Thinking Outside the Bins: Evolution of Galaxy Morphology Over Cosmic Time

Jining Qin*
Advisors: Ann Lee*, Peter Freeman*, Jeffrey Newman†

February 4, 2016

1 Introduction

Galaxy morphology refers to the two-dimensional appearance of a galaxy as projected onto the sky. Astronomers are interested in galaxy morphology because it contains important information about the evolution of the universe and can help constrain theoretical models in cosmology.

Though we have vast amount of image data of the galaxy population we are interested in, direct use of images turns out to be statistically and computationally intractable. Some dimensionality reduction method is needed to enable further analysis in the context of morphology evolution.

Classification is the dimensionality reduction method astronomers have long used in order to divide galaxies into discrete groups. The Hubble sequence (Figure 1) was among the early well-received attempts. Approaching galaxy morphology with classification methods, however, has several drawbacks. Discretization of continuous data obviously leads to loss of information. Also, classification schemes are usually defined based on galaxies in the local universe and do not extrapolate well to older galaxies, since the universe evolves and the galaxy population changes over cosmic time. Moreover, Using human annotators brings problem to the inferential accuracy since a lot of non-expert annotators are used in the labeling process.

Adopting a continuous approach in depicting galaxy morphology can help us avoid these drawbacks, as large-scale sky surveys such as the Sloan Digital Sky Survey[1] program accumulated vast amount of data to support the measure of galaxy morphology on a continuous scale. Various continuous quantitative measures, i.e. feature statistics of galaxy morphology have been proposed that extract low-dimensional features from galaxy images and in principle retain the information present in images. So the question we wish to answer becomes: how do these statistics evolve with time?

*Department of Statistics, Carnegie Mellon University
†Department of Physics and Astronomy, University of Pittsburgh
[1] www.sdss.org

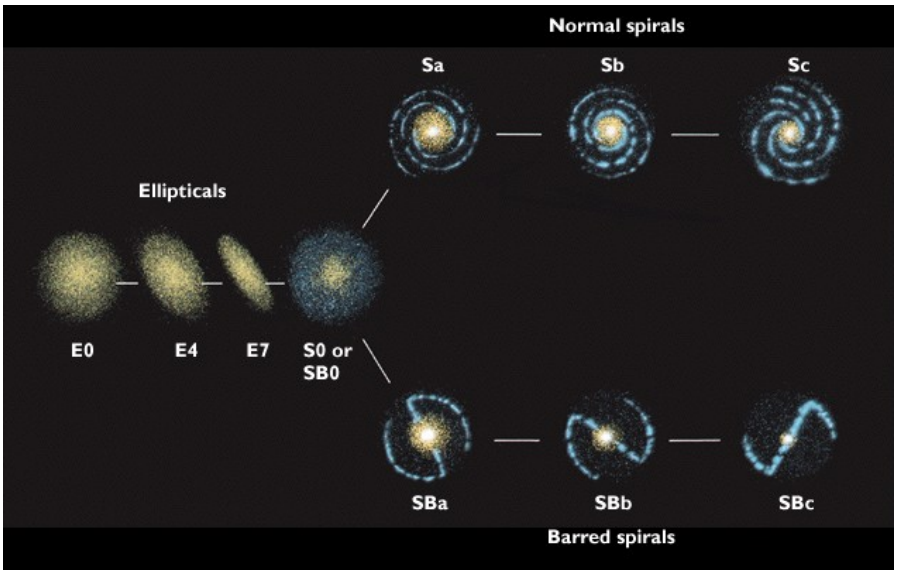

Figure 1: Hubble's tuning fork diagram for galaxy morphology classification.

The standard way of measuring cosmic time in cosmology is redshift (usually denoted by Z). Redshift refers to the increase in wavelength of a light source when it is moving away from the observer. It is defined as following,

$$z = \frac{\lambda_{observed} - \lambda_{emitted}}{\lambda_{emitted}}$$

where $\lambda_{emitted}$ and $\lambda_{observed}$ are the wavelengths of a given photon when it was first emitted by a galaxy and when it was observed, respectively.

The faster a galaxy appears to be moving away from us due to expansion of the universe, the greater its redshift is. The redshift-measured velocity of the galaxy is positively correlated to their distance from the earth. And since light from faraway galaxies takes longer to reach Earth, the images of high redshift galaxies are snapshots indicating how galaxies appeared early in the Universe's history. In other words, as redshift increases, we are going deeper into space and looking at a younger universe.

Therefore, the evolution of galaxy morphology over cosmic time reflects in the change in distribution of morphological feature statistics as a function of redshift. In particular, we want to explore methods to estimate the conditional distribution of morphological statistics given the redshift. We want to see whether there is notable trend in the statistics and whether our estimates are robust to different form of redshift input and image data sources.

Our major challenge comes from the form of data. For large scale sky surveys like the one we are working with, it is not feasible to obtain spectroscopy for all galaxies. Astronomers have to get photometric redshift measurement using broad band filters instead. The photometric redshift in this case is a measurement with uncertainty instead of a precise value. Also, since the images are taken under several different filters, we have several sets of feature statistics for each galaxy. It is not trivial how and when to use each version of the feature statistics.

We illustrate the challenges brought by the particular dataset using some simulated data points in Figure 2. Typically, we should have predictor-response pairs (z_i, y_i), as

in the plot (a). But the precise single value redshift measurement requires taking full spectroscopy of a galaxy and isn't feasible in large scale surveys. Our data set has the photometric redshift measurement. The z-coordinate has a probability distribution instead of a single precise value, as shown in plot (b). Moreover, the image of each galaxy is taken in different filters, leaving us with several correlated sets of image feature statistics for each galaxy, as shown in plot (c).

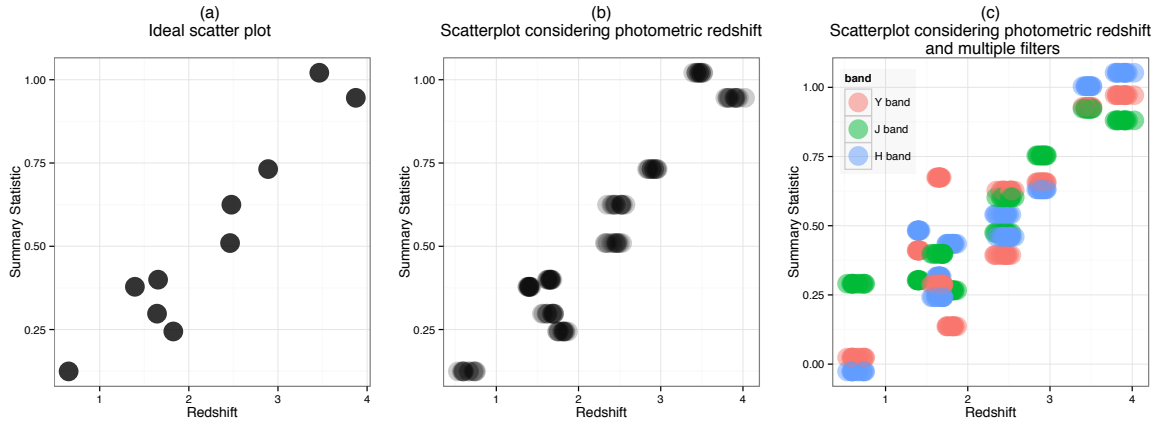

Figure 2: Illustration of challenges in analyzing our dataset. These data points are simulated just to illustrate the form of data we have.

Previously, astronomers used what we call the discrete bin approach, mapping galaxies to ranges in redshift associated with band filters and comparing the morphology of galaxies in different ranges. We extended the estimation of morphology evolution to a continuous redshift scale. In other words, we tried to model the conditional distribution of morphological feature statistics of galaxies given any value of its redshift within appropriate range.

This paper is organized as follows: Section 2 describes the origin and necessary preprocessing of our data. Section 3 introduces the challenges and our strategy in estimating how the distribution of feature statistics changes over cosmic time. In particular, we describe the discrete bin approach and then propose our continuous redshift approach. The results are presented in Section 4, and Section 5 contains our conclusion and further discussion about problems in the current analysis and further steps.

2 Data description

2.1 Data source and preprocessing

The data are galaxies from five fields that were observed by the Hubble Space Telescope as part of the CANDELS program (Koekemoer et al. (2011),Grogin et al. (2011)): COSMOS, EGS, GOODS-North (GOODSN), GOODS-South (GOODSS), and UDS. Galaxies are retained for analysis only if one has a magnitude [2] in the H band less than 25 and

[2]Magnitude is a logarithmic measure of brightness. A smaller magnitude value means a brighter object.

an estimated zero-redshift mass greater than $10^{10} M_{\odot}$.[3] Image data are taken from different wavelength ranges also known as photometric bands. Figure 3 shows the probability for light to be transmitted at each wavelength in each band as a function of wavelength. Table 1 summarizes the dataset available to us.

Only GOODSN and GOODSS field have image data from all five filters available. Also, we notice that images taken in Y, J and H filters are from camera WFC3 while V and i filters are from camera ACS. Exploratory analysis shows discordance between images of the same galaxy from different cameras. Therefore, we will use the images taken in filters Y, J and H for galaxies in field GOODSN and GOODSS.

Table 1: Summary of available cosmic fields

Field	Number of Catalogued Objects	Number of Galaxies After Cuts	Filters for Available Images
COSMOS	38671	3970	V, i, J, H
EGS	41457	4237	V, i, J, H
GOODSN	35451	3899	V, i, Y, J, H
GOODSS	34930	3385	V, i', Y, J, H
UDS	35932	4181	V, i, J, H

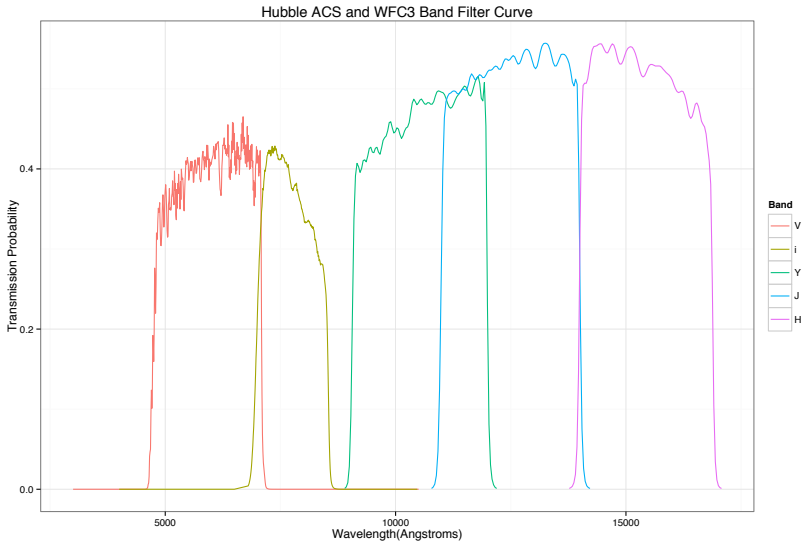

Figure 3: Filter curves for the V, i, H, J and Y Band. [4]

2.2 Redshift measurement for galaxies

The indicator of cosmic time is the redshift Z of a galaxy. The greater the redshift, the farther away the galaxy is from us, and the longer it takes the light it emits to reach our telescope, and thus the older the galaxy is.

[3] $M_{\odot}$ is one solar mass.
[4] Angstrom= 10^{-10}m

Two primary ways of estimating redshift are spectroscopic redshift measurement and photometric redshift measurement. Spectroscopic redshift measurement uses signature spikes and troughs resulting from atomic transitions in spectra of galaxies to infer redshifts of galaxies. Spectroscopic redshifts are usually very precise. However, spectroscopic redshift measurements are costly in terms of time and thus not feasible for large scale sky surveys. A photometric redshift measurement, on the other hand, uses broad-band observations and is much more affordable. The measurement does not have high precision. In fact, photometric redshift measurement provides us with only a probability distribution of a galaxy's redshift. Redshift estimation techniques often fall into two categories: empirical techniques that utilize machine learning algorithms and template fitting algorithms that use a dictionary of representative galaxy spectra.

