Astronomy & Astrophysics manuscript no. ms_pretty
June 25, 2018

Tests of Catastrophic Outlier Prediction in Empirical Photometric Redshift Estimation with Redshift Probability Distributions

E. Jones⋆, and J. Singal

Physics Department, University of Richmond
28 Westhampton Way, University of Richmond, VA 23173

June 25, 2018

ABSTRACT

Aims. We present results of using individual galaxies' redshift probability information derived from a photometric redshift (photo-z) algorithm, SPIDERz, to identify potential catastrophic outliers in photometric redshift determinations. By using two test data sets comprised of COSMOS multi-band photometry spanning a wide redshift range ($0 < z < 4$) matched with reliable spectroscopic or other redshift determinations we explore the efficacy of a novel method to flag potential catastrophic outliers (those galaxies where $|z_{phot} - z_{spec}| > 1.0$) in an analysis which relies on accurate photometric redshifts.

Methods. SPIDERz is a custom support vector machine classification algorithm for photo-z analysis that naturally outputs a distribution of redshift probability information for each galaxy in addition to a discrete most probable photo-z value. By applying an analytic technique with flagging criteria to identify the presence of probability distribution features characteristic of catastrophic outlier photo-z estimates, such as multiple redshift probability peaks separated by substantial redshift distances, we can flag potential catastrophic outliers in photo-z determinations.

Results. We find that our proposed method can correctly flag large fractions of the catastrophic outlier (>50%) galaxies, while only flagging a small fraction (<5%) of the total non-outlier galaxies, depending on parameter choices. The fraction of non-outlier galaxies flagged varies significantly with redshift and magnitude, however. We examine the performance of this strategy in photo-z determinations using a range of flagging parameter values. These results could potentially be useful for utilization of photometric redshifts in future large scale surveys where catastrophic outliers are particularly detrimental to the science goals.

Key words. techniques: photometric - galaxies: statistics - methods: miscellaneous

1. Introduction

Accurate photometric redshift estimates (photo-zs) with well constrained and understood error properties are critical for the current and coming era of large multi-band extra-galactic surveys (e.g. Huterer et al. 2006; Hearin et al. 2010; Bernstein & Huterer 2010), such as the Large Synoptic Survey Telescope (LSST)[1], Euclid[2], Wide Field Infrared Survey Telescope (WFIRST)[3], Hyper-Suprime Cam (HSC)[4], and Kilo-Degree Survey (KiDS)[5] for which precise redshift estimates will be needed for millions or billions of galaxies extending to high redshifts. In particular, photometric redshift accuracy is the primary source of systematic error in weak-lensing surveys (Bernstein & Huterer 2010). Works modeling the error relation between photometric (z_{phot}) and spectroscopic (z_{spec}) redshifts as a Gaussian have found that achieving less than 50% degredation in cosmological parameter uncertainties requires the bias $\langle z_{phot} - z_{spec} \rangle$ and scatter $\langle (z_{phot} - z_{spec})^2 \rangle^{1/2}$ quantities in each redshift bin $\Delta z = 0.1$ to be constrained to roughly 0.003-0.01 (Ma et al. 2006; Huterer et al. 2006; Kitching et al. 2008) with tighter constraints when these distributions are non-Gaussian.

Limiting the occurrence of catastrophic outlier photo-z estimates — those galaxies whose estimated redshift differs substantially from their actual redshift — is a top priority for controlling photo-z errors. In addressing this challenge we present a study directed toward a novel method to flag potential catastrophic outlier photo-z predictions through the utilization of individual galaxy redshift probability information. We utilize SPIDERz (SuPport vector classification for IDEntifying Redshifts — Jones & Singal 2017), a custom implementation of a support vector machine classification model for photometric redshift analysis, which naturally outputs an effective redshift probability distribution for each galaxy[6]. SPIDERz's natural output of an effective redshift probability distribution for each galaxy is not necessarily typical for empirical photo-z estimation methods (which make a predictive model based on a training set with known redshifts), but some other empirical methods which can output probability information are ArborZ (Gerdes et al. 2010), TPZ (Carrasco Kind & Bruner 2013), SkyNet (Bonnett 2015), ANNz2 (Sadeh et al. 2016). The techniques discussed in this work should theoretically be relevant to any photo-z estimation method which provides the requisite redshift probability distribution information for individual galaxies.

The performance of candidate photo-z methods should ideally be demonstrated on test data that is representative of the data anticipated by future large-scale surveys. In particular, some data

⋆ evan.jones@richmond.edu
[1] http://www.lsst.org
[2] http://sci.esa.int/euclid
[3] http://wfirst.gsfc.nasa.gov
[4] http://www.naoj.org/Projects/HSC
[5] http://kids.strw.leidenuniv.nl

[6] available from http://spiderz.sourceforge.net with usage documentation provided there.

sets, such as much of the LSST catalog, will have photometric data for optical bands only while others, such as Euclid, will have, or overlap with, infrared bands. Additionally, some important data sets will span a large redshift range with many high redshift objects. In order to perform an analysis on real data approximating these conditions, we use two relatively large data sets of photometry from the Cosmic Evolution Survey (COSMOS) COSMOS2015 photometric catalog (Laigle et al. 2016) with known redshifts spanning the redshift range $0 < z < 4$. One set consists of the overlap of COSMOS photometry with spectroscopic redshifts from the 3D-HST survey performed with the Hubble Space Telescope and reported in Momcheva et al. (2016) featuring 3704 galaxies, and the other consists of COSMOS photometry with previously reliably estimated (see criteria in §4.1) 30-band photometric redshifts. Furthermore, in order to approximate the photometric redshift conditions of future large scale surveys, we adopt training set sizes that are much smaller than evaluation set sizes.

Photo-z methods have been traditionally divided into two categories: template-fitting and empirical methods. Template-fitting methods rely on fitting galaxy photometry to template spectra evolved with redshift, typically derived using χ^2 minimization, e.g. Le Phare (Arnouts et al. 1999; Ilbert et al. 2006), BPZ (Benítez 2000), HyperZ (Bolzonella et al. 2000), zebra (Feldmann et al. 2006), EAZY (Brammer et al. 2008), gazelle (Kotulla & Fritze 2009), and DELIGHT (Leistedt & Hogg 2017). Template-fitting methods depend critically on the extent to which galaxy spectral energy distributions (SEDs) library templates adequately represent properties of observed SEDs corresponding to target galaxy populations for which one wants to estimate the redshifts; the selection of ill-fitted SED templates provides the greatest source of errors in redshift determinations with these models. Some techniques for template fitting have incorporated the use of training sets of objects with known photometry and spectroscopic redshifts to better calibrate representative SED templates (Benítez et al. 2004; Ilbert et al. 2006, 2009).

Empirical methods, which rely on training sets with known redshifts to derive a mapping from photometry to redshift, depend critically on the extent to which training galaxy populations adequately represent target galaxy populations in terms of the parameter overlap of photometric inputs and true redshift distributions. Early examples empirical photo-z methods utilized relatively simple techniques to achieve such a mapping (e.g. polynomial fitting, Connolly et al. 1995). More recently, models that produce mappings with greater complexity utilizing machine learning have been examined, e.g. artificial neural networks (Firth et al. 2003; Collister & Lahav 2004; Vanzella et al. 2004; Singal et al. 2011; Brescia et al. 2014; Sadeh et al. 2016), support vector machines (Wadadekar 2004; Wang et al. 2007; Jones & Singal 2017), Gaussian process regression (Way & Srivastava 2006), boosted decision trees (Gerdes et al. 2010), random forests (Carrasco Kind & Bruner 2013; Rau et al. 2015), genetic algorithms (Hogan et al. 2015), sparse Gaussian framework (Almosallam et al. 2016), nearest neighbor search (Ball et al. 2007, 2008), and spectral connectivity analysis (Freeman et al. 2009). A review and comparison of a number of existing photo-z methods can be found in Hildebrandt et al. (2010); Abdalla et al. (2011); Sanchez et al. (2014).

