Inferring the Evolution of Galaxy Morphology

Robert Lunde

March 29, 2015

1 Introduction

In astronomy, one of the major goals is to put tighter constraints on parameters in the Λ-CDM model, which is currently the standard model describing the evolution of the Universe after the Big Bang. One way to work towards this goal is to estimate how galaxy structure and morphology evolve; we can then compare what we observe with rates predicted by the standard model via simulation.

We can observe galaxies at different points in time by observing those that vary in distance. It takes a longer time for light from farther galaxies to reach Earth and therefore the light we receive from these galaxies carries information from longer ago compared to closer galaxies. Since a galaxy's distance from Earth is not easily estimated, we instead use redshift (z) as a proxy for distance. Redshift is defined as:

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

where $\lambda_{observed}$ is a photon's observed wavelength and $\lambda_{emitted}$ is its emitted wavelength. Light from distant galaxies are redshifted due to the expansion of the Universe; light waves are stretched as space expands, leading to a higher observed wavelength.

After preprocessing, our data consists of 101x101 pixel images of galaxies. If each pixel is treated as a dimension,estimating a distribution of the galaxy image is a challenging problem due to the curse of dimensionality. One way to proceed is to find a low-dimensional representation of this high-dimensional image data that retains important morphological properties. Our approach to this problem thus far has been to use summary statistics defined in the astronomy literature.

2 Data

The dataset comes from GOODS-South and UDS fields of the Cosmic Assembly Near-infrared Deep Extragalactic Survey (CANDELS), which is a survey of the distant Universe conducted by the Hubble Space Telescope (Grogin et al. 2011, Koekoemoer et al. 2011). The galaxy images in our dataset are an averaged and preprocessed version of numerous scans that have been reduced to a set of 101x101 pixel images for each galaxy. For each galaxy, we have images that were taken at commonly used ranges of wavelengths known as photometric bands. To record images at different wavelengths, a filter is placed on the telescope. The amount of light that passes through the filter is given by a measure known as the transmission coefficient. For the GOODS-S field, we have images from the Y Band, J Band,and H Band and for the UDS field, we have images from the J Band, and the H Band. In Figure 1, the transmission coefficient is plotted against wavelength for each photometric band in our dataset.

For each galaxy in our dataset, we also have an estimated probability mass function for z. For closer galaxies, redshifts can be accurately estimated by observing atomic lines via high resolution

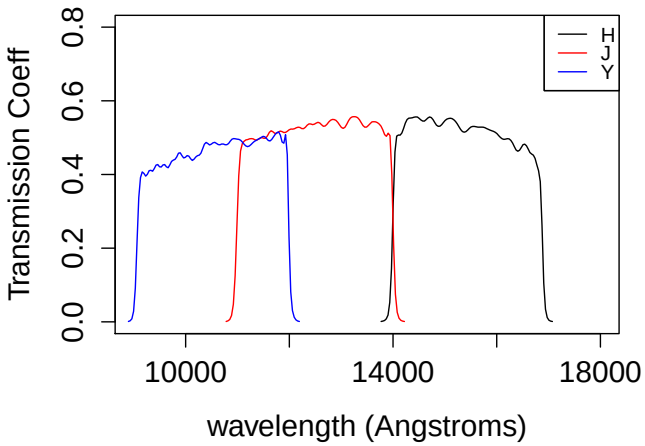

Figure 1: Filter curves for the H, J, and Y Bands, where an Angstrom is defined as 10^{-10} m

spectroscopy. For the galaxies in our our dataset, an algorithm that fits template photometric data to the observed data is used, and we have a probability distribution of potential redshift values (Dahlen 2013). We also have information about various aspects of each galaxy from the catalog, including several different estimates of the mass and a source's signal/noise (S/N) ratio, along with its magnitude (brightness) in the H band.

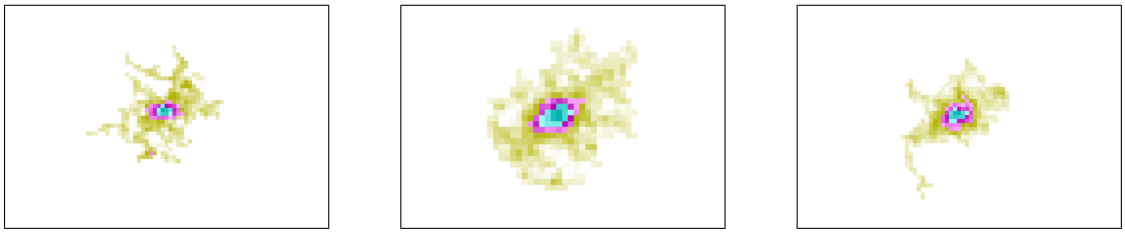

Figure 2: Image of a galaxy taken with, from left to right, the Y-Band, J-Band, and H-Band filters. Blue represents areas with high intensity while yellow represents regions with low intensity.

3 Methods

3.1 Preprocessing

We take several steps to eliminate potential sources of systematic error from our dataset. The first correction we make is to establish a threshold galaxy mass and discard galaxies that are below this threshold. Smaller sources are more difficult to detect at higher redshift ranges; by taking this step, we reduce the risk of biased sampling. Estimates of galaxy mass are provided by CANDELS (Mobasher). For the mass cutoff, we use a measure known as the Wuyts, integrated, exponential declining mass. For the base ten logarithm of the mass, $\log_{10} M$, we use a cutoff of 9.78 for the Y Band, 9.69 for the J band, and 9.55 for the H band (Behroozi). Here, we choose different mass thresholds for different distances to adjust for the change in the median galaxy mass over

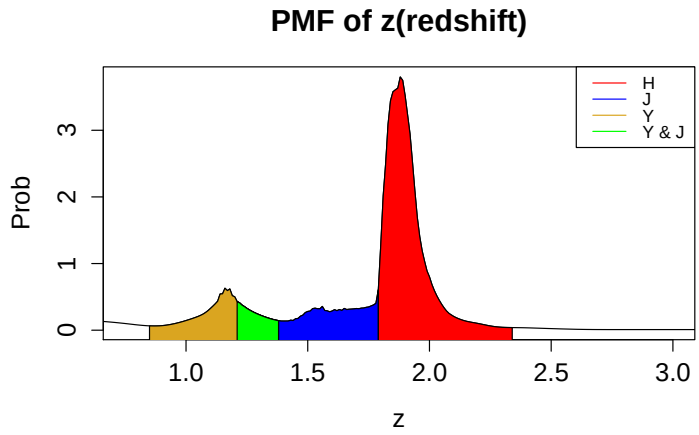

Figure 3: Redshift ranges associated with each band for a given galaxy. The Y and J area corresponds to the overlap between the ranges of the two photometric bands.

time; these cutoffs are equivalent to using a cutoff of 10 for galaxies in the present-day local universe.