The data available include redshift measurements from different methods. In the GOODSS field, we have access to the discretized distribution function of each galaxy's redshift on interval $[0,10]$. In other words, we know the value $\int_x^{x+0.01} f_Z(t)dt$ for $n = 0, 0.01, 0.02, \cdots, 10$, where $f_Z(t)$ is the estimated probability density function for the redshift of a galaxy.

In the other four fields, we have point estimates for the redshifts of galaxies. These point estimates could come either from spectroscopic redshift measurements or photometric measurements. The difference in measurement techniques for this key variable poses an interesting challenge to our analysis, and provides an opportunity for testing the robustness of our methodology.

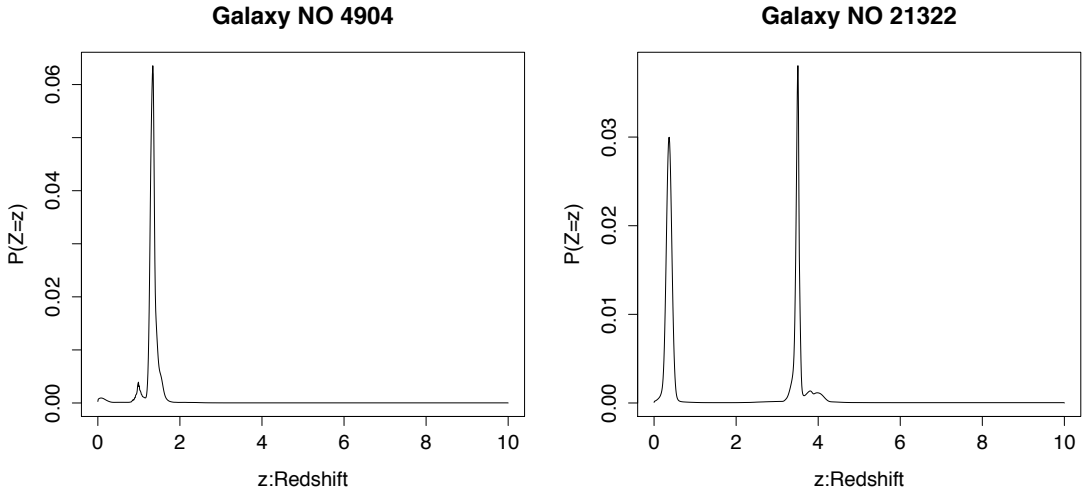

Figure 4: The redshift of most galaxies have only one mode, yet some galaxies do exhibit bimodality or other irregular shapes in the distribution of their redshift estimates.

2.3 feature statistics of galaxy morphology

Instead of using the image pixel data directly, we extract several feature statistics depicting the morphological characteristics of the galaxies. Among the statistics are the Multimode (M), Intensity (I), and Deviation (D) statistics proposed in Freeman et al. (2013) are sensitive to the signatures of two galaxies merging. Gini (G) and M_{20}, described in

Lotz et al. (2004), measure the concentration of light within a galaxy. Concentration (C), which also measures the concentration of a galaxy's stellar light distribution and Asymmetry (A), which measures the degree of asymmetry of a galaxy, are defined in Conselice (2003). Mathematical details for these statistics are given in the Appendix.

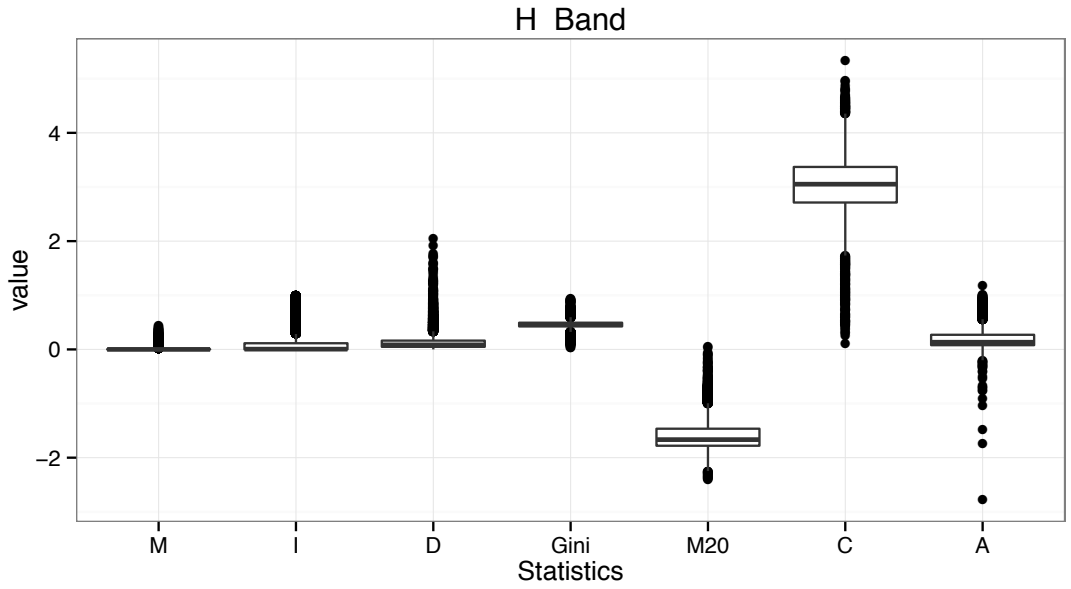

Figure 5: Several statistics are asymmetric. It is apparent that M_n, I D are skewed. Outliers exist for all these statistics.

Figure 5 shows boxplots of the feature statistics for galaxies in the GOODS-South field calculated using images from the H band filter. Distributions of these statistics in the other bands look very similar to the H-band distribution.

2.4 Summary of our use of the data

As shown in Table 2, we have the full photometric redshift measurement for galaxies in field GOODSS, but only point estimates for the other four fields. For only the GOODSS and GOODSN fields, we have feature statistics taken in all 3 bands from the WFC3 camera available. A comparison between analyses on these two fields will be good robustness check for our method.

Table 2: Available redshift measurement and filter bands by field

Field	Available Redshift Measurement	Available Filter Bands
COSMOS	Point estimate	V, i, J, H
EGS	Point estimate	V, i, J, H
GOODSN	Point estimate	V, i, Y, J, H
GOODSS	Probability density	V, i', Y, J, H
UDS	Point estimate	V, i, J, H

3 Methods

Galaxy morphology evolution denotes a change in population of galaxies' appearances in universe over cosmic time. As mentioned earlier, galaxy morphology is depicted by the image feature statistics and cosmic time is measured via redshift. Therefore, galaxy morphology evolution is reflected in the change in conditional distribution of feature statistics given redshift.

The major challenge in our analysis is the fact that redshift measurements are not precise (in other words, they contain measurement error), and there are several sets of feature statistics for each galaxy, due to correlated images taken in different filter bands. In the discrete bin approach often used by astronomers, we associate a range in redshift (we will call them redshift bins) to each filter band and map galaxies to these ranges. Then we compare the statistics of galaxies in each bin taken in the corresponding filter band. With the uncertainty in explanatory variable, we instead map galaxies to wider ranges in redshift and compare the conditional distribution of feature statistics between the redshift ranges. In the continuous approach, we treat each galaxy as a weighted observation with the point estimate of its redshift (the point estimate is either given in the dataset or obtained by us) and feature statistics taken in different filter bands weighted by its probability of falling in each band's corresponding redshift bin. We can then use kernel regression and quantile regression to model the continuous change in distribution of the feature statistics.

We explore the discrete bin approach previously used by astronomers and our continuous redshift approach. We describe our method for recovering continuous density function for z given its discretized version in 3.1. We describe the discrete bin approach comparisons in 3.2 and continuous redshift approach in 3.3.

3.1 Recovering the continuous distribution function for redshift

As mentioned in 2.2, we have the discretized probability density function for redshift of every galaxy in the GOODSS field. Since we want to treat redshift (Z) as a continuous variable with some uncertainty in measurement, we infer its full probability density function using the information in its discretized version. We achieve the goal with reasonable accuracy due to some properties of probability distributions.

The following steps help us in recovering first the cumulative density function and then the probability density function.

1. We calculate the cumulative sum of the discretized distribution function and get point values of the cumulative distribution function $F_Z(z)$:

$$F_Z(z) = \sum_{x+0.01<z} \int_x^{x+0.01} f_Z(t)dt$$

   we get $F_Z(z)$ for $z = 0, 0.01, \cdots, 10$ from this step.

2. We fit a monotonically increasing cubic spline for $F_Z(z)$ over $z \in [0, 10]$ using the method described in Fritsch and Carlson (1980), and get the probability density function $f_Z(z)$ via differentiation:

$$f_Z(z) = \frac{d}{dz} F_Z(z).$$

3.2 Uncovering Morphology Evolution via Discrete Bin Approach

3.2.1 Mapping galaxies to discrete redshift bins

We want to map galaxies to several different groups according to their redshift. In particular, we fix a wavelength value at which the morphologies are most interesting to us. Then we find the band filter most likely to observe the light at this wavelength for each galaxy. Some galaxies can then be mapped to a band filter (and its correpsonding redshift range), while some are discarded since they are not mapped to any band filter.

Recall the definition of redshift,

$$z = \frac{\lambda_{observed} - \lambda_{emitted}}{\lambda_{emitted}}$$

After some small rearrangement it becomes

$$\lambda_{observed} = (1+z) \cdot \lambda_{emitted}$$

The redshift z and observed light wavelength $\lambda_{observed}$ are linearly related if we assume fixed emitted light wavelength (also known as the rest-frame wavelength). Further recall the filters described in 2.1: each band filter is associated with a certain range of observed light wavelength. Therefore assuming a fixed rest-frame wavelength (4500$\mathring{A}$ here), each filter is associated with a certain range in redshift.

$$\lambda_{observed} \in [\lambda_{min}, \lambda_{max}] \implies z \in [\frac{\lambda_{min} - \lambda_{rf}}{\lambda_{rf}}, \frac{\lambda_{max} - \lambda_{rf}}{\lambda_{rf}}]$$

In other words, assuming a $\lambda_{rf} = 4500\mathring{A}$ rest-frame wavelength, there is one band in which photons emitted at $4500\mathring{A}$ are most likely to be observed. Assuming $\lambda = 4500\mathring{A}$ as the rest-frame wavelength, the Y band is associated with redshift range $z \in [1.05, 1.63]$, J band is associated with $z \in [1.46, 2.09]$, and H band is associated to $z \in [2.12, 2.71]$.

For every galaxy, we compute the probability that $4500\mathring{A}$ photons are observed in each band. If the band with the highest probability, has a probability of more than 0.8 (or 0.6), then we select the band or else we throw out the galaxy completely.

For galaxies in field GOODSN, we only have a point estimate for each galaxy's redshift. Therefore we would identify a galaxy with a redshift band if its redshift point estimate is within that band's redshift range. And when the point estimate falls within the overlap of two redshift bins, we pick the redshift bin with its center closest to the point estimate measured by scaled distance (distance divided by width of the redshift bin). About a third of the galaxies are left without falling into one of the five bins in these fields.

Table 3 summarizes the result of binning galaxies in GOODSN and GOODSS field.

3.2.2 Comparing different redshift bins

After assigning galaxies to discrete redshift bins, we can compare the distribution of morphological statistics across them.

As we mentioned in 3.2.1, a galaxy belongs to a redshift bin implies that its $4500\mathring{A}$ light is mostly likely to be observed in the corresponding filter band. So for any galaxy, the morphological statistics taken in that filter would reflect the morphology at $4500\mathring{A}$.