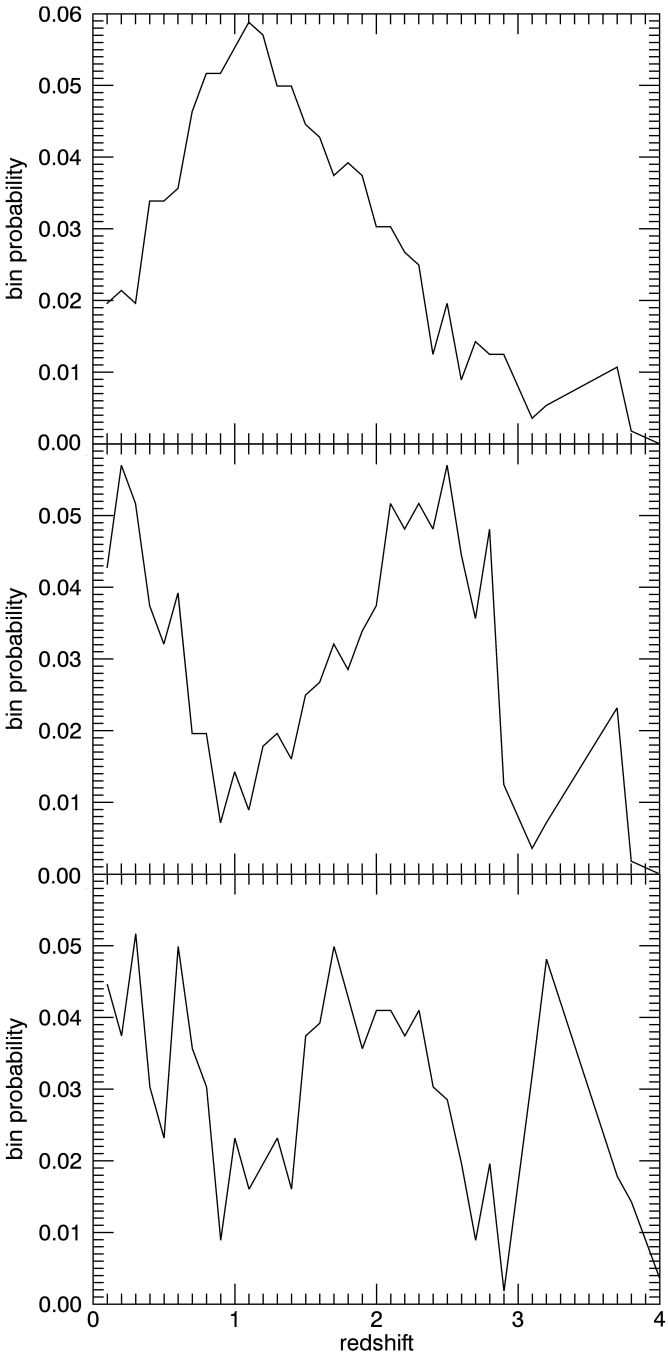

Fig. 1: Examples of EPDFs as determined by SPIDERz for particular individual galaxies in the COSMOSx3D-HST data set described in §4.1. The top panel shows an EPDF with a singular uniform probability peak, which is typical of galaxies with accurate redshift estimates. The middle panel shows a classic doubly peaked EPDF where the spectroscopic redshift is near the slightly lower peak, which is often the case for catastrophic outlier redshift estimates. The bottom panel shows an EPDF without a clear probability peak, which also can be the case for catastrophic outlier redshift estimates.

Here we follow convention (e.g. Hildebrandt et al. 2010) and define "outliers" as those galaxies where

$$Outliers : \frac{|z_{phot} - z_{spec}|}{1 + z_{spec}} > .15, \quad (1)$$

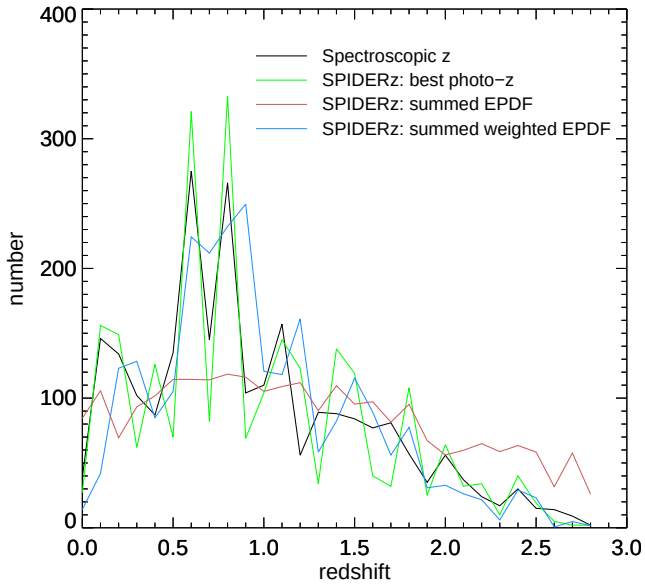

Fig. 2: Reconstructed redshift distributions from a determination with SPIDERz using 1200 training galaxies compared to the actual COSMOSx3D-HST evaluation sample of 2323 galaxies. Test data for the determination shown in this figure only were limited to $z < 2.9$ to prevent the occurrence of unoccupied redshift bins at high redshifts. Distributions are shown for the actual spectroscopic redshift, the single best-estimate (highest probability bin) photo-z, the summed EPDF, and the weighted summed EPDF.

where z_{phot} and z_{spec} are the estimated photo-z and actual (spectroscopically determined) redshift of the object. Although there is not a standard, universal definition of "catastrophic outliers" we use a definition that is typical (e.g. Bernstein & Huterer 2010)

$$O_c : |z_{phot} - z_{spec}| > 1.0. \quad (2)$$

The RMS photo-z error in a realization is given by a standard definition

$$\sigma_{\Delta z/(1+z)} \equiv \sqrt{\frac{1}{n_{gals}} \Sigma_{gals} \left(\frac{z_{phot} - z_{spec}}{1 + z_{spec}} \right)^2}, \quad (3)$$

where n_{gals} is the number of galaxies in the evaluation testing set and Σ_{gals} represents a sum over those galaxies. We also calculate the RMS error without the inclusion of outlier galaxies, referring to this quantity as the "reduced" RMS or R-RMS.

In §2 we present a summary overview of the SVM model implemented in SPIDERz and discuss the probability information produced for each galaxy. In §3 we present a method for flagging potential catastrophic outlier photo-z estimates made by SPIDERz through the utilization of redshift probability information. In §4 we discuss the results of testing SPIDERz on the two test data sets utilizing COSMOS multi-band photometry. We present a discussion in §5.

2. SPIDERz and effective probability distributions

A full discussion of the SPIDERz algorithm, mathematical theory, and a suite of tests with various data sets and comparisons with other photo-z determination methods is available in Jones & Singal (2017). Here, we will provide a brief outline of the machine learning photo-z process for context, but we primarily focus on the utilization of the naturally available probability information for each galaxy produced during photo-z evaluations with SPIDERz. The general technique we propose in this work for utilizing the probability information, however, should theoretically be relevant to any photo-z estimation method which provides the requisite probability information for individual galaxies.

Generally speaking, machine learning photo-z codes perform two main processes: training and evaluation. The output of the training process is a mapping from band magnitudes (and potentially additional information) to redshift. The collection of mappings comprise a predictive model that can be used to make photo-z predictions on evaluation galaxies.

SPIDERz utilizes support vector classification to make photo-z predictions, where bins of redshift are assigned class labels, and photo-z estimation is performed via the solution to a multi-class classification problem. SPIDERz solves the multi-class problem with a "one against one" or "pairwise coupling" approach that treats the complex multi-class problem as a series of simpler binary class problems consisting of every possible pairing of classes (in this case redshift bins). Thus for a system comprised of m distinct classes (m redshift bins in this case), SPIDERz formulates and solves $\frac{m(m-1)}{2}$ separate binary classification problems, choosing the more likely class (redshift bin) in each binary pairing. Each instance of classification in favor of a particular redshift bin can be regarded as a 'vote' for that class. The entire collection of $\frac{m(m-1)}{2}$ votes forms a distribution (see Figure 1 for examples) that we call an 'effective' probability distribution (EPDF) for each galaxy, with the relative probability of each redshift bin proportional to the number of times the corresponding class was chosen as the best binary solution. This EPDF is not continuous, but rather is resolved to the bin-width level. Discrete z_{phot} estimates, if they are desired, can be obtained for each galaxy by simply taking the redshift bin with the highest number of votes.

Examples of actual EPDFs for individual galaxies in the COSMOSx3D-HST data set described in §4.1 are shown in Figure 1. The top panel shows the presence a uniform singular probability peak characteristic of typical cases where $z_{phot} \approx z_{spec}$. The middle and bottom panels show distributions with multiple peaked probabilities throughout wide redshift distances, which is a feature that is typical of many inaccurate z_{phot} estimates.