Another step in the preprocessing phase is to eliminate differences between the images of the two bands caused by different widths in the point spread function (PSF). The point spread function describes an imaging system's response to a point source. We smooth images from the J-Band using a Gaussian kernel with bandwidth of 0.665 pixels and images from the Y-Band using a Gaussian kernel with bandwidth of 0.754 pixels to eliminate this potential source of bias.

We also use a segmentation map algorithm on the raw galaxy image. The segmentation map reduces noise in the dataset by setting intensity values of pixels unrelated to the morphological structure of the galaxy to zero.

3.2 Redshift Bin Approach

We will examine morphological evolution by comparing the distribution of galaxy morphologies belonging to different redshift ranges, or redshift bins. In our approach, we fix a rest wavelength that corresponds to a wavelength that was originally emitted by the source. The rest wavelength is set to exclude image features unrelated to morphological structure, such as star formation. We choose a rest wavelength of 5000 Angstroms (where 1 Å $=10^{-10}$ m) and define the low and high range for a particular bin according to where the filter curves attain the half-maximum. Using this approach, we find that $z \sim 0.85 - 1.38$ corresponds to the Y Band, $z \sim 1.21 - 1.78$ to the J Band, and $z \sim 1.8 - 2.35$ to the H Band.

For each galaxy, we then compute the probability that it lies in one of the redshift ranges. If the probability exceeds a threshold, then that galaxy is included in the corresponding bin. Image statistics are computed using only the images corresponding to the appropriate photometric band. For our dataset, we establish a threshold that the galaxy must have a probability of at least 0.8 to be included in that bin. Together, the mass uts with the binning procedure leads to a sample size of 124 for the H Band, 326 for the J Band, and 111 observations for the Y-Band.

3.3 Summary Statistics

In our analysis, we will use nonparametric measures of galaxy morphology that are commonly used in the astronomy literature. These measures are designed to capture important structural features of the galaxy without making assumptions about its shape. A brief overview of the statistics is given below; a more in-depth treatment can be found in Freeman et al. (2013), Lotz et al. (2004), and Conselice et al. (2003).

3.3.1 Multimodal (M)

The M statistic is constructed to detect galaxies with To calculate M, a quantile q is fixed and all pixels $f_{i,j}$ that are less intense than the quantile value are assigned value 0, while pixels with a higher value are assigned value 1:

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

Then, ratio of the number of pixels in the two largest contiguous regions is calculated. The ratio is then multiplied by the area of the second largest region to make the statistic more robust to noise. We also normalize this quantity by dividing by the number of pixels in the segmentation map, n_{seg}. The expression for this quantity R_l is given below:

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

Finally, the M statistic is defined as:

$$M = \max_l R_l$$

3.3.2 Intensity (I)

One problem with the M statistic calculated above is that the two largest regions may not contain the most intense pixels. Therefore, we compliment the M statistic with the I statistic. With the I statistic, each pixel is associated with a local maximum of the intensity in a gradient-ascent method; the I statistic is given as the ratio of the summed intensities for the regions with the second-largest and largest summed intensities.

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

3.3.3 Deviation (D)

The Deviation statistic is a measure of how much a galaxy's morphology deviates from elliptic symmetry. To calculate the D statistic, one first computes the center of mass of the intensity (x_{cen}, y_{cen}) :

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

Then, normalized Euclidean distance between the center of mass and the location of the local maximum associated with I_1 is computed. This quantity is the D statistic:

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

3.3.4 Concentration (C)

The Concentration (C) statistic measures how much light is in the center of the galaxy opposed to its outer parts. In general terms, the C statistic finds the ratio of the radius that contain 80 % of a galaxy's flux with that of a radius that encloses 20 % of the flux. To calculate the C statistic, we first compute a galaxy's radius r. 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, we define the annular surface brightness $\mu(r)$:

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

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

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

and calculate the r such that

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

where a value of 0.2 is commonly used for ϵ. When calculating the C statistic, we assume a circular aperture of $1.5r$. Then we compute the total amount of light within the radius, along with the radii that contain 20% and 80% of the light. We finally compute the statistic:

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

3.3.5 Asymmetry (A)

The A statistic calculates the rotational symmetry of the image. The A statistic is computed according to the following formula:

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

where O is the original image, R is the image rotated 180°, S is the pixels inside the segmentation map and s' is the set of postage-stamp pixels outside the segmentation map.

3.3.6 Gini Coefficient

The Gini coefficient is a measure of inequality used in many fields. In the discrete case, the Gini coefficient can be expressed as:

$$G = \frac{1}{c\bar{X}(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|$$

Values closer to 1 would represent high inequality (most of a galaxy's light lies within a relatively few pixels) while values closer to 0 would represent low inequality.

3.3.7 $\mathbf{M_{20}}$

The M_{20} statistic is another statistic that measures the concentration of light in a galaxy image. Whereas the C statistic is designed to be detect the concentration of light in the center the galaxy, the M_20 statistic To compute M_{20}, one first computes M_{tot}, given by:

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

where f_i is the flux in each pixel and x_c and y_c are the galaxy's center. x_c and y_c are the minimizers of M_{tot}. Then we rank the galaxy pixels by flux and sum M_i over the brightest pixels until the sum equals 20 % of the total galaxy flux. We then normalize by M_{tot}, as shown in the equation below:

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

3.4 Methods for Comparing Distributions

We will consider several methods to detect signatures of morphological evolution across bins. Our methods can be grouped into two categories. The first category consists of methods that are designed to detect global differences between distributions. Equivalently, these methods attempt to answer the question: are the distributions significantly different from each other? Our second class of methods examines whether there are local differences between the distributions. That is, are the distributions different if we restrict our attention to a subspace of the feature space?

3.5 Global Comparisons

3.5.1 Univariate Tests

By construction, M > 0 and I > 0 is correlated with the presence of a merger. Therefore, one way we can compare distributions is to test whether the proportion of galaxies with these statistics greater than zero is different across bins. If we observe that the proportion is significantly higher with bins associated with higher redshift, this result will be consistent with the hypothesis that more mergers occured in the past. For each band H, J, and Y, we constuct the statistics:

$$\hat{F}_{band,M,N}(0) = \sum_{i=1}^{N} \mathbb{1}(M < 0)$$

$$\hat{F}_{band,I,N}(\epsilon) = \sum_{i=1}^{N} \mathbb{1}(I < \epsilon)$$

where N is the sample size for the redshift bin of interest and $\epsilon = 0.01$. We use a nonzero value for the I statistic. Since $\hat{F}_{band,M,n}(0)$ and $\hat{F}_{band,I,n}(\epsilon)$ are binomially distributed, we can use a two-sample test for difference of binomial means. Let $X_1, \ldots, X_m \sim P$, $Y_1, \ldots, Y_n \sim Q$, $\hat{p}_1$ be equal to the proportion in P, $\hat{p}_2$ be equal to the proportion in Q, and $\hat{p}$ be the pooled mean. Then the two-sample binomial (Score) test is given by:

$$S = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{m} + \frac{1}{n})}}$$

For the other univariate distributions, there is not an obvious value at which to compare the values of the distribution across bands. Instead, we perform nonparametric univariate two-sample tests to compare the distributions of the summary statistics. In particular, we consider the Mann-Whitney U-statistic, which compares the ranks of the two samples. The U-Statistic is defined as:

$$U := \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}(x_i > y_j)$$

P-values are found by using the fact that U is asymptotically normal.