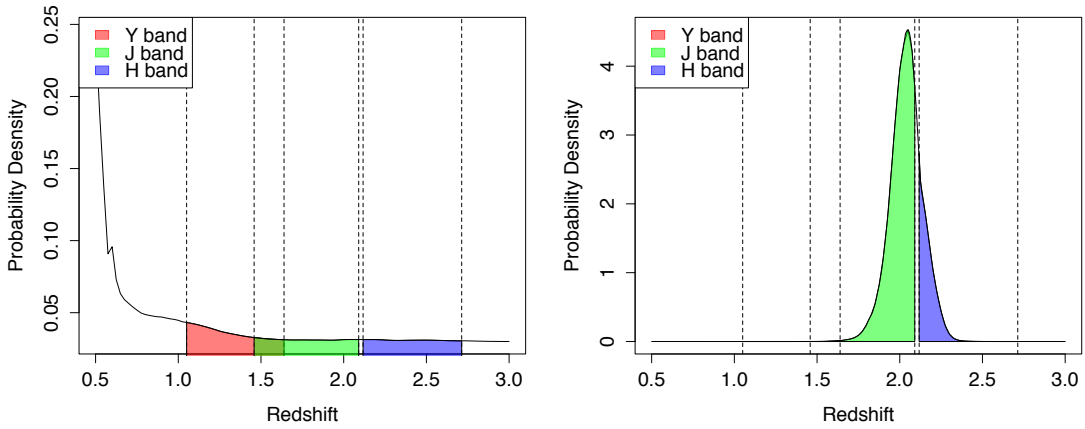

Figure 6: Illustration of mapping galaxies to redshift bins based on photometric redshift measurement. The curve is density function of redshift given by photometric redshift measurement, while shaded areas are probability of the redshift being in the redshift range of each filter band. Here the galaxy on the left is not mapped to any redshift bins while the galaxy on the right is mapped to the J band. We set the probability threshold at 0.8 here.

Table 3: Binning results for galaxies in GOODSN, GOODSS field

Field	Redshift measurement	Number of galaxies				
		Total after cuts	Y band	J band	H band	Not mapped to a band
GOODSN	Point estimate	3852	579	740	806	1774
GOODSS	Photometric	3385	425	524	534	1902

Hence the morphologies of different galaxies become comparable. The three groups of feature statistics reflect the morphology of galaxies at 4500$\mathring{A}$ in redshift range $[1.05, 1.63]$, $[1.46, 2.09]$, and $[2.12, 2.71]$

We can compare the statistics of galaxies in different bins by looking at their respective summary statistics. For example, the M and I statistics are both zero-inflated. Comparing the portion of 0's in different bins would show how frequency of galaxy mergers changed over cosmic time. In general, an exploratory comparison can be easily achieved by showing side-by-side boxplots of statistics within each bin.

3.2.3 Significance test for the bin comparison

We test for significance in the redshift bin comparisons using a bootstrap analog of the ANOVA F test described in Zhou and Wong (2011). We can not use any of the commonly used methods as the distribution of feature statistics include abundance of outliers and so the assumption of normality is violated. The null hypothesis of the test is that the feature statistics in different redshift bins come from the same underlying distribution and the alternative hypothesis is that at least one of the redshift bins differ significantly from the rest.

Under the null hypothesis, feature statistics of galaxies from different redshift bins come from the same distribution, so we pool them together and sample with replacement from the pooled data three new sets of feature statistics $\{Y_i^{(1)}\}, \{Y_i^{(2)}\}, \{Y_i^{(3)}\}$, which correspond to the feature statistics of galaxies in band Y, J and H respectively. We then calculate the following analog of F-statistic, which measures how much of the total variance in the feature statistics is explained by the variance between different redshift bins,

$$F = \frac{\sum_{l=1}^{3} (\bar{Y}^{(l)} - \bar{Y})^2/(3-1)}{\sum_{l=1}^{3} \sum_{i} (Y_i^{(l)} - \bar{Y}^{(l)})^2/(N-1)}$$

where $N = 1483$ is the total number of galaxies associated with Y, J and H band.

Repeat this resampling and calculation enough times, we can obtain the approximate null distribution of this test statistic. Compare this null distribution with the test statistic calculated with original data, we have a significance test for the comparison between redshift bins.

3.3 Continuous redshift approach

The discrete bin approach is not very satisfactory for several reasons. First, as is shown in Table 3 over one thousand galaxies in each field are not mapped to any bin. In other words, at least one third of our data are not utilized in the discrete bin comparison. Second, the comparison does not extrapolate to redshift ranges outside the redshift bins. We will not be able to make general statements about the change in feature statistics as a function of redshift. Third, no information can be obtained about the trend of feature statistics within the range of a redshift bin. However, these local trends could be interesting for astronomers, since the redshift bins are wide in terms of cosmic time (usually ranging billions of years).

Therefore, we would want to estimate the distribution of morphological statistics over a wide continuous redshift range. For a galaxy with photometric redshift measurements, we identify the mode of its redshift probability distribution and treat that its redshift value. We can then use regression methods to obtain conditional distribution of feature statistics over a continuous redshift range. We describe our method of identifying significant modes in 3.3.1.

The correct set of feature statistics to use in this approach is not obvious as in 3.2, since we are trying to retain all galaxies in the analysis, including the thousands of them which do not correspond to any specific filter band. One natural way to think is to weigh different feature statistics taken in different filters with probability of being in the corresponding redshift bins. We outline the method in 3.3.2.

3.3.1 Finding modes for photometric redshifts

As a starting point, we ignore the uncertainty in photometric redshift Z. Instead, we use the most significant mode in each galaxy's redshift probability distribution as its real redshift value, i.e. consider $z_i^* = \arg\max_z f_{Z_i}(z)$. Then the data can be considered as (z^*, Y) pairs, where Y represents the feature statistics of the galaxy. The change in

conditional distribution of feature statistics Y can be estimated using standard nonparametric regression methods. We pay special attention to the mode because the estimated distribution function for redshift is less reliable at other points.

Given the probability density functions for photometric redshift recovered in 3.1, we proceed to identify nontrivial modes of each distribution following the steps as described below.Presence of noise in redshift measurement leads to lots of spurious peaks in density function. So we first smooth the density function and then identify the modes by following the steps given below.

1. We convolve the estimated density function for redshift Z, $f_Z(z)$, with the density function of distribution $N(0, 0.05^2)$ to get smoothed density function of redshift $\widetilde{f_Z}(z)$:

$$\widetilde{f_Z}(z) = \int_0^{10} f_Z(t)\phi(z-t)dt$$

   where $\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2})$.

2. We find the local maxima of $\widetilde{f_Z}(z)$ by checking the first-order conditions:

$$\frac{d}{dz} f_Z(z) = 0$$

$$\frac{d^2}{dz^2} f_Z(z) < 0$$

3. We retain only those z_i's with amplitudes of at least 30% of the global maximum of $f_Z(z)$. [5]

4. We record the local maxima z_i's and their corresponding function value $f_Z(z_i)$'s.

For example, Figure 7 shows the original and smoothed density function for redshift of galaxy number 42. Two significant modes are identified for this galaxy using the method described above.

Table 4: Number of modes identified in each galaxy by count

Number of modes	1	2	3	4	5	6
Count	3266	90	25	3	0	1

When we apply this algorithm to all 3385 galaxies in the GOODSS field, 118 of them have more than one mode. The largest number of modes is six for galaxy number 848. For galaxies with more than one mode identified, we take the most significant mode (the local maximum of density function with the greatest amplitude) as its point estimate of redshift.

[5] The purpose of thresholding the value of $f_Z(z)$ is to eliminate spurious peak points in regions with very low density.

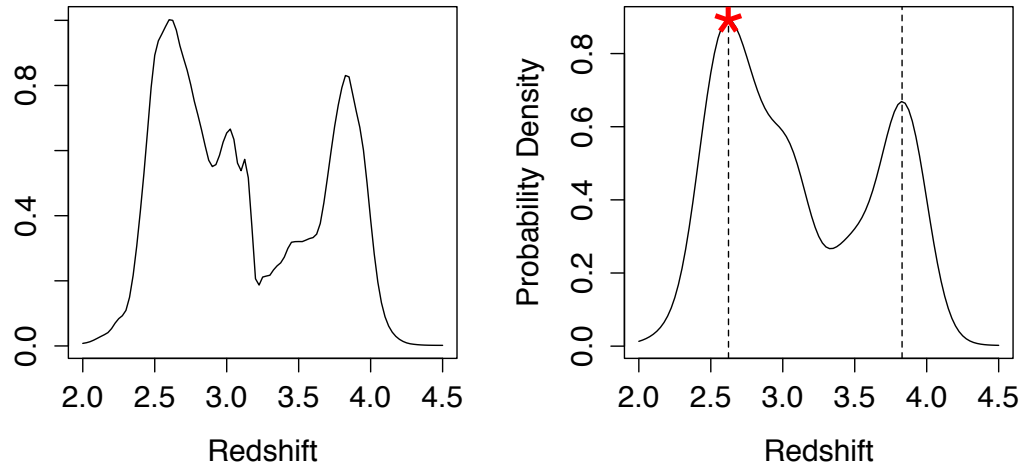

Figure 7: Noise leads to many spurious peaks in the density function for redshift of galaxy number 42. The smoothed version of the density function, however, helps us easily identify the two significant modes at $z = 2.62$ and $z = 3.83$.

3.3.2 Getting weighted statistics for galaxies

Since now we are retaining all galaxies in our analysis, including the ones not mapped to any redshift bins as in 3.2.1. There is no longer one particular filter for which the observed statistics are natural choices to represent a galaxy. One way of addressing the problem is to treat each galaxy as a weighted observation under different filters. We use the probabilities of the galaxy falling into corresponding redshift bins as weights. For galaxies in GOODSS field, the probabilities can be calculated easily using the continuous density function obtained as in 3.1. For galaxies in the GOODSN field, we get the weights by assuming its redshift is has a normal distribution centered at its point estimate with standard deviation 0.27.

Table 5 illustrates how we transform the dataset into a weighted data set with redshift on a continuous scale. Here the binning probability is the probability of a 4500$\mathring{A}$ photon of a galaxy being observed by a given band filter, as described in section 3.2.1. For example, if the probability density function for a galaxy's redshift is $f_Z(z)$, its binning probability for Y band is

$$P(\text{Galaxy falls in Y band}) = \int_{1.05}^{1.63} f_Z(z)dz$$

3.3.3 Estimating $E(Y|Z)$ using kernel regression

To estimate the conditional mean $E(Y|Z)$ we perform a standard univariate regression for each component of the feature vector Y. There is a rich literature covering various methods of nonparametric conditional mean estimation, such as Wasserman (2006). Frequently used methods include local polynomial regression, kernel regression, regression splines, etc. For kernel regression, data-driven approaches have emerged to pick

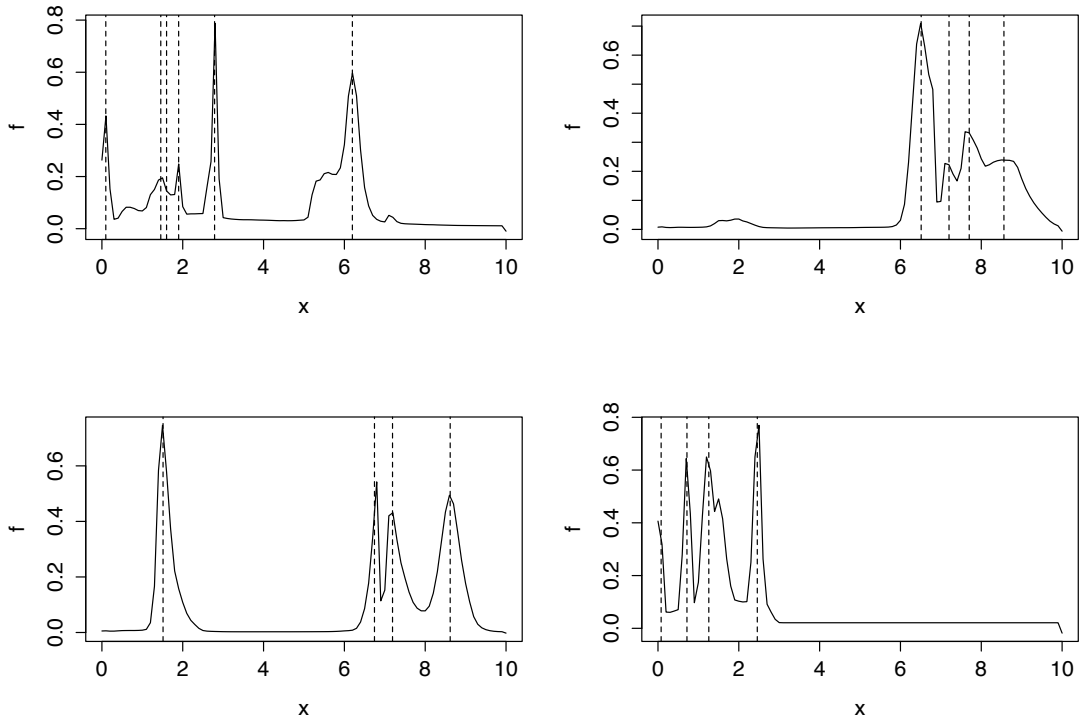

Figure 8: These are some multimodal galaxies identified in the procedure described above. Most of the spurious modes are eliminated.

the optimal tuning parameter (bandwidth h). We use kernel regression in our modeling here.