We use the terminology "effective PDF" because of the way that all bins are used in comparison, thus artificially inflating low probability bins due to the inevitable pairwise comparisons of two low probability bins. However the overall shape of the EPDFs, in regard to higher probability bins which are the only ones relevant in this analysis, approaches that of a true probability distribution.

To breifly illustrate how the EPDF compares to a true probability distribution function, if one desired to mitigate the effect of low probability bin inflation in the EPDFs for comparisons between the summed EPDF for all galaxies and the known $N(z)$ distribution in testing determinations, one would apply weights to the EPDFs that are proportional to the fractional population of training galaxies in each redshift bin relative to the total training galaxy population. Weights are determined for each redshift bin Δz_i by

$$w_i = \frac{N(\Delta z_i)}{N} \quad (4)$$

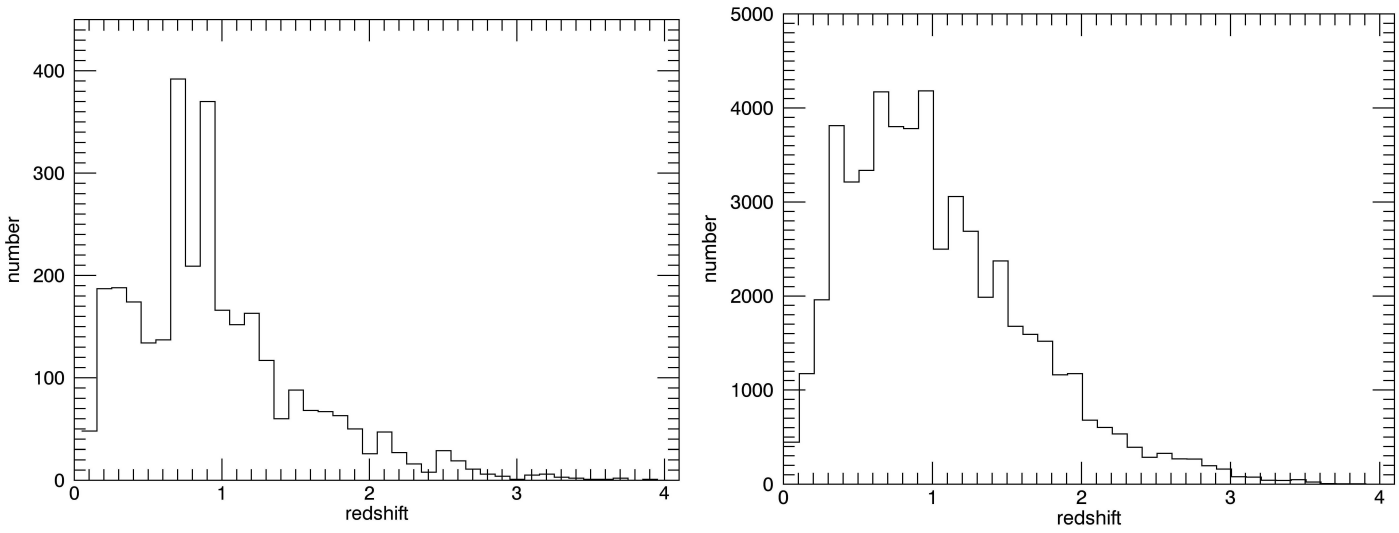

Fig. 3: $N(z)$ distribution for the 3704 galaxies comprising the COSMOSxHST (left) and 58622 galaxies comprising the COSMOS2015 (right) test data sets used in this analysis.

$$\sum_{i=1}^{l} w_i = 1, \tag{5}$$

where $N = \sum_{i}^{l} N(\Delta z_i)$ and l is the number of redshift bins. Weights are applied to the EPDF by

$$P_w(\Delta z_i) = l(\sum_{i=1}^{l} P_{Eff}(\Delta z_i) * w_i)), \tag{6}$$

where $P_w(\Delta z_i)$ is the weighted probability for some redshift bin Δz_i, and $P_{Eff}(\Delta z_i)$ is the probability given by the unweighted EPDF. In this way, as shown in Figure 2, we can see that there is meaningful probability information in the EPDFs and that they can be made, in aggregate, to approach a true probability distribution with weighting. For the present work, however, the degree of fidelity of the EPDFs to true probability distribution functions is not important, as only the highest probabilty bins are relevant, and so no weighting is applied — the analyses in this work simply use the raw EPDFs as output by SPIDERz. The reason for this is severalfold: Firstly, we would like to demonstrate the method of this work with the raw output of a machine learning classifier, for the simplest, most general situation. Further, while it is the case that in the analyses here the training set and the evaluation set have practically the same redshift distribution, that is not necessarily the case for all generic photo-z evaluations going forward, so weighting the individual output galaxy probabilities by the particular redshift distribution of the training set may not be appropriate. Additionally, in this work we are focusing on the utility of individual galaxy probability functions. If one were to weight those functions individually by the cumulative redshift distribution of a given single training set, the amount that high probability peaks are scaled up and down would be highly dependent on the particulars of that training set, and would be different for another training set; therefore values investigated quantitatively here would be entirely training set dependent, and certain training sets would result in a weighting where no individual galaxies have high probability peaks at high redshifts.

We note that to produce Figure 2, due to the relatively limited population of galaxies at high redshifts in the COSMOSx3D-HST data set used in this analysis, the presence of unpopulated redshift bins at high z in a training set is often unavoidable. So in order to present a useful comparison between the summed EPDFs and distribution of discrete most probable estimates $N(z_{phot})$ produced in SPIDERz determinations with the actual redshift distribution $N(z_{spec})$ for this particular data set we utilized a subset of test data galaxies restricted to $z < 2.9$, ensuring all redshift bins are populated, for this particular calculation only.

By default, SPIDERz chooses the most probable (commonly occurring) redshift bin as a single valued photo-z estimate for the galaxy. In this analysis we use this method for discrete photo-z predictions, such as those shown in Figure 4. In this work we seek a method to identify potential catastrophic outliers in such photo-z predictions.

SPIDERz also allows users flexibility in redshift bin size. We generally find determinations have increased accuracy and precision when smaller bin sizes are used, however the optimal bin size for any determination will be dependent on the size and nature of the training set (decreasing the bin size for determinations lowers existing parameter overlap between training and evaluation sets) and can be approached via trial-and-error or approximated with the bin size introduced as an additional parameter in a grid search (see a detailed discussion in Jones & Singal 2017).

3. Strategy for identifying potential catastrophic outliers with EPDFs

To identify potential catastrophic outlier photo-z estimates we focus on the existence of individual galaxies' EPDFs displaying multiple probability peaks or, somewhat equivalently, a 'weak' primary probability peak. There is some ambiguity in what constitutes multiple substantial probability peaks in a galaxy's EPDF. In particular, a secondary peak is more likely to be significant if it is closer in height (probability) to the primary (highest probability) peak, and also if it is located farther away in redshift from the primary peak. Let us denote the ratio of the probability

of a secondary peak i to the primary peak in a galaxy's EPDF as

$$p_f = \frac{p_i}{p_{max}}, \tag{7}$$

where p_{max} is the probability of the primary (highest probability) peak, and let us also denote the redshift distance between that secondary peak and the primary peak as Δz_{peak}. Thus a designated minimum value for p_f ($p_{f,min}$), and a designated minimum value for Δz_{peak} ($\Delta z_{peak,min}$), can serve as filter values above which a multiply peaked EPDF is flagged. If at least one redshift bin in an EPDF distribution satisfies both of the $\Delta z_{peak} > \Delta z_{peak,min}$ and $p_f > p_{f,min}$ criteria, the galaxy is flagged as a potential catastrophic outlier. The optimal values for $p_{f,min}$ and $\Delta z_{peak,min}$ will vary depending on factors such as the redshift range of test data and designated bin size, and the relative importance of flagging more catastrophic outliers versus avoiding spurious flaggings.

The simplest way to deal with flagged galaxies would be to remove them from analyses which rely on photo-zs. This would, of course, remove some fraction of catastrophic outliers and other outliers, along with some fraction of non-outliers. In §4.2 we show that the former number can be relatively high and the latter relatively low. In this analysis going forward we consider flagging being somewhat equivalent to removal from consideration, while acknowledging that other strategies, such as de-weighting while not completely eliminating flagged galaxies in analyses, are possible and likely desirable in some circumstances.