3.5.2 Multivariate Tests

One simple bivariate comparison we can use to compare distributions is to compare the 95% confidence intervals for the correlation coefficient ρ. For the sample correlation $\hat{\rho}$ we can define the transformation:

$$z = 0.5 \log\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$$

The distribution of this statistic quickly approaches a Normal distribution. We can then construct confidence interval for ρ by transforming the confidence interval back to the original scale.

The kernel Maximum Mean Discrepancy (kernel MMD) is a nonparametric two-sample multivariate test (Gretton 2007). Let $X \sim P$ and $Y \sim Q$. The MMD is given by:

$$MMD(P,Q) := \sup_{f \in \mathcal{F}} [\mathbb{E}f(X) - \mathbb{E}f(Y)]$$

where $\mathcal{F}$ is a Reproducing Kernel Hilbert Space (RKHS) in the unit ball. Intuitively, the kernel MMD is a reasonable way of comparing distributions because it is closely related to the definition of equality in distribution. We say that two distributions are equal if:

$$\mathbb{E}f(X) = \mathbb{E}f(Y) \text{ for all bounded continuous } f.$$

In other words, with the kernel MMD, we replace the space of bounded continuous functions with an RKHS and use the supremum of this statistic for hypothesis testing. We will later show that the supremum has a closed form for an RKHS. An RKHS is defined as a Hilbert space $\mathcal{H}$ in which the evaluation functional

$$L_x : f \to f(x)$$

is continuous. In Hilbert spaces, continuous functionals are important due to the Riesz Representation Theorem, which states that for any continuous operator L and $f \in \mathcal{H}$, there exists a unique $g \in \mathcal{H}$ such that:

$$Lf = \langle f, g \rangle$$

Since the evaluation functional is defined to be continuous on an RKHS, we have that $f(x)$ can be expressed in the following form:

$$f(x) = \langle f, k_x \rangle$$

where $k_x \in H$. Since $k_x \in H$, we can apply the Riesz Representation Theorem on k_x itself, which implies:

$$k(x,y) = k_x(y) = \langle k_x, k_y \rangle$$

The fact that we can express any $f \in \mathcal{F}$ as an inner product makes finding the supremum a tractable task. In fact, the MMD^2 can be expressed as:

$$\begin{aligned}
MMD^2 &= \left[\sup_{f \in \mathcal{F}} (\mathbb{E}f(X) - \mathbb{E}f(Y))\right]^2 \\
&= \sup_{f \in \mathcal{F}} \left(\langle u_F, f \rangle - \langle u_G, f \rangle\right)^2 \\
&= \sup_{f \in \mathcal{F}} \langle u_p - u_q, f \rangle^2 \\
&= ||u_p - u_q||^2_{\mathcal{H}} \\
&= \langle u_p, u_p \rangle - 2\langle u_p, u_q \rangle + \langle u_p, u_q \rangle \\
&= \mathbb{E}_{x,x'} k(x,x') - 2\mathbb{E}_{x,y} k(x,y) + \mathbb{E}_{y,y'} k(y,y')
\end{aligned}$$

where we assumed that the expectation is a bounded operator in the RKHS, which implies that it is continuous. Therefore, μ_p and μ_q represent the unique functions in the Hilbert Space $\mathcal{H}$ that correspond to the Riesz Representation of $\mathbb{E}f(x)$ and $\mathbb{E}f(y)$, respectively. We also used the fact that a inner product is maximized when both of the inputs are in the same direction.

The authors propose several different methods for testing using the MMD, each with different computational costs and statistical properties. One test that Gretton et al. propose is based on U-statistics to construct the estimator MMD_n^2 for MMD^2. They use concentration inequalities to establish an inequality of the form

$$\mathbb{P}(|MMD_n^2 - MMD^2)| > \epsilon) \leq \alpha$$

and solve for ϵ.

We also consider the Energy distance (Szekely and Rizzo 2004). The Energy distance is defined as:

$$D(P, Q) = 2\mathbb{E}||X - Y|| - \mathbb{E}||X - X'|| - \mathbb{E}||Y - Y'||$$

where X is a sample drawn from P, Y is a sample drawn from Q, and X' and Y' are i.i.d. copies of X and Y, respectively. The Energy Distance can be thought of as a generalization of L_2 distance for comparing distributions. For cdfs F and G and $d = 1$, it can be shown that:

$$2\int_{-\infty}^{\infty} (F(u)) - G(u))^2 \ du = 2\mathbb{E}||X - Y|| - \mathbb{E}||X - X'|| - \mathbb{E}||Y - Y'||$$

Similar to the Kernel MMD, it can be established that $D(P, Q) = 0 \iff P = Q$; therefore a null hypothesis of $D(P, Q) = 0$ is equivalent to a null hypothesis of $P = Q$. In fact, the Energy Distance has been shown to be a class of kernel MMD. Following Sejdinovic et al. (2013), we will provide a sketch of the proof below.

Let $\mathcal{Z}$ be a topological space , $\rho : \mathcal{Z} \to [0, \infty)$ be a function that satisfies $\rho(x, y) = 0 \iff x = y$ and $\rho(x, y) = \rho(y, x)$. Then ρ is a semimetric on $\mathcal{Z}$.

We will also introduce the notion of a negative definite function. For a function $f : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ if for all $n \in \mathbb{N}$, any $x_i, ..., x_n \in \mathcal{Z}$, and any $\{\alpha_1, ..., \alpha_n\} \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 0$ we have:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j f(x_i, x_j) \leq 0$$

then a function is said to be negative definite. Euclidan distances are known to be a negative definite semimetric. Therefore, if we now denote a negative definite semimetric with ρ, the Energy Distance can be written as an expectation of ρ. In fact, the Energy distance has been generalized to allow for different negative semidefinite semimetrics. It has also been established that there exists a positive semidefinite kernel k such that a nonnegative definite semimetric ρ can be expressed as:

$$\rho(x, y) = k(x, x) - 2k(x, y) + k(y, y)$$

In our proof, we will also use an alternative representation of the Energy Distance. The Energy distance can be written as:

$$\begin{aligned} D(P, Q) &= \int\int 2\rho(x, y) d(P \times Q) - \int\int \rho(x, y) d(P \times P) - \int\int \rho(x, y) d(Q \times Q) \\ &= -\int\int \rho(x, y) \ d(P - Q) \times (P - Q) \end{aligned}$$

where we used the fact that $\rho(x, y)$ is symmetric in its arguments.