We pick the optimal bandwidth by 10-fold cross validation. We randomly split the dataset into ten portions of equal size. Given each bandwidth value, we train ten kernel regression models for each variable, leaving one portion out each time. Then we calculate the L^2 loss of the regression function as

$$L_j(h) = \frac{1}{10}\sum_{i=1}^{10} \left| \hat{f}_h^{(-i)}(Z^{(i)}) - Y_j^{(i)} \right|_2^2$$

$$j = 1, 2, \cdots, 7$$

where $\hat{f}_h^{(-i)}(Z^{(i)})$ is the vector of predicted values for redshift values in the i$^{\text{th}}$ portion, h is the bandwidth used, and $\vec{Y_j^{(i)}}$ is the vector of true values of the j$^{\text{th}}$ statistic in the i$^{\text{th}}$ portion.

We minimize this function with respect to h to get the optimal bandwidth for each statistic.

3.3.4 Confidence bands for kernel regression

We used design matrix bootstrap to get confidence bands for kernel regression. We sampled with replacement from the original data matrix a data matrix with the same

Table 5: Illustration of getting weighted M statistics for Galaxy No. 42

Galaxy No. 42

Band	M Statistic	Binning Probability
Y	0.0756	0.119
J	0.0167	0.234
H	0.0028	0.646

Principal mode of redshift: Z^* =2.62

=>

Z	M Statistic	Weight
⋮	⋮	⋮
2.62	0.0756	0.119
2.62	0.0167	0.234
2.62	0.0028	0.646
⋮	⋮	⋮

number of rows. We then fitted the kernel regression model with the resampled data set, made predictions and obtained confidence interval for prediction at each redshift value by taking percentiles at 2.5% and 97.5% of its repeated predictions.

3.3.5 Significance test for kernel regression

In order to make a formal statement about morphology evolution, we need to test whether the conditional distributions of feature statistics vary significantly with redshift.

Given the kernel regression results we obtained, we would want to know whether the change in conditional mean of feature statistics over redshift is significant. We use the simulation based significance test introduced in Racine (1997).

Here the null and alternative hypothesis are:

$$H_0 : \lambda = E[\frac{\partial E(Y|Z)}{\partial Z}]^2 = 0$$

$$H_A : \lambda = E[\frac{\partial E(Y|Z)}{\partial Z}]^2 > 0$$

In other words, we want to test whether the feature statistics Y change with Z at all in the whole range of redshift $Z \in [0, 10]$. If the null hypothesis holds and Y doesn't vary with Z, then the partial derivative should be uniformly 0. Otherwise the partial derivative should be different from 0 at some point.

The test statistic is constructed as the sample analog of $E[\frac{\partial E(Y|Z)}{\partial Z}]^2$:

$$\hat{t} = \frac{\hat{\lambda}}{SE(\hat{\lambda})}$$

where $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=j}^{p} [\frac{\hat{\beta}_h(z_i)}{SE(\hat{\beta}_h(z_i))}]^2$

$\hat{\beta}_h(z_i)$ is the partial derivative estimate given by kernel regression at the i-th data point. The derivative estimate is summed over different bandwidth values so that the choice of bandwidth won't affect the outcome: if Y does change with Z, the estimate of derivative will be different from 0 at least under appropriate smoothing; if Y doens't change with Z, the estimate of derivative should be 0 no matter the smoothing.

The standard deviation of $\hat{\lambda}$ is obtained through bootstrap. From the sample used to compute the test statistic, resample is drawn and $\hat{\lambda}$ is calculated again. Repeat this enough times (1000 in our case) and we can have an estimate of $SE(\hat{\lambda})$ and perform the one-sided test on $\hat{t}$.

3.3.6 Quantile regression for the interval estimate of feature statistics

As can be seen from Figure 5, most of the feature statistics we are interested in do not have ideal bell-shape distributions. They tend to be heavily skewed and have lots of outliers. Moreover, the range limits on several statistics such as M_n makes common interval estimate methods quite useless. We hence consider nonparametric quantile regression methods to get the conditional quantile estimates of these statistics.

In contrast to ordinary regression, which focuses its interest on conditional expectation, quantile regression helps us get precise information about the conditional quantiles of the response variable. It gives more detailed information about the response variable and is very useful for analyzing data with heteroskedasticity and/or non-Gaussian errors.

The method we use here is developed in Koenker et al. (1994) and the implementation is described in Koenker and Ng (2005). The model estimate is determined by total variation regularization. In other words, it is the solution to

$$\min_{g\in\mathscr{G}} \sum_{i=1}^{n} \rho_\tau(y_i - g(Z_i)) + \lambda \left(\int_0^{10} |g''(Z)| dx \right)$$

where $\rho_\tau(y) = y(\tau - I(y < 0))$, τ is the desired quantile for the regression.

Three quantiles we are especially interested in is the 25th percentile, the median, and the 75th percentile, since they are usually the benchmark quantiles of a distribution. Also, they correspond directly to the quantiles shown in box plots.

3.3.7 Confidence intervals for quantile regression

We used design matrix bootstrap for obtaining error estimates for the quantile regression predictions. In other words, given a fixed band filter (H, J, or Y), for any redshift level Z, we draw a sample of the same size from the empirical distribution of our data points. Get a new model estimate along with its fitted value $\hat{g^*}(Z)$, and then repeat. This process can be repeated B (500 in our case) times, to yield bootstrap sample $\hat{g^*}(z)_{(1)}, \cdots, \hat{g^*}(z)_{(B)}$.

We can then estimate the variance of the fitted value and thus give a confidence interval using the pivotal method, or we can simple get the percentiles at 2.5% and 97.5% to get the 95% confidence interval of the fitted value at each redshift level.

4 Results

4.1 Discrete bin approach

4.1.1 Comparison between two binning methods

As we mentioned in 3.2, since different data are available for different fields, we used different methods for mapping galaxies into redshift bins. For GOODSS galaxies whose photometric redshift measurements are entirely available, we associate a galaxy with a bin if the probability of its redshift being within that bin exceeds a certain threshold (either 80% or 60%). For the GOODSN ~~field~~ galaxies (use this instead of field for consistency), since we only have a point estimate for redshift, a galaxy is associated with a redshift bin if the point estimate is within its redshift range. For the overlapping redshift range between Y band and J band, we scale the distance of the point estimate to the center of the band by the width of the

New paragraph (separate idea)

band's redshift range and put the galaxy into the closest redshift bin. One might wonder whether using these two methods will lead to significantly different results. And since we also have access to point estimates of redshift for galaxies in field GOODSS, we have a natural way of testing whether the two methods give similar binning results using the GOODSS field galaxy redshift measurements.

Can you expand on this a little bit? What do you mean by different results? How might this look in your confusion matrices?

We bin each galaxy in GOODSS field just as we described in 3.2.1. We also bin each galaxy in GOODSS field using the most significant mode identified as in 3.3.1 and associate a galaxy with a redshift bin if its most significant mode is within the redshift range of that bin. We show the confusion matrix between the two binning methods below in Table 6.

Might this be better communicated with a graph? With bar plots perhaps? I think this because the specific numbers don't mean much to me. What I'm interested in is which one is higher or lower. I want to make comparisons. A graph might make that easier. (Tufte would agree, I think.)

Table 6: Comparison between binning results

$Bin_{point} \backslash Bin_{photo}$	Y	J	H	No bins
Y	423	0	0	75
J	2	524	0	63
H	0	0	534	183
No Bins	0	0	0	1581

Note: Bin_{photo} refers to the binning result given by the method using the whole distribution and Bin_{point} represents the result of binning using only the point estimate of a galaxy. Here Bin_{photo} are determined using threshold 0.8.

$Bin_{point} \backslash Bin_{photo}$	Y	J	H	No bins
Y	476	1	0	21
J	3	571	0	15
H	0	0	651	66
No Bins	0	1	0	1580

Note: Bin_{photo} refers to the binning result given by the method using the whole distribution and Bin_{point} represents the result of binning using only the point estimate of a galaxy. Here Bin_{photo} are determined using threshold 0.6.

Most of the binning results match for the two methods since both confusion matrices are diagonally dominant. When using the 0.8 threshold, 3062 out of 3385 galaxies remain in the same bin (or do not belong to any bin). When using the 0.6 threshold, 3278 galaxies have the same binning result between the two methods. Moreover, most of the discrepancies between two methods are galaxies that are not assigned a redshift bin by the point estimate method but are assigned ones by the photometric method.

4.1.2 GOODSS field

For GOODSS field galaxies, we group the galaxies into redshift bins and then compare the statistics in each bin through exploratory visualization. Figure 10 shows the bin approach comparisons using the 0.8 threshold. The difference is very small between the results using 0.8 threshold and those using 0.6 threshold, which indicates the bin approach is quite robust to change in the threshold used in our binning step.

This shows it's robust to changes from 0.6 to 0.8, but not in general.

We summarize the significant test as described in Section 3.2.3 in Table 7 and Figure 9.

