Comparing Distributions of Galaxy Morphologies

Ilmun Kim

Adviors : Ann Lee[1], Peter Freeman[1] and Jeffrey Newman[2]

1 Introduction

A principal goal of astronomy is to describe and understand how galaxies evolve as the Universe ages. To understand the processes that drive evolution, one needs to investigate the connections between various properties of galaxies−such as mass, star-formation rate (SFR), and morphology−in a quantitative manner. The last of the these properties, morphology, refers to the two-dimensional appearance of a galaxy projected onto the plane of the sky.

The classic description of galaxy morphology is the *Hubble sequence* (Figure 1), in which galaxies exist in two main classes: ellipticals and spirals. While this discrete measurement of galaxy morphology is efficient to describe nearby galaxies, it is not flexible enough to capture the full range of morphologies that galaxies have exhibited over the Universe's history. To overcome this limitation, various authors have developed so-called "nonparametric" summary statistics that attempt to concisely preserve morphological information (e.g., Conselice 2003, Lotz et al. 2004, Freeman et al. 2013). By combining summary statistics with catalogs that arise from large-scale sky surveys, it is now possible to carry out comprehensive quantitative studies about galaxy formation and evolution.

One of the major questions we would like to address in this study is how galaxy assembly occurs over cosmic time. However, we are limited by not being able to trace galaxies through time; rather, we have many snapshots gleaned at different times that we wish to stitch together into a panorama. Within each snapshot, it is particularly interesting to investigate relationships of galaxy morphologies to other physical properties of galaxies, such as stellar mass and star-formation rate (SFR). Another interesting question is whether the existence of such relations, or their nature, breaks down or changes at some point. In this study, we focus on the first topic from the standpoint of comparing distributions of galaxies (e.g., high- versus low-mass galaxies or high- versus low-SFR galaxies) at the same time point; the

[1]Department of Statistics, Carnegie Mellon University
[2]Department of Physics and Astronomy, University of Pittsburgh

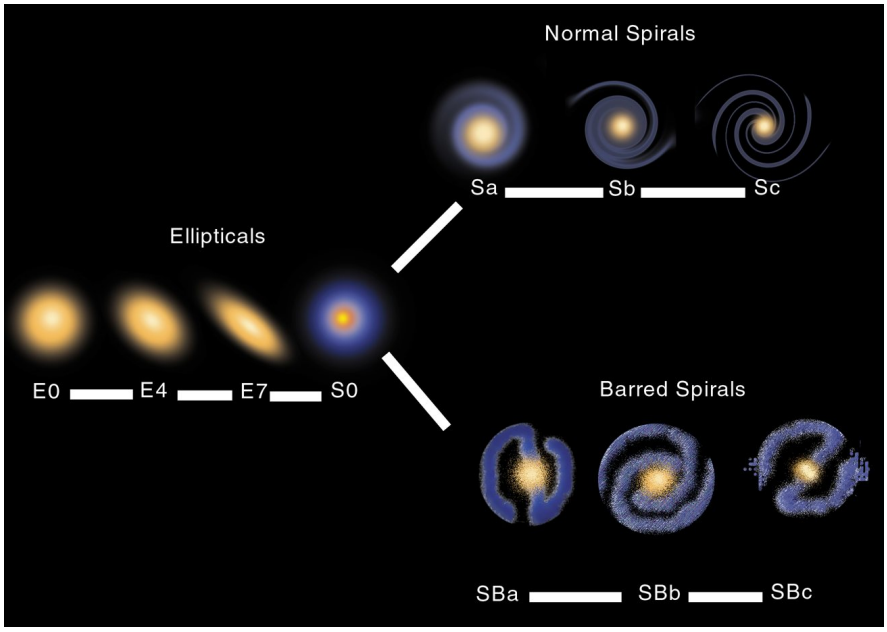


Figure 1: Hubble tuning fork diagram, showing the three major morphological classes for massive galaxies in the local Universe: ellipticals, spirals, and barred spirals.

methods we develop can then later be applied to matched samples at different redshifts to test for evolution.