Plugging in our expression for ρ into the Energy Distance and letting $\nu = P - Q$, we have that:

$$\begin{aligned} D(P,Q) &= -\int\int \left[k(x,x) + k(y,y) - 2k(x,y)\right]\ d\nu(x)\ d\nu(y) \\ &= \int\int 2k(x,y)\ d\nu(x)\ d\nu(y) \\ &= 2MMD^2(P,Q) \end{aligned}$$

where we used an alternative representation of the MMD:

$$\begin{aligned} MMD^2 &= \mathbb{E}k(x,x') - 2\mathbb{E}k(x,y) + \mathbb{E}k(y,y) \\ &= \int\int k(x,y) d\nu(x)\ d\nu(y) \end{aligned}$$

and the fact that the signed measure ν has measure 0 on the whole space $\mathcal{Z}$.

The Energy Distance has the property that it is rotationally invariant, making it a useful distance for nonparametric high-dimensional tests. A hypothesis test is derived using a permutation test framework.

3.6 Local Comparisons

Even if two-sample tests that test whether two distributions are equal are inconclusive, it may be the case that the distributions are significantly different locally. In the following section, we will describe how density ratios can be used to analyze local differences between distributions.

3.6.1 Density Ratios

Suppose $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_m \sim Q$ and we are interested in whether $X|X \in A \overset{\mathrm{d}}{=} Y|Y \in A$ for some interval $A \in \mathbb{R}^d$. The *density ratio* is defined as:

$$\beta(u) = \frac{f_y(u)}{f_x(u)}$$

We can look at the density ratio and determine whether it is significantly different from 1 within our region A. In particular, we can consider the mean density ratio over a region:

$$E(\beta(u)|u \in A)$$

However, we do not know the set A a priori. Methods to find a set A that do not affect the calculation of confidence intervals will be discussed further in a later section.

3.6.2 Estimation of Density Ratio

When estimating a density ratio, a common procedure in the literature is to construct the estimators $\hat{f}_x(x)$ and $\hat{f}_y(y)$ using nonparametric density estimation methods and then form the estimator:

$$\beta(u) = \frac{\hat{f}_y(u)}{\hat{f}_x(u)}$$

However, this method has two important shortcomings. The denominator need not be positive definite, which may lead to $\hat{\beta}(z) = \infty$ for some values of z. In addition, division of two estimators may compound the error of the estimator. Therefore, it seems likely that estimators that do not rely on estimating both the numerator and denominator separately would seem ideal. In fact, Izbicki, Lee, and Schafer (2014) provides empirical evidence that one step methods such as k-NN density ratio estimation perform better than the two step methods for problems related to the estimation of photometric redshift

3.6.3 k-NN Density Ratio Estimation

The density ratio can be estimated using k nearest neighbors as follows (Zhao and Liu 1985):

$$\hat{\beta}(u) = \frac{n}{m} \quad \frac{\mathbb{1}(y \in V_x^k)}{k}$$

where $u \in \mathbb{R}^d$, m,n are sample sizes from P, Q respectively, $d(\dot{})$ represents the distance function, and

$$V_x^k = \{x \in \mathbb{R}^d : d(x,u) \leq d_{(k)}(y,u)\}$$

Intuitively, we can think of the k-NN estimator as a ratio of the number of data points in a hypersphere of size depending on magnitude of $f_y(y)$.

We choose k is chosen by using 5-fold Cross-Validation. The loss function used is:

$$L(\beta, \hat{\beta}) = \int (\beta(u) - \hat{\beta}(u))^2 dQ$$

$$\hat{L}(\beta, \hat{\beta}) = \frac{1}{m}\sum_{i=1}^{m} \hat{\beta}^2(y) - \frac{2}{n}\sum_{i=1}^{n} \hat{\beta}(x)$$

For each pairwise comparison, G is chosen to be the band with the most observations. We scale the dataset to avoid implicit variable importance during computation of nearest neighbors and use standard Euclidean distance.

3.6.4 Adaptive Grid

In our analysis, the density ratio was evaluated at data points instead of over a d-dimensional grid. We provide further justification for this procedure later in this section. To visualize the behavior of the density ratio in an interval, we consider two-dimensional feature spaces. We will define the regions A_i as elements of a partition of $\mathbb{R}^2$. In our application, we want to be able to make inferences in each grid. If the value of both densities is small within a region, then we may not have enough power to detect a difference in that grid.

One solution is to pool nearby regions of low density together. This can be done via an adaptive grid where each grid contains roughly the same number of points. A simple way is to ensure that the data is partitioned into groups of equal sizes along each dimension. For visualization purposes, a minimum grid size will also be set. After some experimentation, a value of 1/16 of the difference between the smallest and largest observation for a particular dimension was used. This method does not guarantee equal number of points in each grid, but it does along each dimension if the minimum grid length were 0. We can then take the mean of the density ratio in this region and that should give us an approximation of the conditional mean density ratio via the plug-in principle:

$$E(\beta(U)|u \in A) \approx \sum_{u \in A} \frac{1}{\mathbb{1}(u \in A)} \beta(u) \approx \sum_{u \in A} \frac{1}{\mathbb{1}(u \in A)} \hat{\beta}(u)$$

To make this more precise, u is randomly drawn from the mixture model given by:

$$\begin{aligned} f_u(u) &= \frac{\int_A f_x(x)dx}{\int_A f_x(x)dx + \int_A f_y(y)dy} f_x(u) + \frac{\int_A f_y(x)dx}{\int_A f_x(x)dx + \int_A f_y(y)dy} f_y(u) \\ &\approx \frac{n}{m+n} f_x(u) + \frac{m}{m+n} f_y(u) \end{aligned}$$

where $f_x(u)$ and $f_y(u)$ are also conditioned on $u \in A$. Our conditional expectation of the log density ratio can be written as:

$$\begin{aligned} E(\log \beta(u)|u \in A) &\approx \int \log\left(\frac{f_y(u)}{f_x(u)}\right) \frac{n}{m+n} f_x(u) + \frac{m}{m+n} f_y(u) du \\ &= \frac{m}{m+n} KL(Q, P|u \in A) - \frac{n}{m+n} KL(P, Q|u \in A) \end{aligned}$$

We can interpret this expression as follows. A naive comparison of whether two samples are different would involve counting the number of observations from each sample. Here, this number is smoothed by the KL divergence. However, this is assuming that the sample sizes are equal, which is not true for any comparison involving the J Band. We may need to take a weighted average instead of a naive mean, where each point is weighted according its sample size. As is, however, averaging reduces the variance of our estimator over a contiguous region and also reduces the multiple comparison problem.

3.6.5 Clustering

One problem with the grid approach is that it will not generalize very well to higher dimensions due to curse of dimensionality. In addition, as we increase dimension, it loses its utility as a visualization tool. To continue using the plug-in approximation for the mean, we either need to consider data splitting or use a two step method in which the way the partition is defined is independent from the estimation of the density ratio so that the confidence intervals are valid.