As is believed in cosmology, the younger universe is more chaotic than the current one in the sense that there are more galaxy mergers in it. Both M and I statistics are

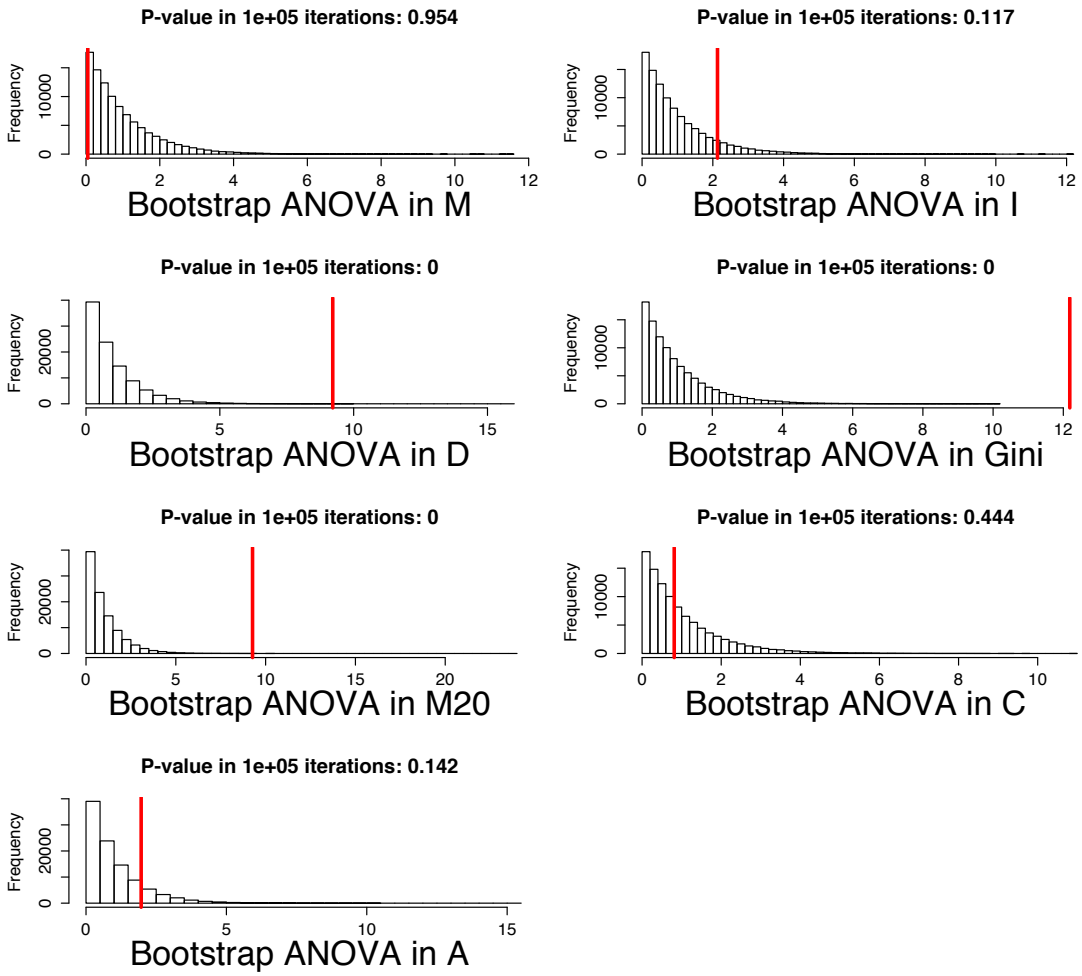

Figure 9: Bootstrap ANOVA for feature statistics in GOODSS field. The histogram is approximate null distribution of the test statistic. The red vertical line represent the statistic value calculated with the observed data.

expected to increase with redshift. Also, proportion of zeroes is expected to decrease in both statistics from band V to band H. The illustrated comparison met all these expectations.

Should this be obvious from the figures? I'm not sure I understand which part of the figures or the analysis showed that the cosmologists' expectations were fulfilled.

4.1.3 GOODSN field

GOODSN is another field where we have statistics taken under all five filter bands available. We map all galaxies to five redshift bins using their redshift point estimate and compare the conditional distributions of feature statistics amongst these redshift bins. Figure 11 shows the comparison. The significance test result is also summarized in Table 7.

This figure is very far away from the text that cites it. Could you say something like, "If the reader wants to see additional confirmation of this she can refer to the appendix"?

4.2 Continuous redshift approach

We show results of bandwidth selection and significance test as described in Section 3.3.3 and 3.3.5 in Table 8.

Table 7: Bootstrap ANOVA test results for discrete bin approach

Statistic	P-value in bootstrap ANOVA test	
	GOODSS field	GOODSN field
M	0.954	0.849
I	0.117	0.307
D	0	0.022
Gini	0	0
M_{20}	0	0.046
C	0.444	0
A	0.142	0

Would it be useful to combine these two tables?

Table 8: Bandwidth selection and significance test results for kernel regression

Statistic	Optimal Bandwidth	P-value in Significance Test
M	0.218	0.01
I	0.184	0.005
D	0.218	$<10^{-12}$
Gini	0.0545	$<10^{-12}$
M_{20}	0.179	$<10^{-12}$
C	0.109	0.058
A	0.065	$<10^{-12}$

Beyond the summary statistic C, where test is borderline significant, all tests have very low P-values. Put in context, we know the evolution in morphology over cosmic time is significant shown by the evolution of the seven feature statistics.

Maybe there is a better way to say this. Something like, "In other words, we have evidence for our hypothesis that there is evolution in galaxy morphology over cosmic time."

4.3 Comparison between our two estimation approaches

An important question we are interested in answering is whether the two approaches would give quantitatively similar results. Since discrete bin approach is standard practice in astronomy, it is important that our method gives consistent results in the same range.

Again, in terms of what?

Same comment as above. The figures are very far away. Do you really need to show them? If you do then you might want to put them in text.

We compare the our approaches by showing their results in the same plot in Figure 12 and Figure 16. In the redshift range of each redshift bins, the predicted conditional mean, the conditional median, and the conditional 25th and 75th percentiles roughly match the mean, median, 25th and 75th percentiles. Trends in all features statistics are captured by the continuous redshift approach.

Looking at conditional expectation predictions first, Figure 13 shows comparison between kernel regression predictions with confidence bands and means of each filter band from the discrete bin approach. The 95% confidence intervals cover the bin means most of the time. The same is true for the comparison using data in GOODSN field from Figure 15.

What is the implication of this?

The result obtained by design matrix bootstrap is shown in Figure 14. The 95% intervals cover the quantiles obtained from discrete bin approach most of the time. Also, as redshift increase to high values (~ 4) the confidence bands become very large, indicating the conditional quantile estimates in high redshift region are no longer reliable.

In the GOODSN field, we make the similar comparison between two approaches. And similarly, the two approaches do not match well using all five band filters, yet they match well using only Y, J, H bands, in Figure 16. And as in GOODSS field, kernel regression shows high prediction power in estimating the conditional mean of feature statistics over different redshift value, as shown in Figure 15.

Similarly we compare the confidence bands obtained by bootstrap in the GOODSN field with conditional quantiles obtained via discrete bin approach. The result is shown in Figure 17.

In brief, the above results indicate the continuous redshift approach gives largely consistent results compared to the discrete bin approach. We could rely on the result from the continuous redshift approach as an extension of the discrete bin approach.

This is good.

5 Conclusion

5.1 Conclusion in the context of galaxy morphology evolution

As shown by both discrete bin approach and continuous redshift approach, we detect significant sign of galaxy morphology evolution over cosmic time. In other words, the distributions of feature statistics do evolve as a function of redshift.

The M, I and D statistics all increase significantly with redshift, indicating merger activity is more frequent in earlier universe. However, the conditional 25th percentile and the median do not rise much with redshift increasing, only the 75th percentile rise significantly with redshift, indicating a relatively small portion of galaxies are driving the trend in M, I and D via their merging activities. The Gini, M_{20} and C statistics first rise in the low redshift ($Z < 0.5$) area and then fall slowly, which might mean the galaxies in the early time universe have their light more evenly spread out. The A statistic decreases with redshift, indicating galaxies in the early time universe are more symmetric compared to galaxies in later time.

5.2 Conclusion in terms of methodology

Our results show that when we have access to photometric redshift measurements, we can extend the discrete bin approach to a continuous estimate of the distributions of feature statistics as a function of redshift. This method retains all data points, gives more powerful test results while yielding consistent results compared to the discrete bin approach. Also, we can depict the change in the distributions beyond the redshift bins associated with band filters. We consider this our most important contribution in this work, since with the same data set we are able to give a more detailed depiction of galaxy morphology evolution.

In the continuous redshift approach, we tried to use feature statistics calculated with images from all three filters available. That requires essentially combining three values to get a true value of a galaxy's feature statistic. For example, when we are estimating the conditional expectation of feature statistics, the value of a galaxy's M statistic would be the average of its M statistic calculated with images in three filters, weighted by the probability its redshift falls in the range of each filter. This combining process would require the following two implicit assumptions.

First, there has to be a certain linearity in the feature statistic of galaxies. Otherwise, the weighted average of feature statistics calculated from images taken in different filters wouldn't be a plausible approximation for a galaxy's true feature statistic in rest-frame wavelength. Figure 13 shows that the conditional expectation estimates differ the least between two approaches for feature statistic M, I, and D, which indicates these statistics fit the linearity requirement better than the other statistics.

Second, the weight obtaining process requires reliable photometric redshift measurement over the entire range of redshift, otherwise inaccurate weights would make the results systematically wrong. Figure 12 shows that the continuous approach in general deviate farther from the discrete bin approach in the H band redshift range. This indicates there might be systematic bias in the photometric redshift measurements we have.

In short, the consistency of our continuous redshift approach is roughly satisfying yet the degree of consistency varies from statistic to statistic and from filter to filter. We would think it comes from properties of feature statistics and the photometric redshift measurement we are using. A possible direction for future work would be resolving this discrepancy.

References

Conselice, C. J. (2003). The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1.

Freeman, P., Izbicki, R., Lee, A., Newman, J., Conselice, C., Koekemoer, A., Lotz, J., and Mozena, M. (2013). New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, page stt1016.

Fritsch, F. N. and Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246.

Grogin, N. A., Kocevski, D. D., Faber, S., Ferguson, H. C., Koekemoer, A. M., Riess, A. G., Acquaviva, V., Alexander, D. M., Almaini, O., Ashby, M. L., et al. (2011). Candels: The cosmic assembly near-infrared deep extragalactic legacy survey. *The Astrophysical Journal Supplement Series*, 197(2):35.

Koekemoer, A. M., Faber, S., Ferguson, H. C., Grogin, N. A., Kocevski, D. D., Koo, D. C., Lai, K., Lotz, J. M., Lucas, R. A., McGrath, E. J., et al. (2011). Candels: The cosmic assembly near-infrared deep extragalactic legacy surveythe hubble space telescope observations, imaging data products, and mosaics. *The Astrophysical Journal Supplement Series*, 197(2):36.

Koenker, R. and Ng, P. (2005). A frisch-newton algorithm for sparse quantile regression. *Acta Mathematicae Applicatae Sinica*, 21(2):225–236.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4):673–680.

Lotz, J. M., Primack, J., and Madau, P. (2004). A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163.

Racine, J. (1997). Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics*, 15(3):369–378.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Zhou, B. and Wong, W. H. (2011). A bootstrap-based non-parametric anova method with applications to factorial microarray data. *Statistica Sinica*, 21(2):495.

A Appendix: Definition of feature statistics

1. **Multimode (M) statistic**
   Given a quantile q_l(such that l of the pixels have intensity not greater than q_l) for pixel-wise intensity of an image, define the following new image,

$$g_{i,j} = \begin{cases} 1 & : f_{i,j} \geq q_I \\ 0 & : otherwise \end{cases}$$

   where $f_{i,j}$ is the intensity of pixel (i, j).

   Then let $A_{l,(1)}$ and $A_{l,(2)}$ denote the number of pixels in the two largest groups of contiguous pixels given the quantile q_l. Let n_{set} denote the number of pixels in segmentation map, i.e. the galaxy mask. Define the area ratio for each quantile

$$R_l = \frac{A_{l,(2)}^2}{A_{l,(1)} n_{seg}}$$

   and the M statistic is the maximum R_l value over all quantile l's,

$$M = \max_l R_l$$

   The M statistic characterizes whether a galaxy is multimodal. Values closer to 0.5 means the galaxy might have more than one mode.

2. **Intensity (I) statistic**
   In computing the I statistic, each pixel in the image is associated with a local maximum in pixel intensity using the gradient ascent method (also known as mode clustering). And the I statistic is the ratio between the total intensity of the two regions with greatest total intensity,

$$I = \frac{I_{(2)}}{I_{(1)}}$$

   I is the statistic to complement M in characterizing whether a galaxy has more than one bright regions, taking the intensity of regions into account.

3. **Deviation (D) statistic** The intensity centroid of a galaxy is defined as

$$(x_{cen}, y_{cen}) = \left(\frac{1}{n_{seg}} \sum_i \sum_j i f_{i,j}, \frac{1}{n_{seg}} \sum_i \sum_j j f_{i,j} \right)$$

   where the summations are over all n_{seg} pixels in the image, and the D statistic is

$$D = \sqrt{\frac{\pi}{n_{seg}}} \sqrt{(x_{cen} - x_{I_{(1)}})^2 + (y_{cen} - y_{I_{(1)}})^2}$$

   where $(x_{I_{(1)}}, y_{I_{(1)}})$ is the center of the pixel group with highest summed intensities. By calculating the normalized distance between the center of the whole image and the center of the brightest region, the D statistic depicts how a galaxy deviates from a spherical or disc shape.
   M, I, D statistics are introduced in Freeman et al. (2013).