From a statistical point of view, the challenge is to develop tools that are able to quantify the detailed structural differences between two different groups of galaxies. Most two-sample tests only tell us whether two samples are different, rather than exactly where and by how much they differ. At the same time, in astronomy, standard approaches to comparing galaxy morphologies mostly rely on by-eye comparisons of one-dimensional or two-dimensional plots of feature statistics; such approaches do not identify statistically significant differences and also do not consider more than two features jointly.

In this work, we develop and apply a new approach to comparing distributions of galaxy morphologies. The method is fully nonparametric and finds locally significant differences with statistical confidence in a multivariate feature space. The basic idea is to detect locally significant differences with a test statistic based on class posteriors. The details on the methodology can be found in Section 3. We apply our method on data collected by the CANDELS program, and use it to explore interrelationships between galaxy morphologies, stellar mass and SFR at a fixed cosmic time.

The rest of this work is organized as follows. In Section 2, we present a brief description of the data, which include summary statistics for measuring the morphologies and structures

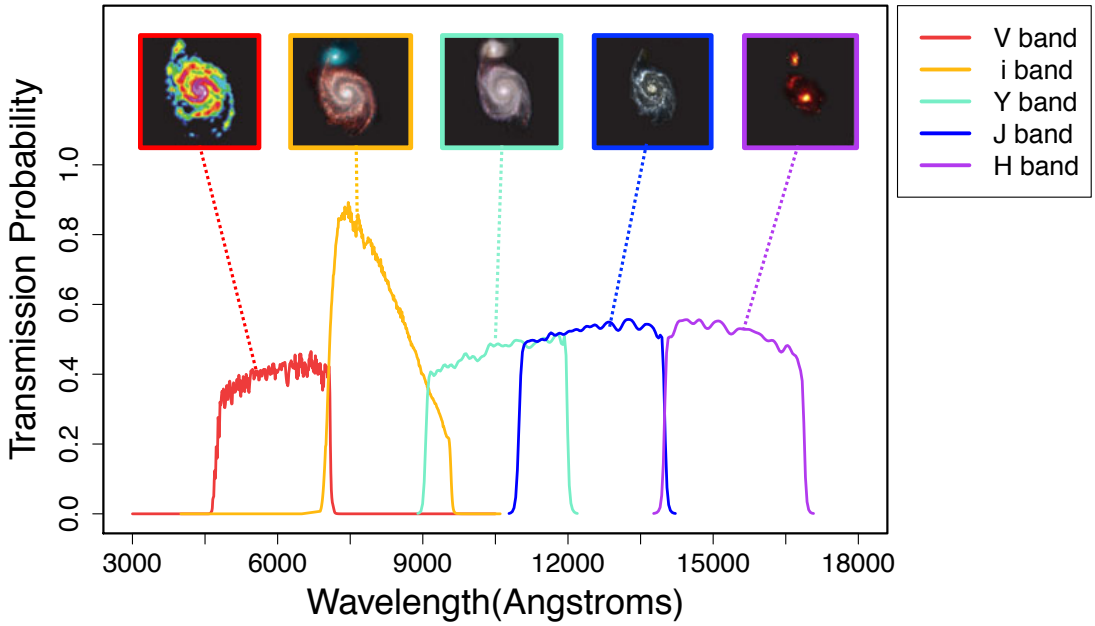


Figure 2: Filter curves for the V, i, Y, J, and H bands, as observed by the ACS and WFC3 instruments on board the Hubble Space Telescope. The transmission probabilities are denoted as a function of wavelength in units of Å, where 1 Å= 10^{-10} meters. (The human eye observes light with wavelengths between ≈ 4,000 and 7,000 Å.) Five images of one galaxy are include to illustrate how morphology changes as a function of wavelength.

of galaxies. In Section 3, we propose a framework and methodology for detecting local significant differences. Then, in Section 4, we apply our method to describing how galaxy morphologies and other physical properties are interrelated . Finally, in Section 5, we provide a summary and future directions.