4. Results

In this section, we present the results from our study of using EPDFs to identify probable outlier and catastrophic outlier galaxy estimates as discussed in §3. We begin with a discussion of the two test data sets used in these photo-z analyses. Next we provide results from photo-z determinations performed with SPIDERz on the test data sets — both with and without application of the EPDF outlier identification method discussed in §3. Metrics of performance of this method are provided for a range of values for the identification criteria, assuming here a simple removal of flagged galaxies.

4.1. Test Data Sets

To obtain a data set of real galaxies with publicly available spectroscopic redshifts containing sources throughout a large redshift range including higher redshifts we use spectroscopic redshifts from the 3D-HST survey performed with the Hubble Space Telescope and reported in Momcheva et al. (2016) that overlap with photometry from the COSMOS2015 photometric catalog (Laigle et al. 2016) which reports photometry for over half a million objects in the COSMOS field (Scoville et al. 2007). For spectroscopic redshifts we use the reported "best available" redshift measurement and eliminate those flagged as having their redshift obtained from photometry or as being stars. This results in a data set of 3704 galaxies, of which 383 (10.3%) have $z > 2$ and 948 (25.6%) have $z > 1.5$. The $N(z)$ distribution for this data set is shown in Figure 3. These data span an i-band magnitude range from 27.05 to 18.16 with a median of 23.74.

In order to form an additional test set with a significantly larger number of real galaxies, we also utilize galaxies from the COSMOS2015 photometric catalog that contain particularly reliable, previously estimated photometric redshifts derived from a large number of photometric bands. As the COSMOS2015 catalog provides photometry for some galaxies in up to 31 optical, infrared, and UV bands, those galaxies with (i) magnitude values for at least 30 bands of photometry, and (ii) for which the stated χ^2 for the redshift estimate is < 1, and (iii) for which the stated photo-z value from the minimum χ^2 estimate is less than 0.1 redshift away from the stated photo-z value from the peak of the pdf, can be considered to have highly reliable previous redshift estimates. Applying these criteria result in a data set of 58622 galaxies spanning an i band magnitude range from 27.17 to 19.00 with a median of 24.08. For shorthand purposes we will

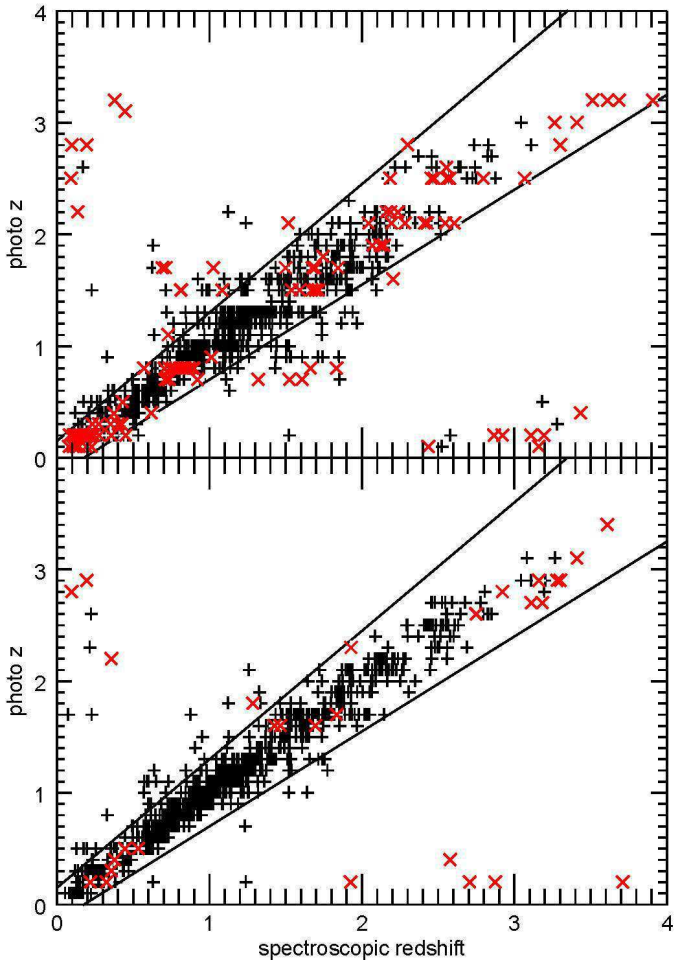

Fig. 4: The best discrete photo-z estimation (most probable redshift, as discussed in §2) as determined by SPIDERz versus the actual redshift for the COSMOSx3D-HST data set discussed in §4.1 for a realization of the five-band (*top*) and ten-band (*bottom*) cases. The catastrophic outlier identification method discussed in §3 was employed for these determinations with the $\Delta z_{peak,min} = 0.6$ criteria and the flagged galaxies are shown by red crosses. These determinations were performed with a training set consisting of 1200 galaxies chosen at random and an evaluation testing set consisting of the other 2504 galaxies. A bin size of 0.1 was used. Outliers in a determination are defined by equation 1, shown as those points lying outside of the two diagonal lines. The density of points within the lines is quite high — only 2.6% of points lie outside of the lines as outliers for the ten-band case (BOTTOM) before flagging and 6.7% for the five-band case (TOP).

refer to this set here as the "COSMOS-reliable-z" test data set. The $N(z)$ distribution for this data set is also shown in Figure 3.

Although the COSMOS2015 catalog provides photometry in a potentially large number of optical, infrared, and UV bands, we choose to restrict our test analyses to the u, B, V, r, i, z+, Y, H, J, and Ks bands, and a subset of five of these bands, because with data sets approaching 30 bands of photometry, the distinction between photo-z estimation and spectroscopic redshift determination is somewhat muddled, and in any case this does not represent a realistic photometric situation for upcoming large surveys such as LSST, even for subsets which would have infrared survey overlap. In the following sections we refer to test data consisting of only five optical bands (u, B, r, i, z+) as the 'five-band case', which could resemble the default situation for obtaining photometric redshifts from a very large optical survey, and similarly refer to test data comprised of all ten aforementioned bands as the 'ten-band case', which could resemble the situation for obtaining photometric results from a large optical survey that overlaps a large near-infrared survey. For these bands we use aperture magnitudes measured in a 3" aperture. The depths of the photometry for the bands are given in Table 1 of Laigle et al. (2016). We have not utilized galaxies with missing photometry values in these bands — for the COSMOSx3D-HST test set the number of galaxies where this is the case is negligible, while for the COSMOS-reliable-z test set applying this filter has almost no effect since this data set by definition contains 30 reliable bands of photometry.

Unless otherwise noted, all determinations are performed with randomly selected training and testing set populations of 1200 and 2504 galaxies, respectively for the COSMOSx3D-HST data set, and 5000 and 53622 galaxies for the COSMOS-reliable-z data set. Increasing the training population size beyond 1200 for the COSMOSx3D-HST data set produced only marginal improvements in photo-z accuracy. For the COSMOS-reliable-z data set we chose to maintain a training set to evaluation set size ratio of below 1:10 to more closely approximate the photo-z conditions of future large scale survey analyses than would be achieved with doing analyses with larger ratios.

We note that the galaxies in these data sets span the largest redshift range of publicly available real galaxy photo-z test data with photometry down to these magnitudes of which we are aware. We also note that a significant limitation is posed on the performance accuracy of SPIDERz due to inadequate parameter overlap between training and evaluation galaxies in sparsely populated redshift regions, which, among other restrictive influences, imposes a lower limit on the redshift bin size that can be effectively used.

4.2. Results for various parameter choices

Figure 4 displays the estimated SPIDERz photo-z versus actual redshift for an example of typical determinations with the five-band and ten-band cases for the COSMOSx3D-HST data set discussed in §4.1. The EPDF outlier identification method discussed in §3 was then employed for these determinations with particular flagging parameters $\Delta z_{peak,min} = 0.6$ and $p_{f,min} = 0.95$. Red data points indicate flagged potential catastrophic outlier estimates in these cases. Estimates with the ten-band case are of course significantly better than with the five-band case.

To examine the influence of our proposed method for flagging potential catastrophic outliers in photo-z determinations, we performed an extensive analysis with test determinations on the five-band and ten-band cases for both the COSMOSx3D-HST and COSMOS-reliable-z data sets using a range of $\Delta z_{peak,min}$ values and $p_{f,min}$ values, redshift bin sizes, and training population sizes.