4. **Gini (G) statistic**
   The Gini coefficient of the distribution of absolute flux values is defined as

$$G = \frac{1}{|\bar{X}|n(n-1)} \sum_i (2i - n - 1)|X_i|$$

   where X_i's are the absolute flux values of each pixel.
   The Gini coefficient describes the inequality of a galaxy, i.e. whether a majority of a galaxy's flux is concentrated within a small area. The higher the Gini coefficient, the more concentrated light is among the image of a galaxy.

5. **M_{20}**
   The total second-order moment M_tot is the flux in each pixel f_i multiplied by the squared distance from the center of the galaxy.

$$M_{tot} = \sum_i M_i = \sum_{i=1}^{n} f_i[(x_i - x_c)^2 + (y_i - y_c)^2]$$

   where (x_c, y_c) is the center of the galaxy, which is calculated by minimizing M_{tot}. M_{20} is defined as the normalized second-order moment of the brightest 20% of the galaxy's flux.

$$M_{20} = \log_{10}\left(\frac{\sum_i M_i}{M_{tot}}\right) :, where \sum_i f_i < 0.2 f_{tot}$$

   The M_{20} statistic depicts the distribution of bright regions off the center of the galaxy.
   Both G and M_{20} are defined in Lotz et al. (2004).

6. **Concentration (C) statistic**
   Concentration is defined as the ratio of 80%-20% curve of growth radii (r_{80}, r_{20}) normalized using a logarithm.

   First the value for the galaxy's radius r is needed. Let $I(r, \theta)$ be the intensity of a galaxy's light, where r is defined relative to the galaxy catalog position. To calculate the radius, define the annular surface brightness $\mu(r)$ as follows,

$$\mu(r) = \frac{\int_0^{2\pi} \int_{r-\delta r}^{r+\delta r} I(r', \theta) r' dr' d\theta}{\int_0^{2\pi} \int_{r-\delta r}^{r+\delta r} r' dr' d\theta}$$

   define the average surface brightness $\bar{mu}(r)$,

$$\bar{mu}(r) = \frac{\int_0^{2\pi} \int_0^{r+\delta r} I(r', \theta) r' dr' d\theta}{\int_0^{2\pi} \int_0^{r+\delta r} r' dr' d\theta}$$

   r is then defined as the solution of equation

$$\frac{\mu(r)}{\bar{\mu}(r)} = \varepsilon$$

The concentration statistic is then defined as

$$C = 5 \times \log(r_{80\%}/r_{20\%})$$

The concentration statistic measures the concentration of a galaxy's stellar light distribution.

7. **Asymmetry (A) statistic**
   The asymmetry statistic compares the image of a galaxy and its appearance after a 180° rotation. It is defined as follows

$$A = \frac{\sum_S |O_{i,j} - R_{i,j}| - (n_S/n_{S'}) \sum_{S'} |O_{i,j} - R_{i,j}|}{2 \sum_S |O_{i,j}|}$$

   where O represents the galaxy's image and R its image after a 180° rotation. S is the set of pixels inside the segmentation map and S' is the set of postage-stamp pixels outside the segmentation map.
   By comparing these two images, the A statistic depicts the degree of asymmetry of a galaxy. Both A and C statistics are defined in Conselice (2003).

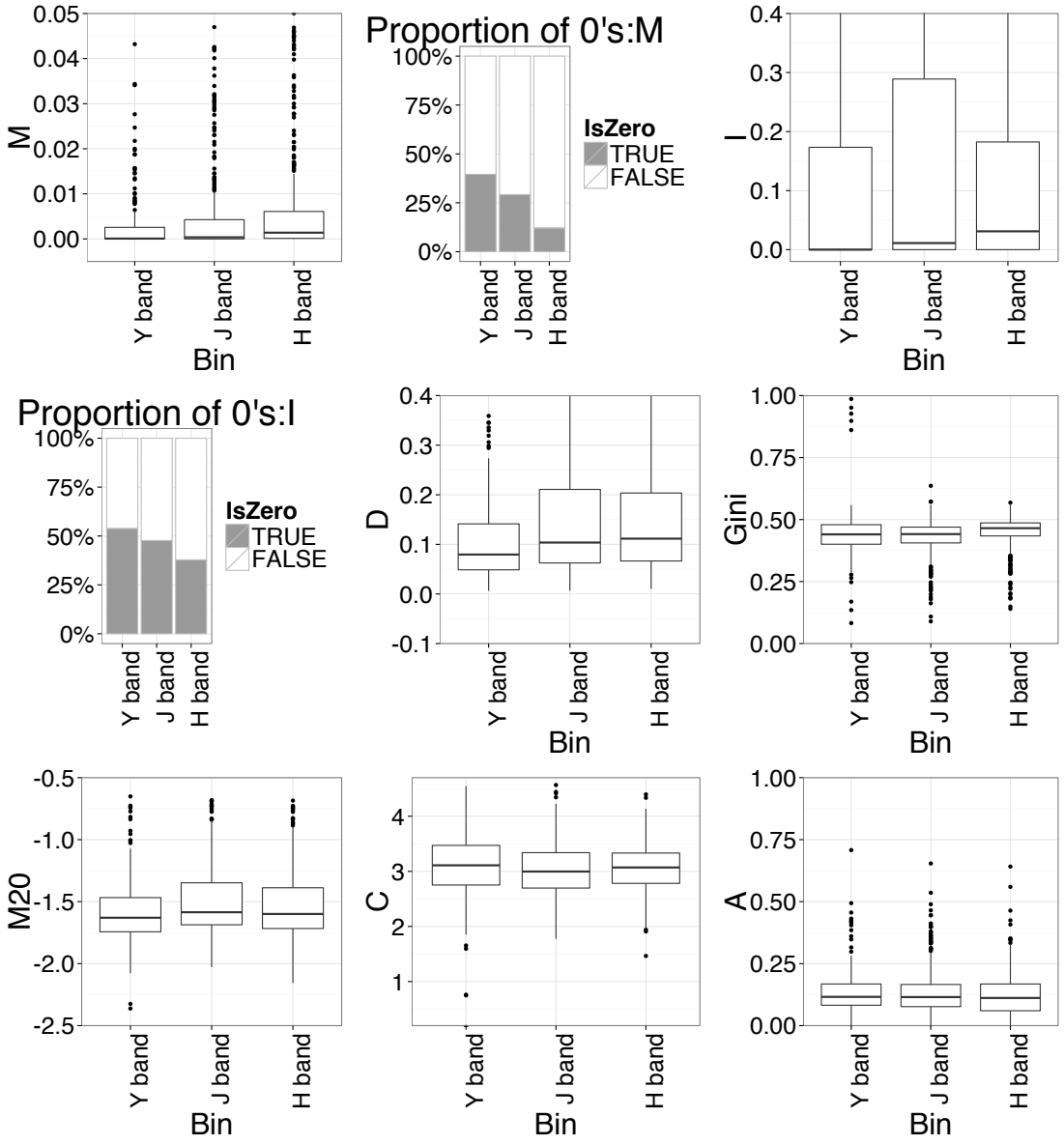

Figure 10: Side-by-side box plots for feature statistics for GOODSS field galaxies in different redshift bins. A 0.8 threshold is used in determining redshift bins for each galaxy. Since the M and I statistics are both zero inflated, we show box plots as well as bar plots to show portion of zeroes in the statistics.

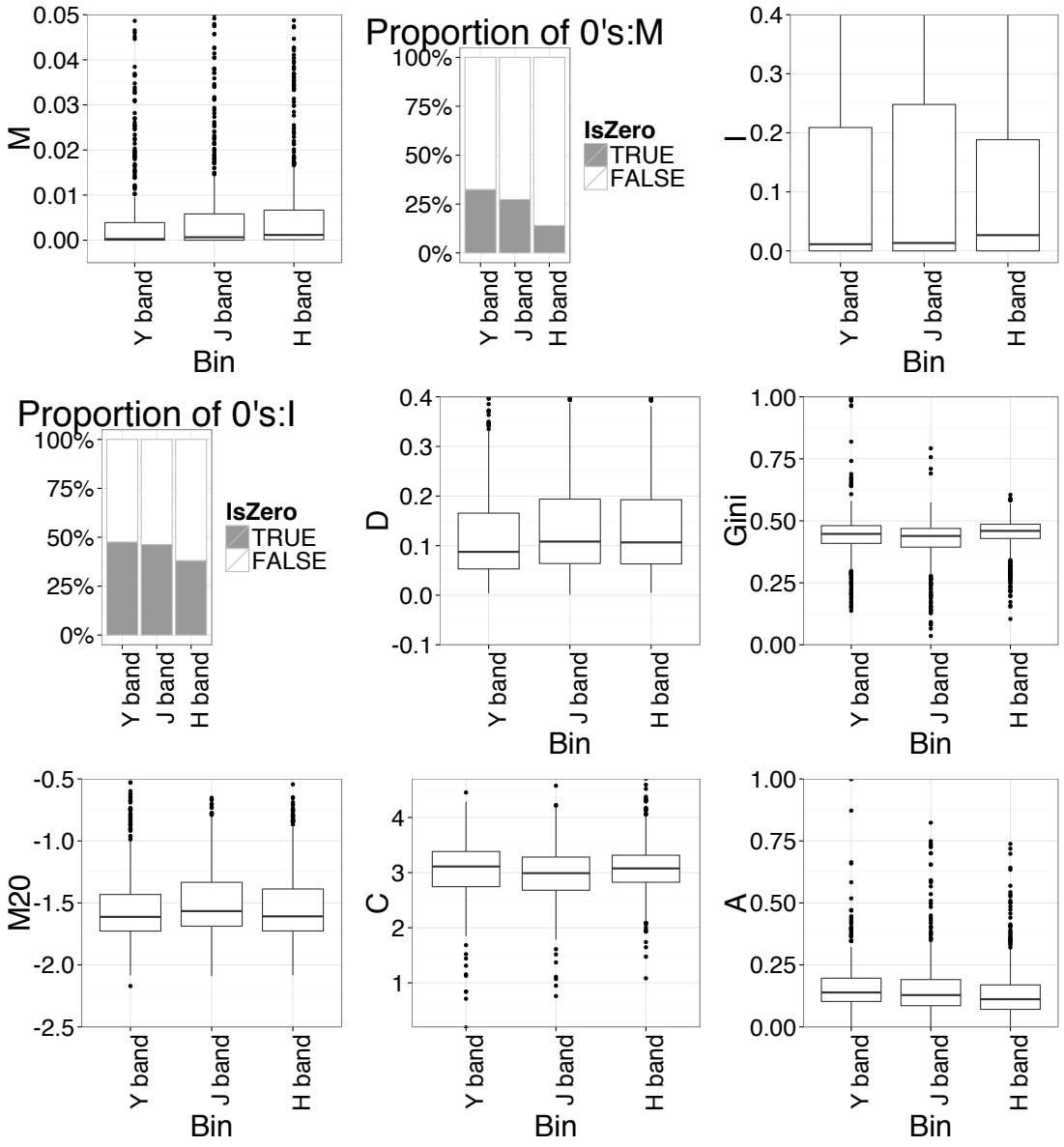

Figure 11: Same as Figure 10, for the GOODSN field. We see slight differences in behavior of portion of zeroes in M and I, but the trend is still largely consistent with Figure 10.