2 Data

We analyze galaxy morphologies from galaxy images. The images were collected from four fields oberved by the *Hubble Space Telescope* (*HST*) as part of the CANDELS program: the Cosmic Evolution Survey field (COSMOS), the Extended Groth Strip (EGS), the Great Observatories Origins Deep Survey North field (GOODSN), and the Ultra Deep Survey field (UDS). Each galaxy is observed at up to five wavelengths via different filters. Specifically, we have galaxy images from five filters−V, i, Y, J, H−and the transmission probabilities as a function of wavelength for each are shown in Figure 2.

2.1 Redshift Estimation

In expanding space, the velocity at which a galaxy is moving away from us is proportional to its distance from us. This movement results in the observed wavelengths of galaxies shifting towards the red end of the spectrum with the increase in wavelength, a phenomenon referred to as redshift. Objects with higher values of redshift are more distant from us and it takes longer for the emitted light to reach our telescope. More specifically, redshift is defined as

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e} \tag{1}$$

where λ_o and λ_e represent observed and emitted wavelengths of a photon, respectively.

Two ways to estimate redshift are via spectroscopy and via photometry. Spectroscopy usually provides a precise estimate ($\Delta z/z \sim 10^{-5}$), but it is not appropriate for large–scale sky surveys because of time and cost. On the other hand, photometry is less time consuming but usually imprecise so it is common practice to estimate a probability distribution for the redshift of a galaxy from its photometric data (Izbicki et al. 2016 and references therein). Estimated density functions are usually unimodal but some exhibit several spikes or irregular shapes. Four examples of density functions are shown in Figure 3.

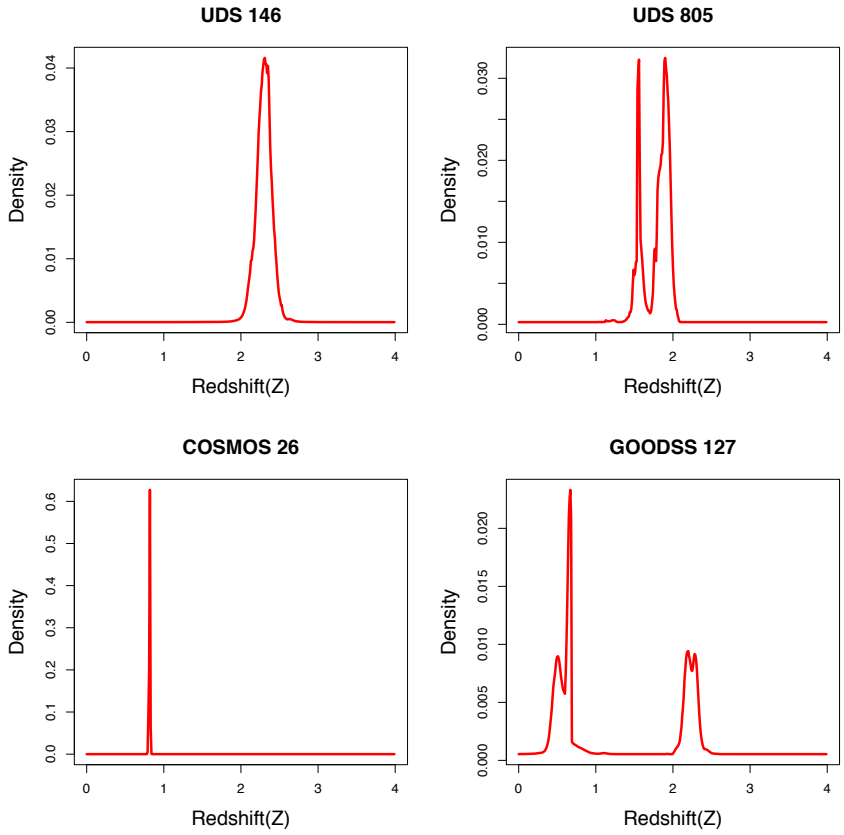


Figure 3: Illustration of redshift density estimates. Most estimates exhibit only one mode, such as those for UDS 146 and COSMOS 26 (upper and lower left), while some such as UDS 805 and GOODSS 127 exhibit more complex behaviors (upper and lower right).