Perhaps surprisingly, we determine that appropriate values of $p_{f,min}$ are quite high, with any values below $p_{f,min} = 0.9$ resulting in an unacceptably large number of spurious flaggings. We find variations in the designated value for $\Delta z_{peak,min}$ greatly influence the performance of the outlier identification method, as measured by the relative numbers of correct outlier identifications versus spurious removal of non-outliers, however variations in $p_{f,min}$ produced marginal difference in the range $0.90 \leq p_{f,min} \leq 0.98$.

We also find that discrete photo-z accuracy is generally highest on this test data when using redshift bin sizes between 0.1 and 0.05; the use of larger bin sizes significantly reduced photo-z precision across all z values and particularly at lower zs, as expected, while the use of bin sizes less than 0.05 produced a significant number of unoccupied bins at higher redshifts and deteriorated parameter overlap between training and evaluation sets.

Figures 5 and 7 and Tables 1 and 2 show various performance metrics from determinations with SPIDERz using the EPDF outlier identification method on COSMOSx3D-HST test data. Table 1 highlights the percentage of outliers, percentage of outliers removed, percentage of removed galaxies that are outliers, percentage of non-outliers removed, percentage of catastrophic outliers removed, and finally the percentage of catastrophic outliers remaining for determinations on five-band and ten-band cases for this data set, with a range of values for $\Delta z_{peak,min}$ and a fixed $p_{f,min}$ of 0.95, while figure 5 provides a visual compendium of some of those quantities for the five-band case. Table 2 shows various metrics for several combinations of $\Delta z_{peak,min}$ and $p_{f,min}$ values. Figure 7 shows a redshift histogram of the reduction in the number of catastrophic outliers and outliers present in a typical determination with the five-band case with one particular parameter value choice. Figure 6 shows performance metrics from determinations with SPIDERz using the EPDF outlier identification method on the COSMOS-reliable-z test data set. Comparing Figures 5 and 6 it is clear that results from the two test data sets are quite similar but with significantly smaller error bars in the COSMOS-reliable-z case as would be expected from a significantly larger data set.

We see that certain choices for $\Delta z_{peak,min}$ and $p_{f,min}$ result in successfully flagging a high percentage (> 50%) of the catastrophic outliers while flagging a small percentage (2-4%) of the non-outlier galaxies. On the other hand, low values of $\Delta z_{peak,min}$ result in the flagging of a large percentage of the non-outlier galaxies.

It is also of interest to explore whether this method flags an excessive fraction of galaxies at higher redshifts and/or higher magnitudes. In Figure 8 we show the percentage of non-outliers flagged in bins of 0.1 in redshift (left panel) and in sextiles of i-band magnitude (right panel) for the COSMOS-reliable-z test data set with flagging parameter values $p_{f,min} = 0.95$ and $\Delta z_{peak,min} = 0.6$. It is seen that less than 15% of non-outliers are flagged in the highest magnitude (dimmest flux) sextile but in a few of the least populated redshift bins in the sample roughly half of non-outliers are flagged. This suggests that steps could be taken to mitigate this effect within certain low population redshift bins, as discussed in §5.

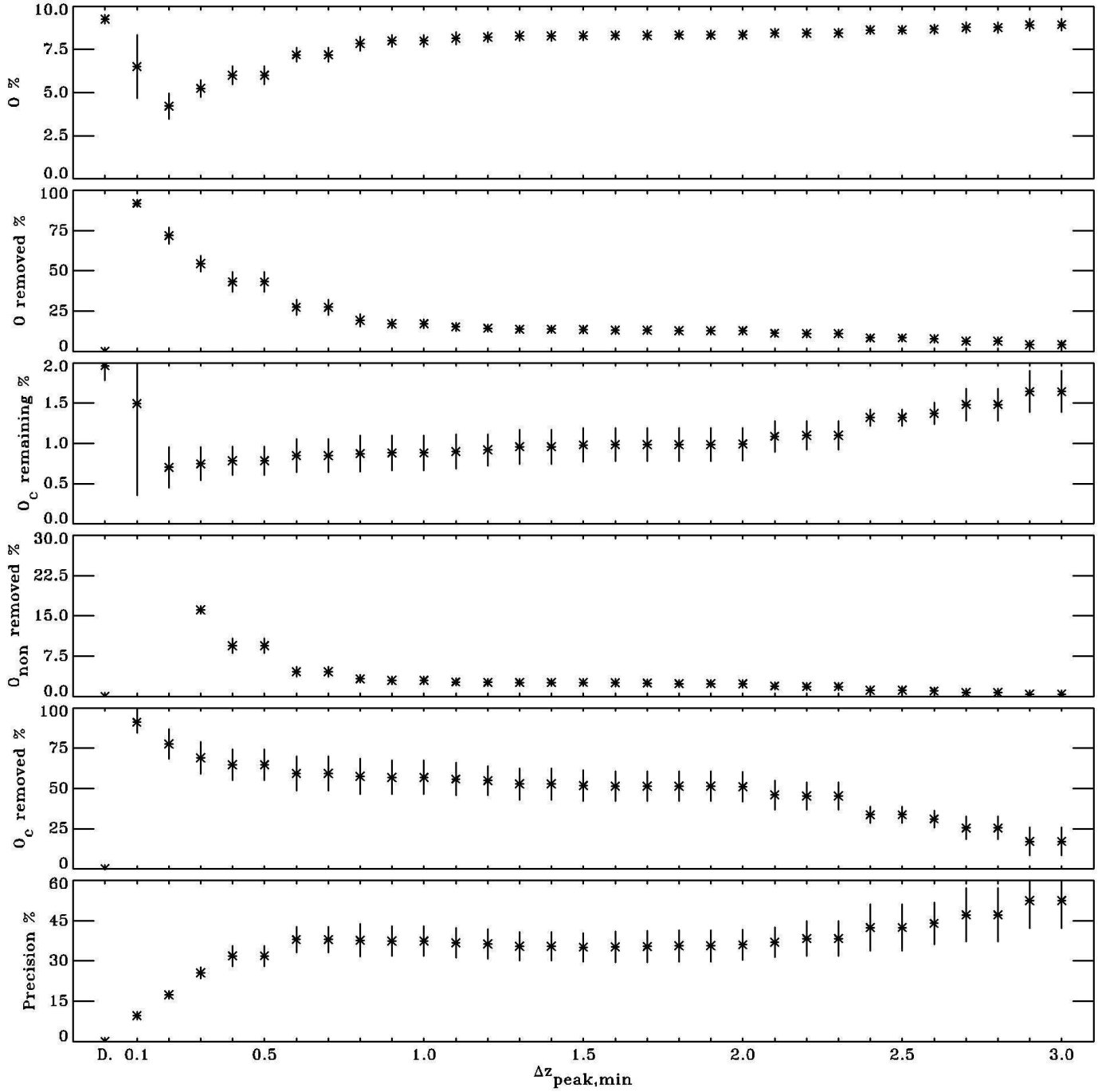

Fig. 5: Visualization of photo-z performance metrics from determinations performed by SPIDERz on the COSMOSx3D-HST data set discussed in §4.1 for the five photometric band case using a range of $\Delta z_{peak,min}$ values and fixed $p_f = 0.95$, considering that all flagged galaxies would be removed from an analysis that relied on accurate photo-zs. We also include the performance for the default case of no flagging on the left-most portion of the x-axis labeled "D.". The determinations were performed with a bin size of 0.1, and a training set consisting of 1200 galaxies chosen at random and an evaluation testing set consisting of the other 2504 galaxies, with results averaged over six determinations. The performance metrics shown include the percentage of outliers (TOP), followed by the percentage of outliers removed (2nd from TOP), followed by the percentage of catastrophic outliers remaining (3rd from TOP), followed by the percentage of non-outliers removed (3rd from BOTTOM), followed by the percentage of catastrophic outliers removed (2nd from BOTTOM), and finally the percentage of removed galaxies that are outliers (BOTTOM). The variance in performance across the six randomized realizations is indicated.