One approach would be to use clustering after removing the labels. Removing the labels would make the two-step inference valid without the need for data splitting. Galaxies naturally have therefore our distribution can be thought of as some kind of mixture model. If the mixture components have different probabilities, then this method should intuitively be sensitive to those differences. We will use hierarchical clustering and use statistics such as the Dunns validity index or Davies-Bouldin validity index to find a reasonable place to cut the dendogram.

3.6.6 Inference for Density Ratios

We are interested in construction a confidence interval for the quantity:

$$\sum_{u \in A} \frac{1}{\mathbb{1}(u \in A)} \hat{\beta}(u)$$

To do this, we will use the bootstrap.We will construct 500 bootstrap samples of the original data. For each bootstrap sample we follow the procedure below:

1. Draw a bootstrap sample $u_1^{(b)}, u_{n+m}^{(b)}$ from the joint distribution formed by combining the two bands of interest.
2. Recompute the statistic $\hat{\beta}^{(b)}$ on $u_1^{(b)}, u_{n+m}^{(b)}$ at the resampled points $u_1^{(b)}, u_{n+m}^{(b)}$.
3. Pool bootstrap sample points that lie in the same region, calculate the mean density ratio for that region.
4. Repeat for $b = 1, ..., 500$
5. Compute the α and $1 - \alpha$ quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of $\hat{\beta}^{(b)}$ and form the confidence interval:

$$[L, U] = [2\hat{\beta} - q_{1-\alpha/2}, 2\hat{\beta} - q_{\alpha/2}]$$

4 Discussion

4.1 Global Comparisons

4.1.1 Univariate Distributions

The Score test for testing the difference of binomial menas for M and I are significantly different from zero for the H-J Band and H-Y Band comparisons. These results are summarized in the table below:

	HJ	HY	JY
$M > 0$	68.694	45.794	18.735
$I > 0.01$	46.140	235.430	56.795

Table 1: Score Test Results, where a statistic of 1.96 would indicate significance at α level equal to 0.05.

The Mann-Whitney U tests also reject null hypothesis for the H-J comparison for the M and the I statistics and for the D statistic for the $H - Y$ and $J - Y$ comparison.

Overall, univariate results are consistent with the findings in Freeman et al (2013): that is the M, I, and D statistics seem to be variables that are strong indicators of disturbance. The other statistics have been found to also have predictive power for galaxy classification. The nonparametric univariae tests may not have enough power to detect these differences.

4.1.2 Mann-Whitney U Test Results

	HJ	HY	YJ
M	0.014^*	0^*	0.508
I	0.001^*	0.439	0.518
D	0.09	0.018^*	0^*
Gini	0.155	0.001^*	0.252
M_{20}	0.094	0.444	0.605
C	0.362	0.281	0.56
A	0.493	0.138	0.435

Table 2: Mann-Whitney Test computed to compare H and J Bands (HJ), H and Y Bands (HY), and Y and J Bands (YJ). Asterisks represent p-values < 0.05.

4.1.3 Bivariate Distributions

The confidence intervals for correlations are overlapping between bands; therefore we cannot reject the null hypothesis that the correlations are the same.It is likely that we are comparing mixture distributions where the mixing proportions are different across bands. A test using the correlation will likely not be a powerful test .

4.1.4 Multivariate Distributions

The Kernel MMD with any reasonable choice for the kernel fails to reject null hypothesis for any of the pairwise comparisons between bands. However, the Energy Test rejected the null hypothesis for the H-J comparison (p-value: 0.008) and H-Y comparison (p-value: 0.005). The Energy Test

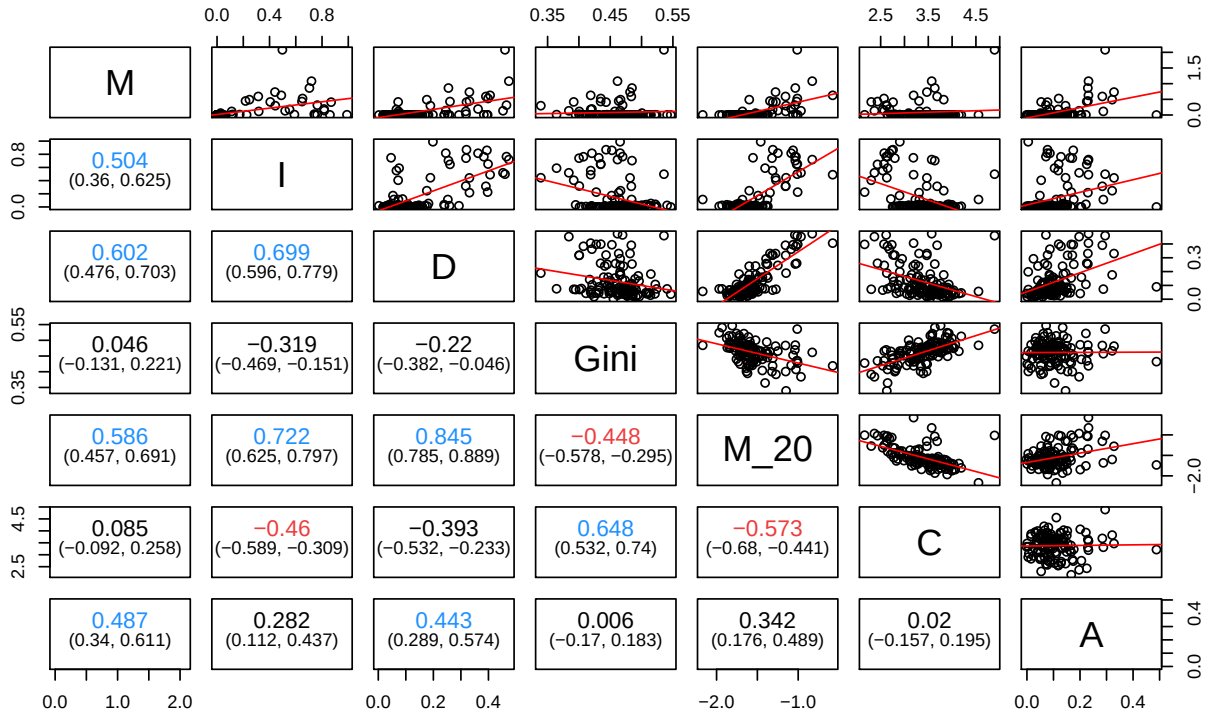

Figure 4: Correlation Plot for H-Band, with correlations values and 95% confidence intervals on the lower diagonal Correlations above 0.4 are in blue and those below -0.4 are in red.