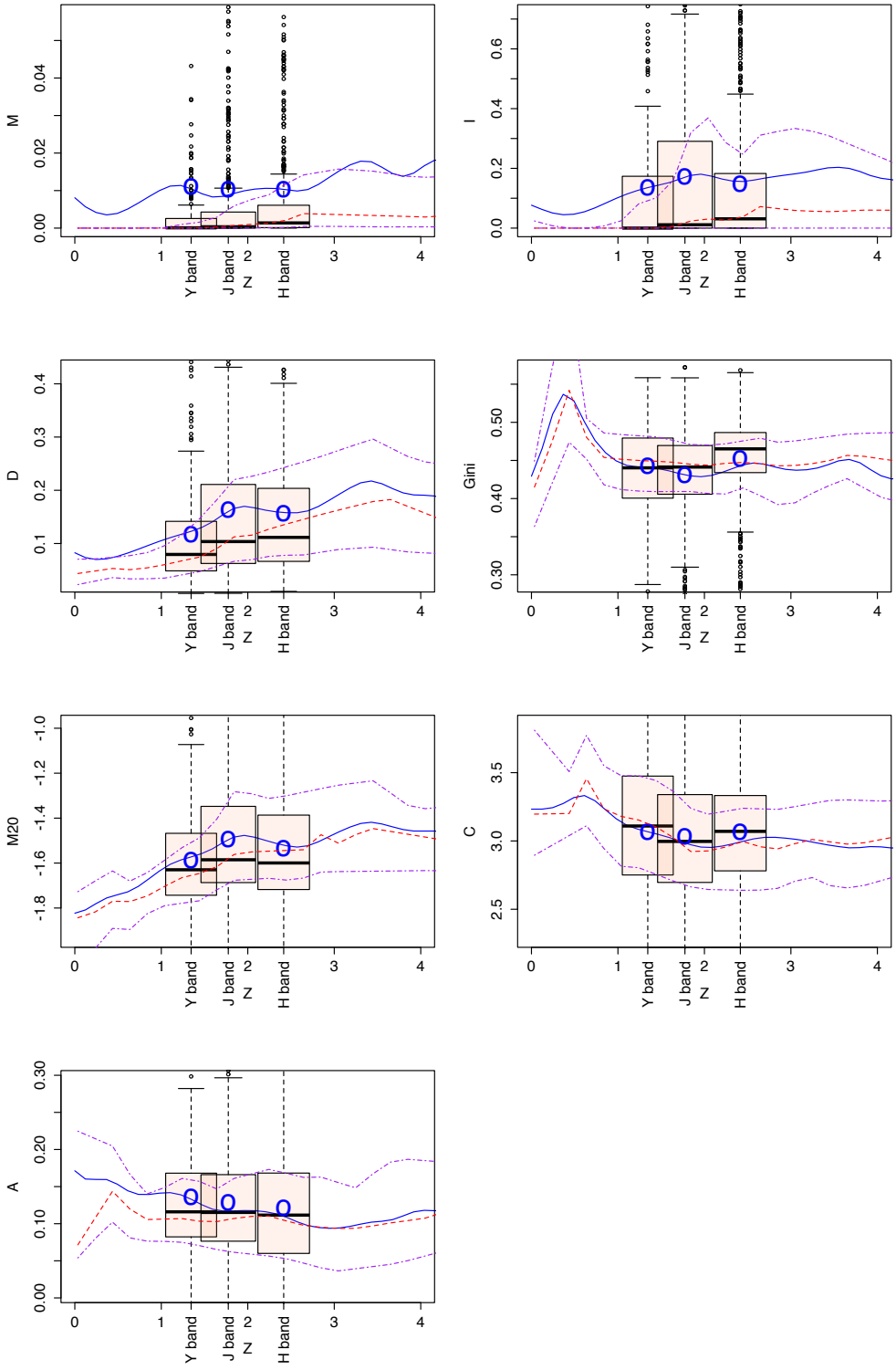

Figure 12: Comparison between continuous approach and bin approach for galaxies in GOODSS fields. The horizontal line in the middle of the box represent the median of each bin. The O's represent the mean of each bin. Purple curves correspond to the 25th and 75th percentile quantile regression. The yellow curve corresponds to the median regression. The blue curve corresponds to the conditional mean curve given by kernel regression.

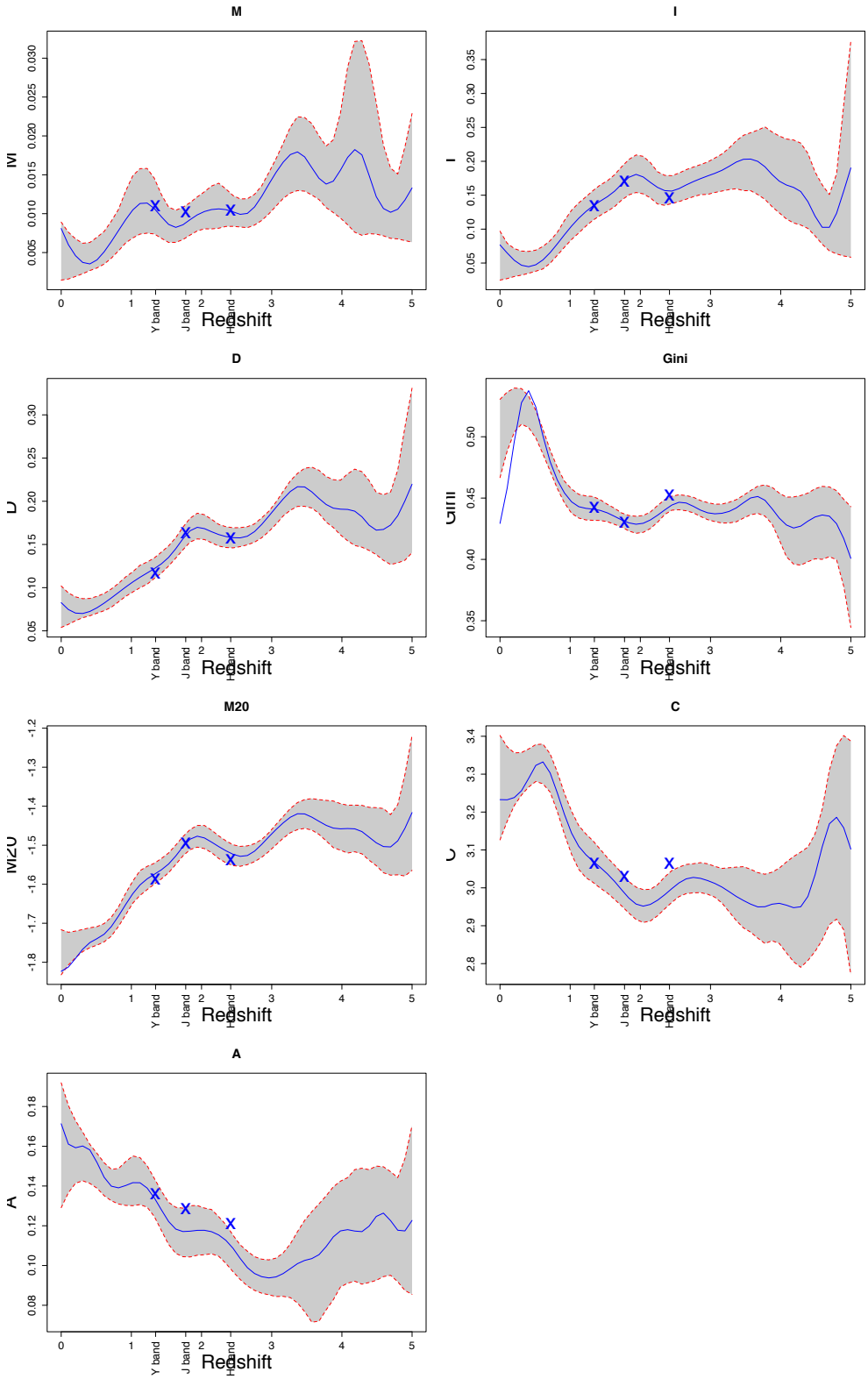

Figure 13: Comparison between the mean of each redshift bin and confidence band of the kernel regression using galaxies in the GOODSS field.

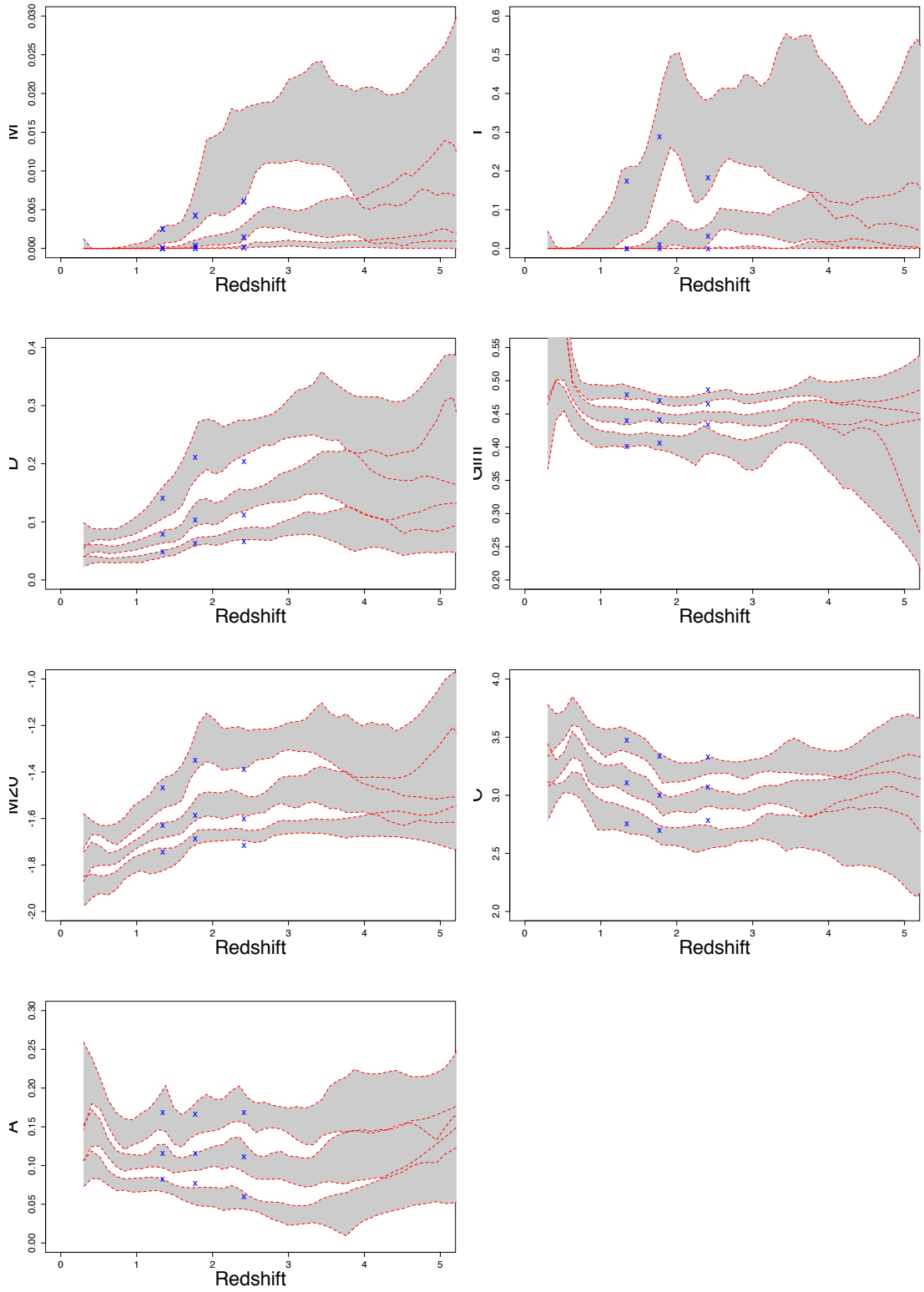

Figure 14: For each conditional τ-quantile, we provide its corresponding interval estimate via design matrix bootstrap, resampling from the galaxy population of GOODSS field 500 times. The blue X's are the 25th percentile, median and 75th percentile for three redshift bins, located at the middle of each redshift bin.