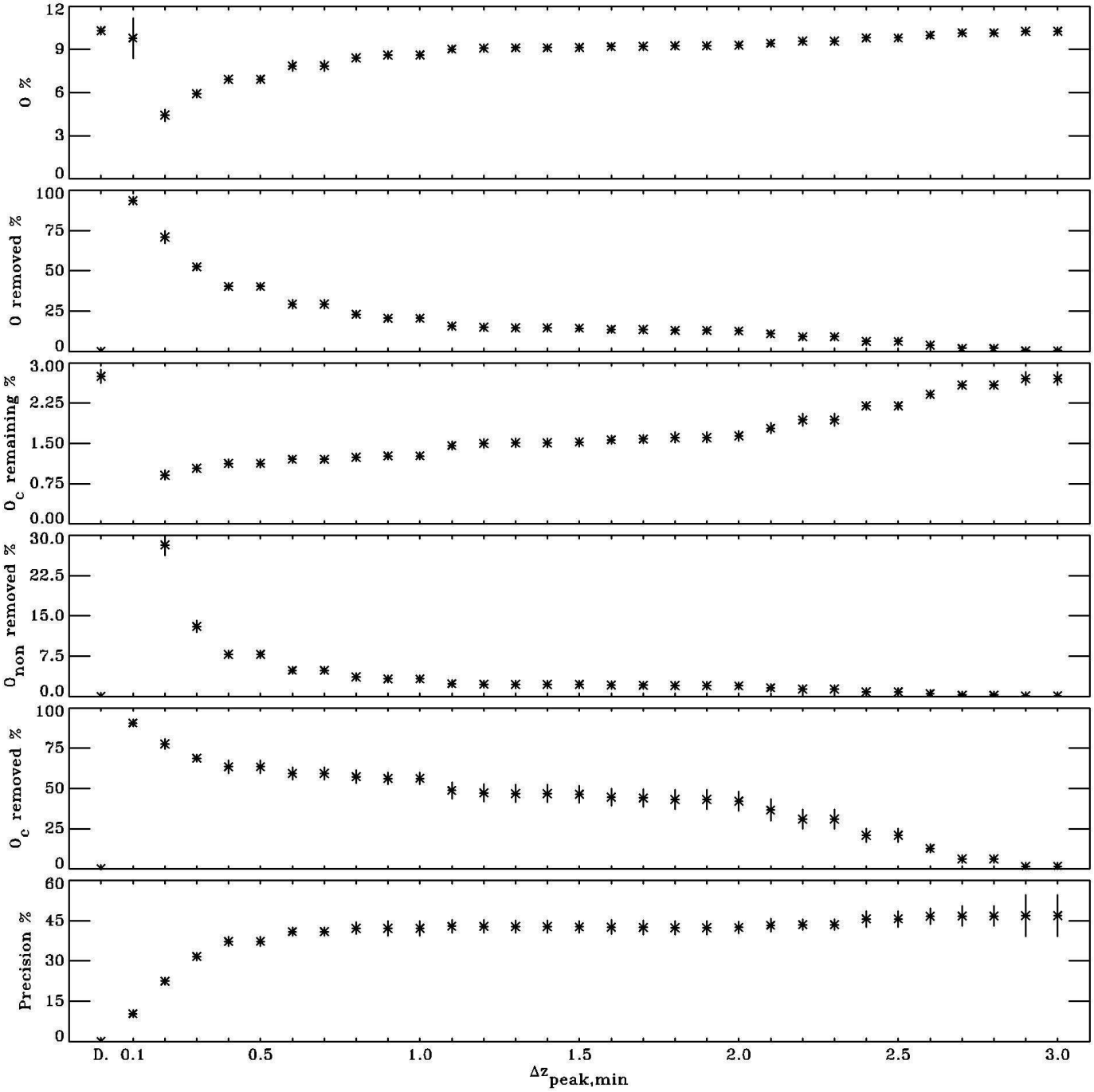

Fig. 6: Same as Figure 5 but for the COSMOS-reliable-z data set discussed in §4.1. The variance in performance across the six randomized realizations is indicated. Results from this data set are quite similar to those from the COSMOSx3D-HST data set shown in Figure 5 but with smaller error bars as would be expected from a much larger data set.

Table 1: Results for analyses performed with SPIDERz on the five- and ten photometric band test data sets derived from the COSMOSx3D-HST data discussed in §4.1. Determinations feature 1200 galaxies used for training and the remaining 2504 galaxies used for evaluation. Six determinations were performed for every case, each with randomized training and evaluation testing sets, and results averaged. Results are shown for the default cases with no flagging, and also with implementation of the EPDF outlier flagging method discussed in §3 using a range of $\Delta z_{peak,min}$ values and fixed $p_f = 0.95$, assuming that all flagged galaxies would be removed from a data set that relied on accurate photo-zs, to illustrate the percentage reduction in outlier and catastrophic outlier galaxies achieved at the cost of incorrectly removing a percentage of non-outlier galaxies. Here we use the shorthand O and O_c for outliers and catastrophic outliers, respectively, which are defined by equations 1 and 2, and O_{non} for non-outliers. The 'Precision' refers to the percentage of flagged galaxies which are outliers. The RMS and reduced RMS errors are also included for each case and defined by equation 3 as discussed in §1.

$\Delta z_{peak,min}$	O %	O^{removed} %	O_c %	O_c^{removed} %	$O_{\text{non}}^{\text{removed}}$ %	Precision %	σ_{RMS}	σ_{R-RMS}
Five photometric bands (0.1 bin size)								
Default	9.25	-	1.96	-	-	-	0.221	0.052
0.2	4.22	54.4	0.702	64.2	35.0	17.3	0.110	0.035
0.3	5.24	43.4	0.736	62.4	16.1	25.5	0.122	0.043
0.4	6.00	35.1	0.783	60.0	9.46	31.9	0.131	0.046
0.5	6.00	35.1	0.783	60.0	9.46	31.9	0.131	0.046
0.6	7.19	22.3	0.847	56.8	4.62	37.9	0.139	0.049
0.7	7.19	22.3	0.847	56.8	4.62	37.9	0.139	0.049
0.8	7.83	15.4	0.872	55.5	3.23	37.7	0.142	0.050
0.9	8.00	13.5	0.882	55.0	2.98	37.3	0.143	0.050
1.0	8.00	13.5	0.882	55.0	2.98	37.3	0.143	0.050
Ten photometric bands (0.05 bin size)								
Default	4.17	-	1.08	-	-	-	0.144	0.047
0.2	0.938	77.5	0.146	86.5	53.1	6.89	0.064	0.025
0.3	1.49	64.3	0.129	88.1	20.5	13.4	0.069	0.037
0.4	2.02	51.6	0.180	83.3	9.59	20.8	0.078	0.042
0.5	2.21	47.0	0.198	81.7	7.48	23.4	0.079	0.043
0.6	2.66	36.2	0.218	79.8	3.96	30.8	0.085	0.045
0.7	2.76	33.8	0.237	78.1	3.44	32.3	0.087	0.045
0.8	2.95	29.3	0.257	76.2	2.92	33.0	0.088	0.046
0.9	3.07	26.4	0.277	74.4	2.74	32.3	0.090	0.046
1.0	3.12	25.2	0.290	73.1	2.66	32.1	0.090	0.046

Table 2: Improvements in RMS and R-RMS (defined by equation 3), and the percentage of catastrophic outliers (O_c, defined in equation 2) after flagging potential catastrophic outlier EPDFs in SPIDERz determinations on COSMOSx3D-HST test data for the five photometric band case for a range of $\Delta z_{peak,min}$ and $p_{f,min}$ values, with a redshift bin size of 0.1, assuming removal of flagged galaxies. Six determinations were performed for every case, each with randomized training and evaluation testing sets consisting of 1200 and 2504 galaxies respectively, and results averaged. The default case is for no flagging. We also show the percentage of non-outliers (O_{non}) flagged.