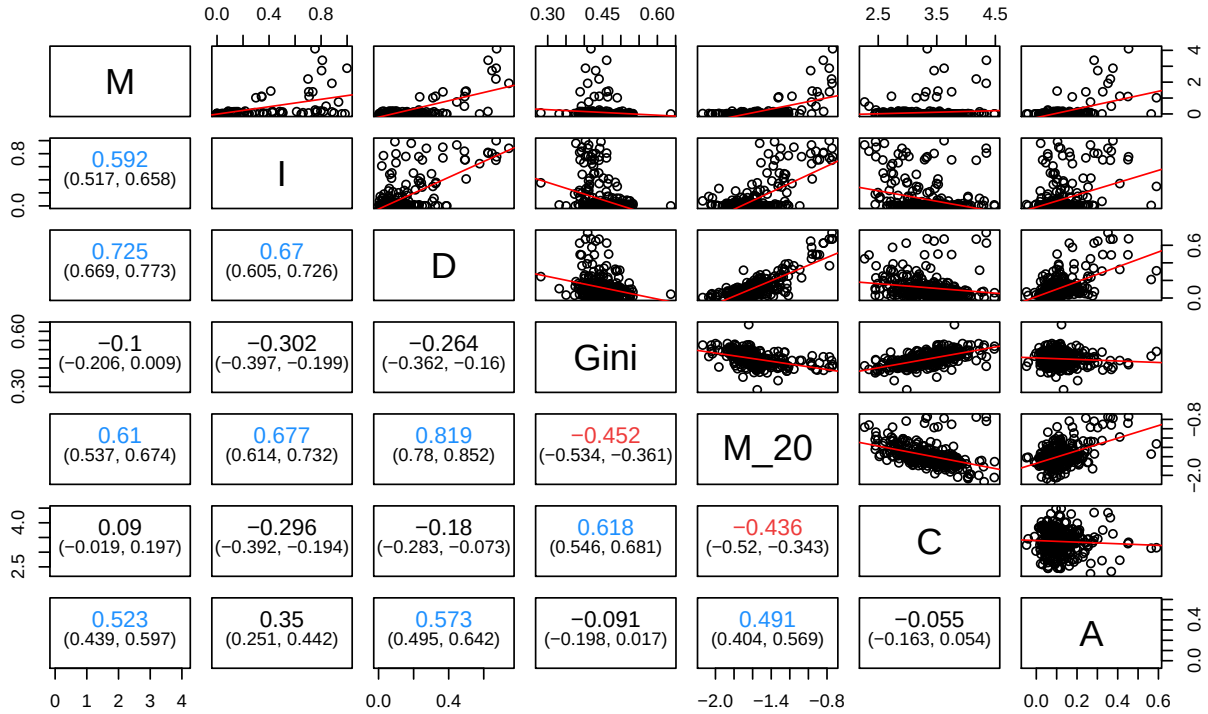

Figure 5: Same as figure 4, except for J-Band galaxies

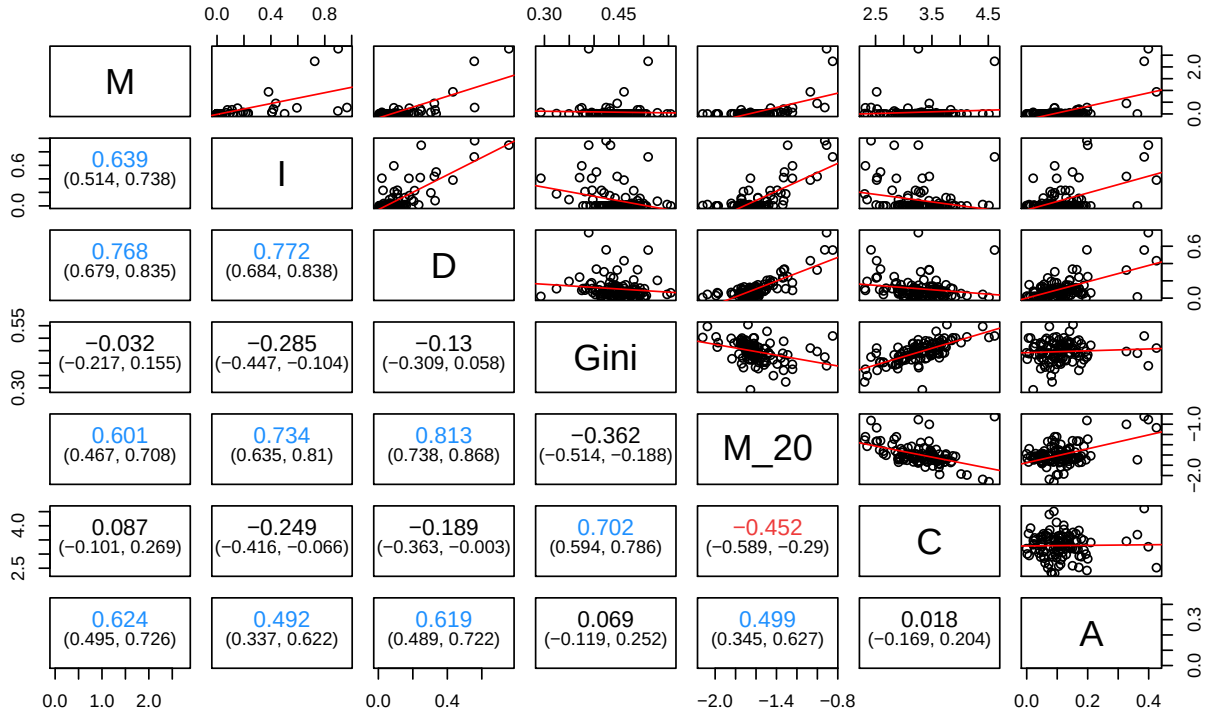

Figure 6: Same as figure 4, except for Y-Band galaxies

failed to reject the null hypothesis for the J-Y comparison (p-value: 0.527). Therefore, overall the multivariate high dimensional tests are inconclusive. One possible takeaway is that kernel choice may make a very large difference on the test result. One may also consider using an implementation of the Kernel MMD procedure with higher power at the expense of higher computational cost.

4.2 Local Differences

4.2.1 Density Ratios - Adaptive Grid

Instead of producing plots for all 63 pairwise comparisons of statistics between two bands, we will quickly summarize some trends and display several representative plots.

Statistics that showed significant differences even with the univariate tests have a density ratio greater than 1 in a region that we would expect to be correlated with the presence of a merger. For comparisons between the H and J and H and Y Bands, we also see that the upper right region of Gini and M_{20} is significantly greater than 0, which would also be consistent with an increasing merger rate. In previous studies, all statistics used have been shown to be at least weak predictors of mergers. The grid approach may reveal some of these differences more than the univariate approaches considered previously.