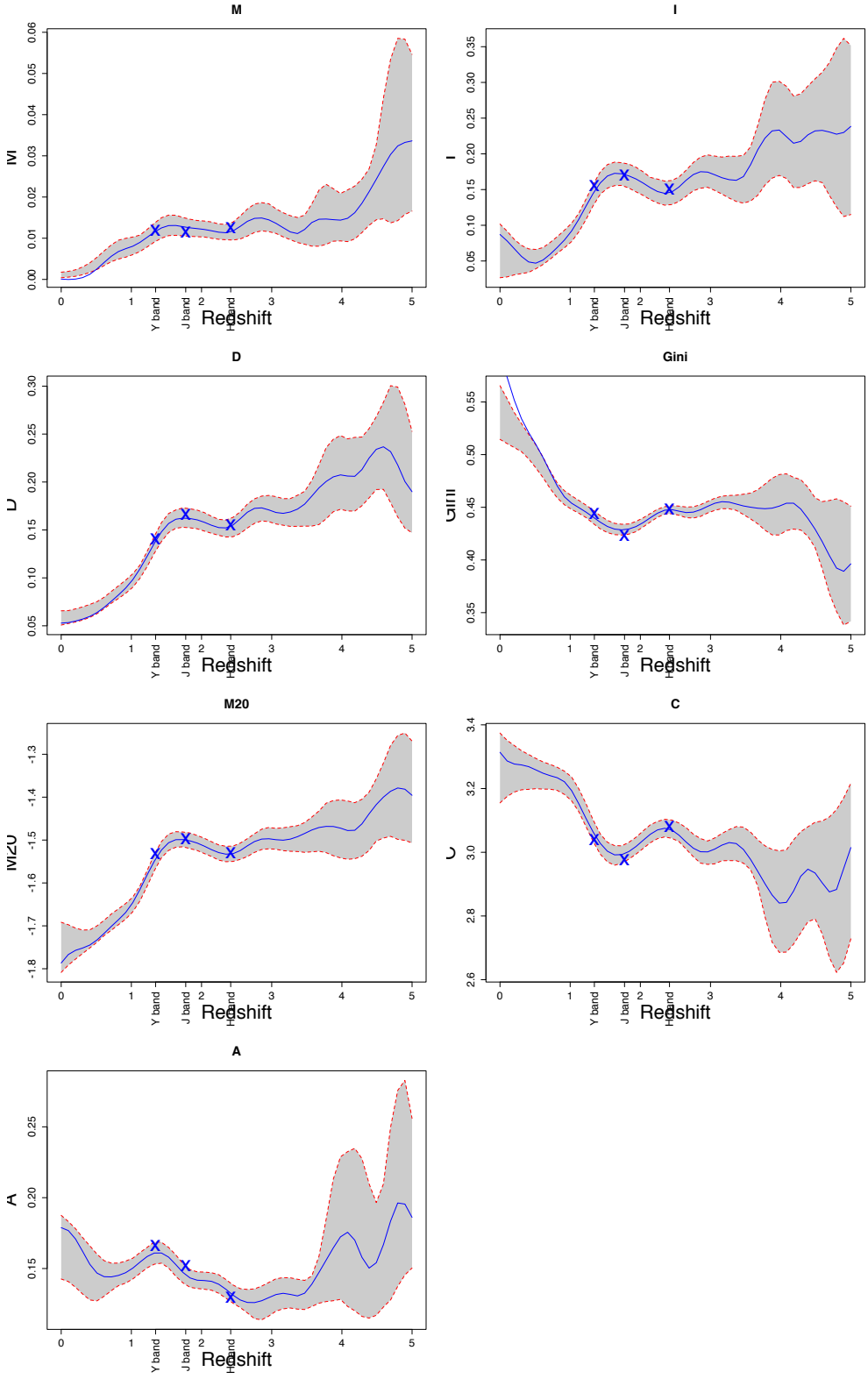

Figure 15: Same as Figure 13, for GOODSN field.

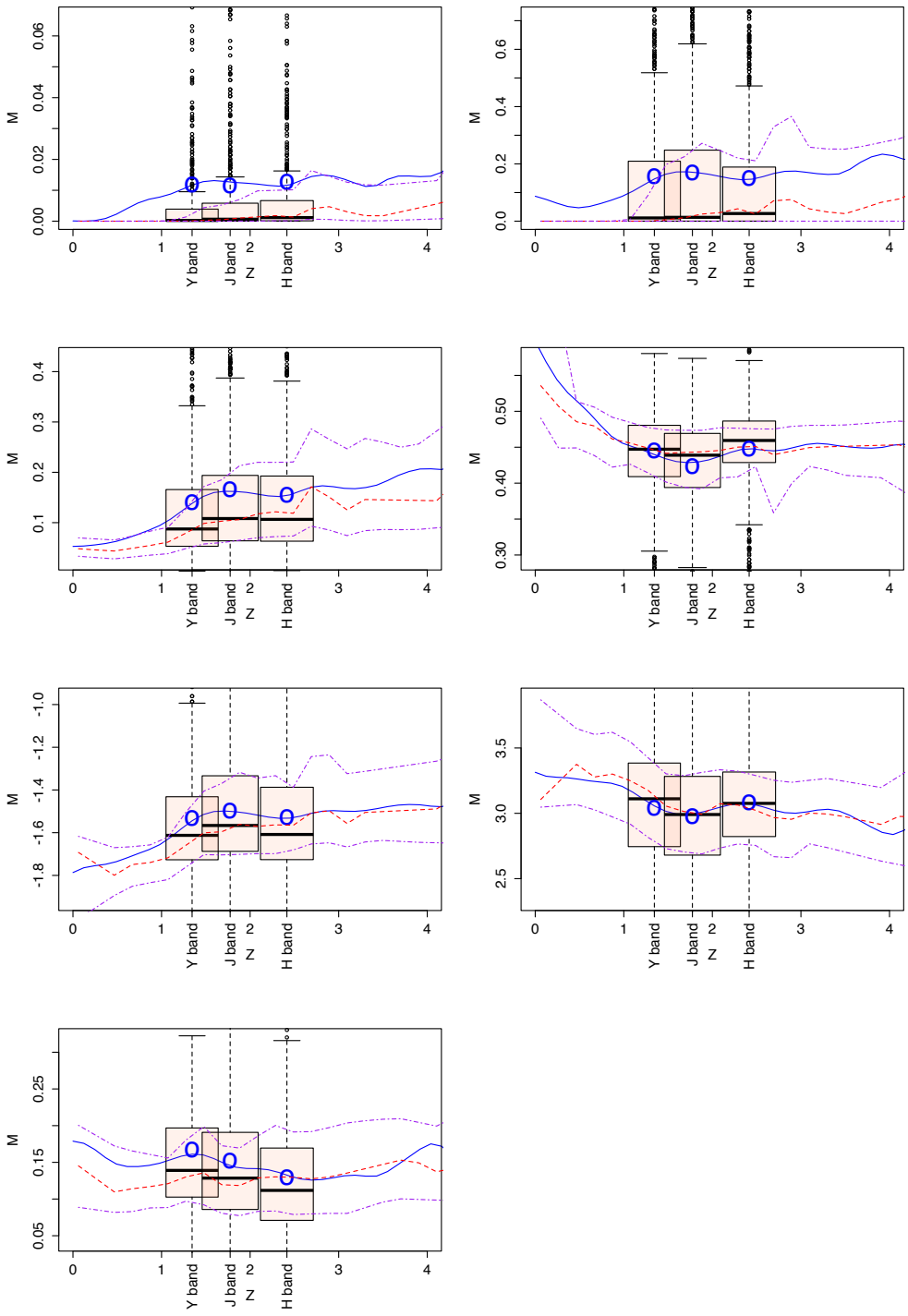

Figure 16: Same as Figure 12, but for GOODSN field using only data from Y, J and H band.

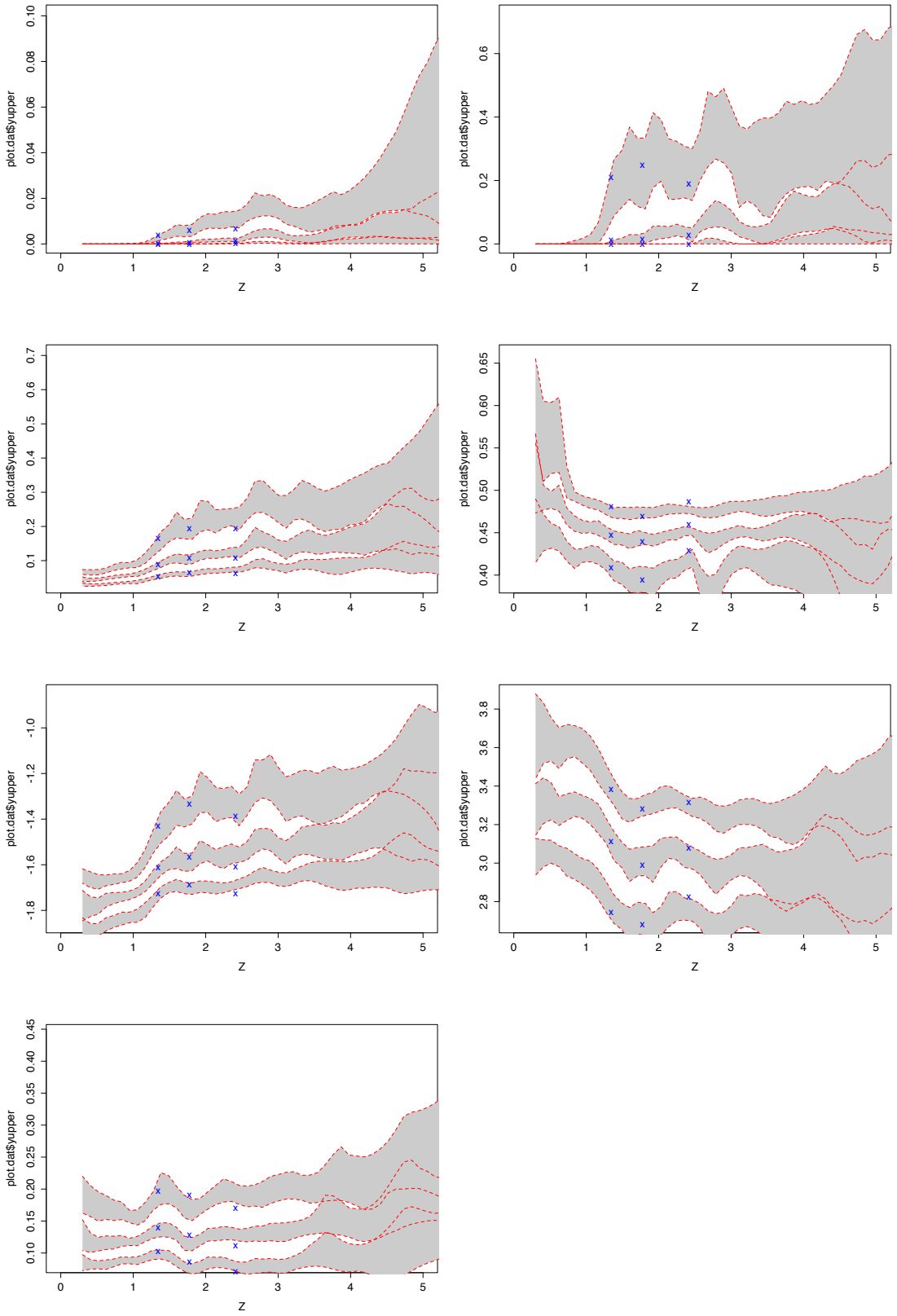

Figure 17: For each conditional τ-quantile, we provide its corresponding interval estimate via design matrix bootstrap, resampling from the galaxy population of GOODSN field 500 times.