$\Delta z_{peak,min}$	$p_{f,min}$	O_c %	Δ%	σ_{RMS}	Δ%	σ_{R-RMS}	Δ%	$O_{\text{non}}^{\text{removed}}$ %
Default	-	2.08	-	0.213	-	0.0445	-	
0.2	90	0.483	-76.8	0.078	-63.4	0.0274	-38.4	55.1
0.2	95	0.703	-66.2	0.110	-48.4	0.0346	-22.2	35.0
0.2	98	0.636	-69.4	0.102	-52.1	0.0349	-21.6	32.5
0.3	90	0.486	-76.6	0.092	-56.8	0.0392	-11.9	28.8
0.3	95	0.746	-64.1	0.122	-42.7	0.0426	-4.27	16.1
0.3	98	0.664	-68.1	0.118	-44.6	0.0418	-6.07	17.0
0.4	90	0.543	-73.9	0.100	-53.1	0.0463	4.04	16.0
0.4	95	0.783	-62.4	0.131	-38.5	0.0462	3.82	9.46
0.4	98	0.659	-68.3	0.122	-42.7	0.0457	2.70	9.82
0.5	90	0.562	-73.0	0.103	-51.6	0.0481	8.09	12.8
0.5	95	0.783	-62.4	0.131	-38.5	0.0462	3.82	9.46
0.5	98	0.659	-68.3	0.122	-42.7	0.0457	2.70	9.82
0.6	90	0.627	-69.9	0.115	-46.0	0.0512	15.1	7.77
0.6	95	0.847	-59.3	0.139	-34.7	0.0491	10.3	4.62
0.6	98	0.703	-66.2	0.128	-39.9	0.0486	9.21	4.73
0.7	90	0.625	-70.0	0.116	-45.5	0.0519	16.6	6.60
0.7	95	0.847	-59.3	0.139	-34.7	0.0491	10.3	4.62
0.7	98	0.703	-66.2	0.128	-39.9	0.0486	9.21	4.73
0.8	90	0.630	-69.7	0.118	-44.6	0.0526	18.2	5.07
0.8	95	0.872	-58.1	0.142	-33.3	0.0499	12.1	3.27
0.8	98	0.743	-64.3	0.135	-36.6	0.0494	11.0	3.20

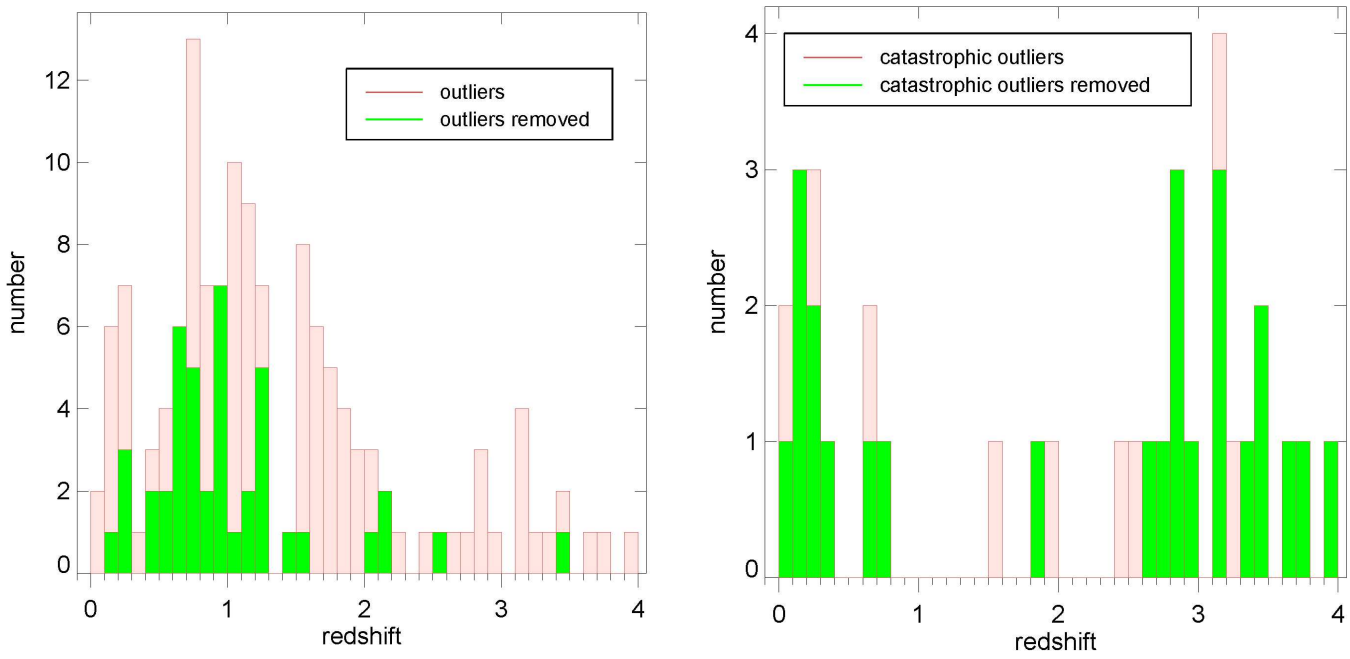

Fig. 7: Redshift histogram of the number of outliers (left) and catastrophic outliers (right), both as defined in equations 1 and 2 respectively, present in one particular typical determination with the five photometric band case for the COSMOSx3DHST test data set compared to the numbers flagged through the use of the EPDF flagging method with flagging parameter values $p_{f,min} = 0.95$ and $\Delta z_{peak,min} = 0.6$.

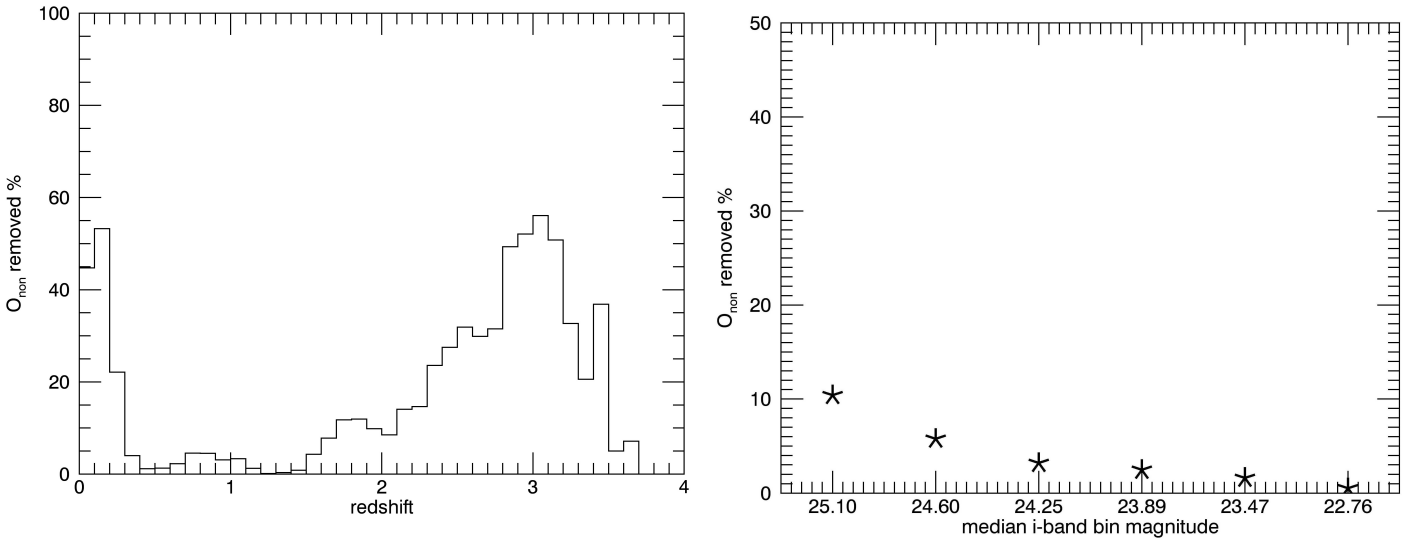

Fig. 8: The percentage of non-outliers flagged through the use of the EPDF flagging method in bins of 0.1 in redshift (left) and in equally populated sextiles of *i*-band magnitude from highest to lowest magnitude indexed with the median magnitude (right) for the COSMOS-reliable-*z* test data set with flagging parameter values $p_{f,min} = 0.95$ and $\Delta z_{peak,min} = 0.6$. The results here are for one particular representative determination. The standard deviation from averaging over multiple determinations would be smaller than the plotting symbols in the *i*-band case and small in the redshift case. The redshift bins where a large fraction of non-outliers are flagged are those which are least populated in the sample generally.

5. Discussion

In this work, we have considered the utilization of SPIDERz's effective redshift probability distributions for flagging likely catastrophic outlier photo-z predictions — gross mis-estimations defined by $|z_{phot} - z_{spec}| > 1$ — by considering galaxies with multiple or ill-defined peaks in photo-z probability separated by redshift. We introduced a formalism with two threshold criteria: the minimum redshift separation of multiple peaks ($\Delta z_{peak,min}$) and the minimum probability ratio of secondary probability peaks to the highest probability peak ($p_{f,min}$), as discussed in §3, to preemptively flag potential catastrophic outlier estimates. We implemented this method in SPIDERz photo-z determinations performed with real galaxy test data spanning a wide redshift range $0 < z < 4$ and utilizing limited photometric bands to estimate photometric redshift (see §4.1), testing a range of threshold values $\Delta z_{peak,min}$ and $p_{f,min}$ (see §3).

