiklind, T., Dickinson, M., Ferguson, H. C., et al. 2008, ApJ, 676, 781

Wilk, M. B., & Gnanadesikan, R. 1968, Biometrika, 55, 1

Wirth, G. D., Willmer, C. N. A., Amico, P., et al. 2004, AJ, 127, 3121

Wirth, G. D., Trump, J. R., Barro, G., et al. 2015, AJ, 150, 153

Wittman, D., Bhaskar, R., & Tobin, R. 2016, MNRAS, 457, 4005

Wolf, C., Meisenheimer, K., Kleinheinrich, M., et al. 2004, A&A, 421, 913

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579