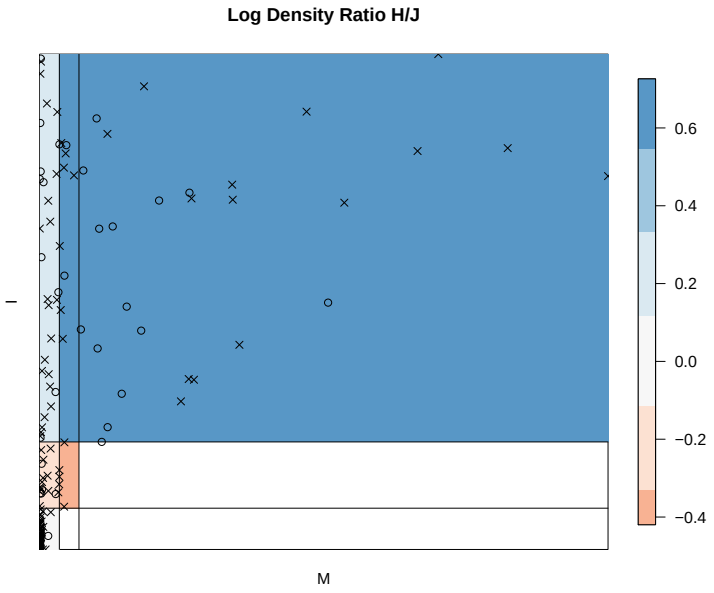

Figure 7: Grid for HJ M and I statistics. Blue represents regions with log density ratio greater than 0 and red represents regions with log density ratio less than 0. The x's indicate galaxies from the the H Band and the points indicate galaxies from the J-Band.

The HJ M-I plot is consistent with what we found using univariate tests; that is high M and and I, both of which are associated with the presence of a merger, is more likely in the H Band.

However, we see more disproportionately more galaxies from the J Band than from the H Band for very large values of M. Therefore, it is possible that very high M and I values may be uninformative; that is they may be galaxies that are outliers but not necessarily disturbed morphologies. Manual examination of a small sample of galaxies from this region of very large M and I indeed reveals some galaxies where part of the galaxy image has been cut off.

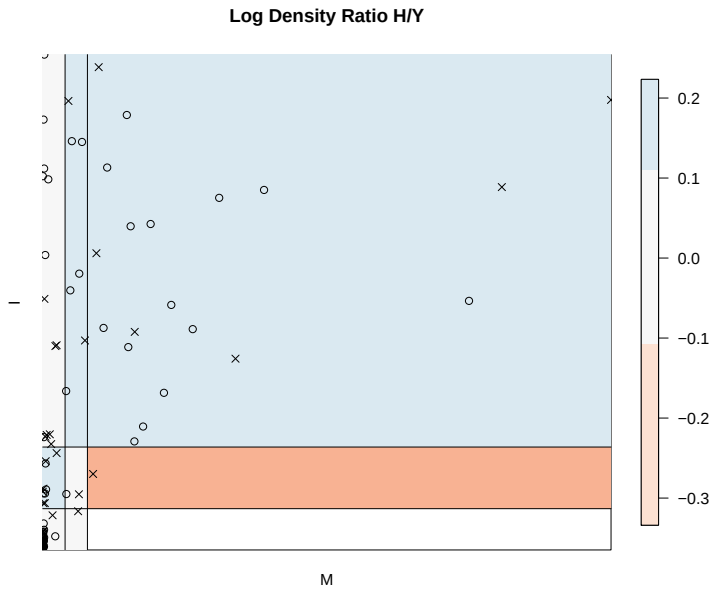

Figure 8: Same as Figure 7 for HY M and I statistics.

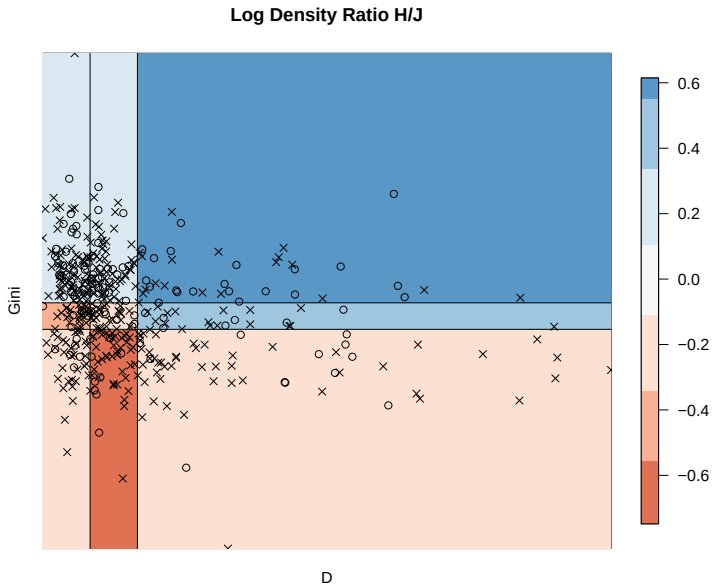

Figure 9: Same as Figure 7 for HJ Gini and and D statistics.

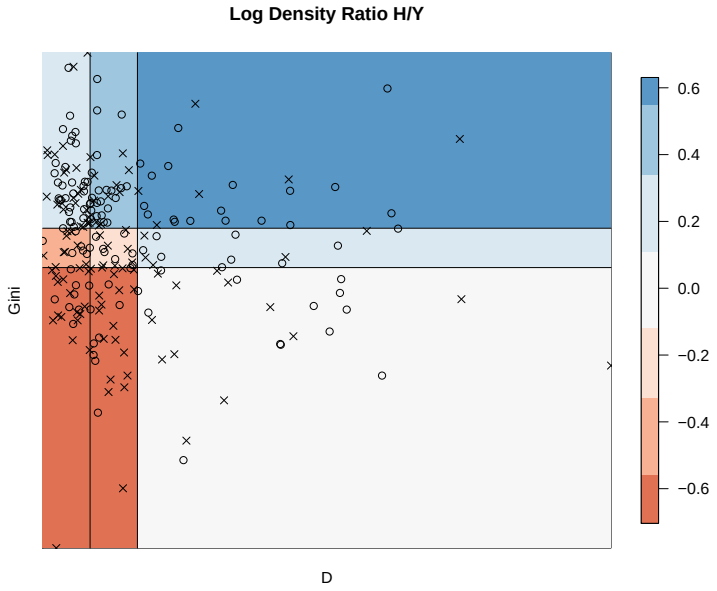

Figure 10: Same as Figure 7 for HJ Gini and D statistics.

Figures 11 and 12 display plots of galaxies from the M and I HJ comparison. We confirm that a random sample of galaxies from the upper right portion of the plot (region of interest) contains more disturbed morphologies than galaxies sampled randomly from outside this area. The HY MI plot also is consistent, but the density ratio is not as large as it is for the HJ comparison. This is most likely due to small sample size.

Part of the problem may be that the space was not partitioned as efficiently as the univariate case. In the univariate case, we use our knowledge about the M and I statistic to partition the space whereas in the grid method we enforce a more even bin size. However, the HJ and HY D-Gini plot both reveal differences that we did not detect with the nonparametric univariate tests.

In Table 3, we provide confidence intervals for the mean log density ratio all upper right regions examined in the bivariate plots. We see that they are all signficantly greater than 0, although we must be cautious due to concerns about the validity of the bootstrap and the small sample size, particularly for the HY M-I comparison.