We found $\Delta z_{peak,min}$ to have the greatest influence on the fraction of catastrophic outliers which were flagged, while $p_{f,min}$ was sub-dominant in this regard but most strongly correlated with flagging precision, with low values of $p_{f,min}$ leading to a higher number of non-outliers flagged. Optimal values for $\Delta z_{peak,min}$ and $p_{f,min}$ for any given application would result from striking an acceptable balance between more thoroughly flagging catastrophic outlier galaxies and reducing the number of spuriously flagged non-outlier galaxies.

We present results for a variety of choices of $\Delta z_{peak,min}$ where this trade-off can be seen, particularly in Figure 5 and Tables 1 and 2. There are a range of options away from the lowest values of $\Delta z_{peak,min}$ where the percentage of catastrophic outliers flagged is quite high and the percentage of non-outliers flagged is relatively low. For all parameter choices, more non-outliers are flagged than outliers, but this is likely inevitable considering that in the default case more than 90% of the galaxies in the five-band case and 95% in the ten-band case are non-outliers.

We have seen that with proper choices for $\Delta z_{peak,min}$ and $p_{f,min}$ EPDFs can be utilized to flag potential catastrophic outlier photo-z predictions with a high degree of overall effectiveness in determinations performed on a data set which spans a wide redshift range and contains realistic photometry in a limited number of wavebands. As discussed in §3, in a future large scale survey utilizing photometric redshifts, the simplest use of such flagging information would be to simply remove the flagged galaxies from science analyses in which catastrophic outlier redshift predictions are detrimental, such as weak-lensing cosmology. Another simple option for utilization of flagging information could include de-weighting of potential catastrophic outliers in cosmological probes.

If such flagged galaxies are simply removed from analysis, there is, necessarily, a trade-off between more complete removal of actual catastrophic outliers and spurious removal of non-outliers. In this work we present various options for the parameters $\Delta z_{peak,min}$ and $p_{f,min}$ (discussed in §3) which lead to different points on this trade-off continuum. We show the various results for catastrophic outliers removed, spurious removals, and other metrics in Tables 1 and 2 and visualizations in Figures 7 and 5. It is seen that for a range of flagging parameter values a favorable ratio of total genuine catastrophic outlier flagging to spurious non-outlier flagging is obtained, for example flagging of significantly more than half of catastrophic outliers while spuriously flagging only 2-4% of non-outliers. With the need to obtain precise redshift estimates satisfying photo-z error constraints for probing cosmological parameters and the abundance of galaxies that will be observed in future large photometric surveys, it may be reasonable in many cases to accept a slightly larger (although still low) percentage of overall spurious removals in exchange for maximizing the number of removed catastrophic outlier photo-z estimates.

It is important to note, however, that as seen in Figure 8 a significant fraction (approaching half in the most dramatic cases) of non-outliers are flagged in a few of the more sparsely populated redshift bins, including some of those at higher redshifts. This points toward a possible strategy beyond simple removal of flagged galaxies in these particular redshift bins in order to not lose for cosmological analyses such a large fraction of high redshift galaxies in a data set. We will explore possible weighting strategies for this in a future work. We do also note two crucial caveats regarding this: (1) that in this work, as mentioned in §1, in order to approximate the photo-z conditions applying to future large scale surveys, we utilize much larger evaluation sets than training sets in this study. Thus it is likely that by adopting a larger training to evaluation set size ratio than here, as has been done in many other photo-z studies in the literature, one could reduce the percentage of spuriously flagged non-outliers in the sparsely populated redshift bins given a similarly sized test data set. Also, (2) it is likely the case that, for a given training to evaluation set size ratio and $N(z)$ distribution, there will be a lower percentage of spuriously flagged non-outliers in relatively sparsely populated redshift bins given a larger overall test data set. However even with a very large training set high redshift bins will contain a higher proportion of potential catastrophic outliers and therefore spurious removals due to the degeneracy between Balmer and Lyman breaks in galaxy spectra.

While this analysis focused on utilization of EPDFs provided by SPIDERz, there is no reason that it should not be generalizable with analagous parameters to any photo-z estimation method which provides redshift probability distribution information for each galaxy. While the parameters we used in this work to flag EPDF features, $\Delta z_{peak,min}$ and $p_{f,min}$, were effective in distinguishing likely catastrophic outliers, the optimal values of these parameters for a given purpose may be data set dependent to some extent. Also other photo-z estimation codes and probability determination methods may or may not necessitate alternate parameter values and/or definitions to those employed in this work. We also note that in general results in empirical photo-z estimation methods often depend on the degree of representativeness of the training set relative to the evaluation set.

References

Abdalla, F. B., Banerji, M., Lahav, O., Rashkov, V. 2011, MNRAS, 417, 1891
Almosallam I. A., Lindsay S. N., Jarvis M. J., & Roberts S. J. 2016, MNRAS, 455, 2387
Arnouts, S., Cristiani, S., Moscardini, L., Matarrese, S., Lucchin, F., Fontana, A., Giallongo, E. 1999, MNRAS, 310, 540
Ball, N. M., Brunner, R. J., Myers, A. D., Strand, N. E., Alberts, S. L., & Tcheng, D. 2008, ApJ, 683, 12
Ball, N. M., Brunner, R. J., Myers, A. D., Strand, N. E., Alberts, S. L., Tcheng, D., & Llora, X. 2007, ApJ, 663, 774
Benítez, N. 2000, ApJ, 536, 571
Benítez, N., et al. 2004, ApJ, 150, 1
Bernstein, G. & Huterer, D. 2010, MNRAS, 401, 1399
Bolzonella M., Miralles J.-M., Pello' R. 2000, å, 363, 476
Bonnett, C. 2015, MNRAS, 449, 1043
Brammer G. B., Dokkum P. G. v., Coppi P. 2008, ApJ, 686, 1503
Brescia M., Cavuoti S., Longo G., & De Stefano V. 2014, å, 568, A126
Carrasco Kind, M. & Brunner, R. 2013, MNRAS, 432, 2
Collister A. A. & Lahav O. 2004, PASP, 116, 345
Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., & Munn, J. A. 1995, AJ, 110, 2655
Feldmann R., et al. 2006, MNRAS, 372, 565
Firth A. E., Lahav O., & Somerville R. S. 2003, MNRAS, 339, 1195

Freeman, P. E., Newman, J. A., Lee, A. B., Richards, J. W., & Schafer, C. M. 2009, MNRAS, 398, 2012
Gerdes, D., et al. 2010, ApJ, 715, 823
Hearin, A., Zentner, A., Ma, Z., & Huterer, D. 2010, ApJ, 720, 1351
Hildebrandt, H. et al. 2010, å, 523, 832
Hogan R., Fairbairn M., & Seeburn N. 2015, MNRAS, 449, 2040
Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS, 366, 101
Ilbert, O., et al. 2006, A&A, 457, 841
Ilbert, O., et al. 2009, ApJ, 690, 1236
Ivezic, Z. et al. 2008 arXiv:0805.2366
Kotulla R., Fritze U. 2009, MNRAS, 393, L55
Jones, E. & Singal, J. 2017, A&A, 600, A113
Kitching, T. D., Taylor, A. N., & Heavens, A. F. 2006, MNRAS, 332, 788
Laigle, C. et al. 2016 ApJS, 224, 24
Laureijs, R. et al. 2011 arXiv:1509.03318
Leistedt B. & Hogg D. W. 2017, ApJ, 838, 5
Ma, Z., Hu, W., & Huterer, D. 2006 ApJ, 636, 21
Momcheva, I. et al. ApJS, 225, 27
Rau M. M., Seitz S., Brimioulle F., Frank E., Friedrich O., Gruen D., & Hoyle B. 2015, MNRAS, 452, 3710
Sadeh, I. et al. 2016, PASP, 128, 968
Sanchez, C. et al. 2014, MNRAS, 445, 1482
Scoville, A., et al. 2007, ApJS, 172, 1
Singal, J., Shmakova, M., Gerke, B., Griffith, R.L., & Lotz, J. 2011 PASP, 123, 615
Sadeh I., Abdalla F. B., & Lahav O. 2016, PASP, 128, 104502
Vanzella et al. 2004, A&A, 423, 16
Wadadekar, Y. 2004, PASP, 117, 79
Wang, D., Zhang, Y., Liu, C., & Zhao, Y. 2007, CJAA, 7, 43
Way M. J. & Srivastava A. N. 2006, ApJ, 647, 102