Region	[a	b]
HJ (M,I)	0.255	1.634
HY (M,I)	0.129	1.549
HJ (D,Gini)	0.370	1.716
HY (D,Gini)	0.295	1.375

Table 3: 95 % confidence intervals for the log density ratio of upper right regions displayed in Figures 7-10.

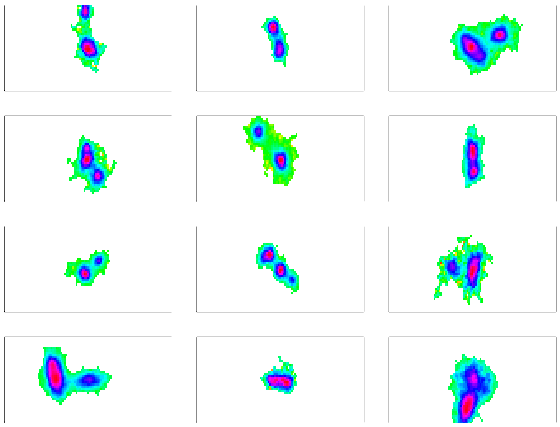

Figure 11: Galaxies From Region of Interest for M and I

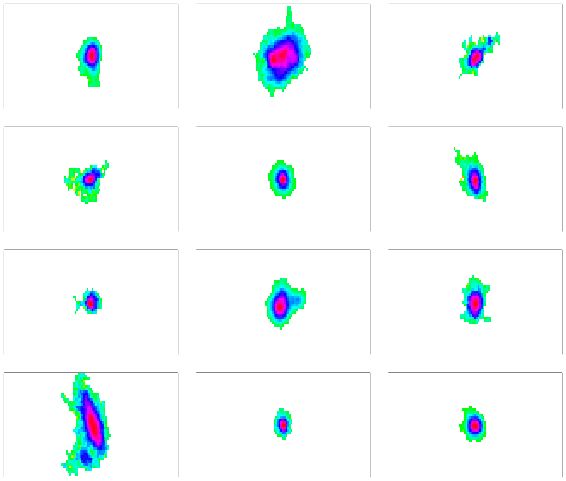

Figure 12: Galaxies From Outside of Interest for M and I

4.2.2 Clustering

The Dunn-Davies index and Bouldin index suggest cutoffs of 2 and 5 respectively for average hierachical clustering. Below we will present the cluster means and the confidence intervals for the mean density ratio in a given region.

We see that we have one cluster with high I, D, and M_{20} with a significantly different average density ratio from zero for the HY Band, which may point to the existence of region where distributions differ. A better method would potentially also detect difference with HJ comparison. In particular, as mentioned before, we may need to perform a weighted mean or over/under-sampling of the data to deal with the imbalanced dataset. Also, confidence intervals may understate uncertainty if there are issues concerning the validity of the bootstrap, so we must interpret results cautiously.

Cluster	size	M	I	D	M_{20}	Gini	C	A
A	48	0.250	2.028	0.956	-1.084	1.378	-1.550	0.586
B	389	-0.196	-0.340	-0.236	0.141	-0.271	0.158	-0.152
C	9	3.566	2.460	3.111	-0.009	2.967	1.092	2.131
D	4	8.070	3.183	4.431	-0.663	3.156	0.759	2.946

Table 4: HJ Mean of Hierarchical Cluster

Cluster	[a	b]
A	-0.328	1.006
B	-0.254	0.075
C	-0.099	1.457
D	-1.191	2.317

Table 5: HJ 95 % Confidence Interval for Log Density Ratio

Cluster	size	M	I	D	M_{20}	Gini	C	A
A	21	0.023	0.644	0.369	0.440	-1.076	2.843	0.195
B	132	0	0.013	0.073	0.476	-1.692	3.582	0.105
C	79	0.002	0.105	0.110	0.420	-1.530	3.036	0.087
D	3	0.142	0.708	0.591	0.4780	-0.924	4.254	0.360

Table 6: HY Mean of Hierarchical Clusters

Cluster	[a	b]
A	0.338	1.834
B	-0.103	0.418
C	-0.512	0.258
D	-0.384	2.413

Table 7: HY 95 % Confidence Interval for Log Density Ratio

5 Conclusion

We observe that the frequency of disturbed morphologies appear to increase with redshift, which is consistent with previous results. Univariate tests with the M and I statistics reveal differences

between the H and J and H and Y Band. Nonparametric (Whitney-U) tests also indicate that the D statistic informative; however, the other tests are not powerful enough to detect differences between distributions.

The multivariate tests, on the other hand, were largely inconclusive; in particular, the Kernel MMD with Gaussian Kernel and Energy Test have contradictory results. Using local analyses, however, we were able to observe differences between distributions that were inaccessible with the global methods. In fact, local analyses show that even statistics that are not strongly informative by themselves contain information.

References

[1] Behroozi, P.S. et al (2010). *A comprehensive analysis of uncertainties affecting the stellar mass-halo mass relation for* $0 < z < 4$. Astrophysical Journal 717 (1), 379.

[2] Conselice, C.J. (2003). *The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories.* Astrophysical Journal 147 1.

[3] Dahlen, T. et al (2013). *A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation.* Astrophysical Journal, 775.

[4] Freeman, P.E. et al (2013). *New Image Statistics for Detecting Disturbed Galaxy Morphologies at High Redshift* Monthly Notices of the Royal Astronomical Society, v. 434, pp. 282-295 (2013).

[5] Guo, Y. et. al (2013). *CANDELS Multi-wavelength Catalogs: Source Detection and Photometry in the GOODS-South Field.* Astrophysical Journal, Letters 24.

[6] Gretton, A. et al. (2007). *A kernel method for the two-sample problem.* NIPS 513 520. MIT Press, Cambridge, MA.

[7] Grogin, N.A. et al. (2011). *CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey.* Astrophysical Journal, 197 35.

[8] Koekemoer, A.M. et al. (2011). *CANDELS: The Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey - The Hubble Space Telescope Observations, Imaging Data Products and Mosaics* in Astrophysical Journal, 197 2.

[9] Izbicki, R., Lee, A.B., and Schafer, C.M. (2014). *High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation.* Journal of Machine Learning Research (AISTATS Track), 33, 420-429.

[10] Lotz, J.M., Primack, J., and Madau, P. (2004). *A New Non-Parametric Approach to Galaxy Morphological Classification.* The Astronomical Journal, 128:163182.

[11] Mobasher, B. et al., in preparation.

[12] Sejdinovic, D. et al (2013). *Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing.* Annals of Statistics, 41 No.5 2263-2291.

[13] Szekely, G. J. and Rizzo, M (2004). *Testing for Equal Distributions in High Dimension.* InterStat, Nov. (5).

[14] Zhao, L. and Liu, Z. (1985). *Strong consistency of the kernel estimators of conditional density function.* Acta Mathematica Sinica, 1(4):314318.