2.2 Sample Selection

As mentioned before, there are five sets of images for each galaxy from five different filters. We use redshift tomography to select an appropriate image for each galaxy based on its estimated distance from us. (We also call this the discrete bin approach.) Specifically, after rearranging eq. (1), the observed wavelength can be expressed as

$$\lambda_o = (1 + z) \times \lambda_e.$$

According to Figure 2, we can determine a range for λ_o for each filter (see Table 1).

Table 1: Range of observed wavelength for five HST filters.

	V-filter	i-filter	Y-filter	J-filter	H-filter
Minimum λ_o	4643.04	7020.00	9226.99	11063.50	14027.50
Maximum λ_o	7167.95	9550.00	11876.99	13908.50	16710.50

One can then determine a redshift range given a fixed emitted wavelength:

$$\lambda_o \in [\lambda_{min},\ \lambda_{max}] \iff z \in \left[\frac{\lambda_{min} - \lambda_e}{\lambda_e},\ \frac{\lambda_{max} - \lambda_e}{\lambda_e}\right]$$

In this project, λ_e is chosen to be 4,500Å, a wavelength where the appearance is not greatly affected by star formation ($\lambda_e \ll$ 4,500Å) or dust ($\lambda_e \gg$ 4,500Å). The corresponding redshift ranges are presented in Table 2.

Table 2: Range of redshift for five HST filters given $\lambda_e = 4{,}500\mathring{A}$.

	V-filter	i-filter	Y-filter	J-filter	H-filter
Minimum z	0.03	0.56	1.05	1.46	2.12
Maximum z	0.59	1.12	1.64	2.09	2.71

Note that when a spectroscopic redshift is available, we assign a galaxy to the corresponding filter on the basis of that single precise estimate. However, most galaxies have a photometric redshift estimate, with a density function of the sort shown in Figure 3. In this case, we estimate the probability that a galaxy lies in one of the redshift ranges by integrating the density estimate. If the probability is greater than a certain threshold, for example 0.8 in our case, then the galaxy is associated to the corresponding filter. Otherwise, the galaxy is removed for the entire analysis. Furthermore, when a galaxy is included in two different filters due to overlapping areas in the redshift range, we select the one for which the probability

Table 3: Results of binning in the five CANDELS fields. Y-band data are not available for the COSMOS, UDS, and EGS fields, and i-band data is not available for the GOODSS field.

Fields	V	i	Y	J	H	Not matched	Total
GOODSS	158	NA	475	598	570	1,828	3,629
GOODSN	391	938	456	472	327	1,984	4,568
COSMOS	273	995	NA	921	431	2,095	4,715
UDS	145	612	NA	769	517	2,270	4,313
EGS	189	594	NA	761	346	3,823	5,713
Total	1,156	3,139	931	3,521	2,191	12,000	22,938

is higher than the other. In Figure 4, we illustrate the discrete bin approach using four examples.

2.3 Summary Statistics

As an quantitative approach to studying the morphological content of galaxies, it is common to extract important features of a galaxy image. We also adopt this approach by considering seven statistics that summarize galaxy images in a nonparametric way: Multimode (M), Intensity (I), Deviation (D), Gini coefficient (G), M_{20}, Concentration (C), and Asymmetry (A). Each statistic is sensitive to particular aspects of galaxy morphology. In brief, the M, I, D statistics proposed by Freeman et al. (2013) capture galaxies with disturbed morphologies. G and M_{20} in Lotz et al. (2004) are useful to describe the variance of a galaxy's stellar light distribution. C and A, given in Conselice (2003), measure the concentration of light and asymmetry of a galaxy, respectively. More details and mathematical definitions of summary statistics are provided in Appendix C.

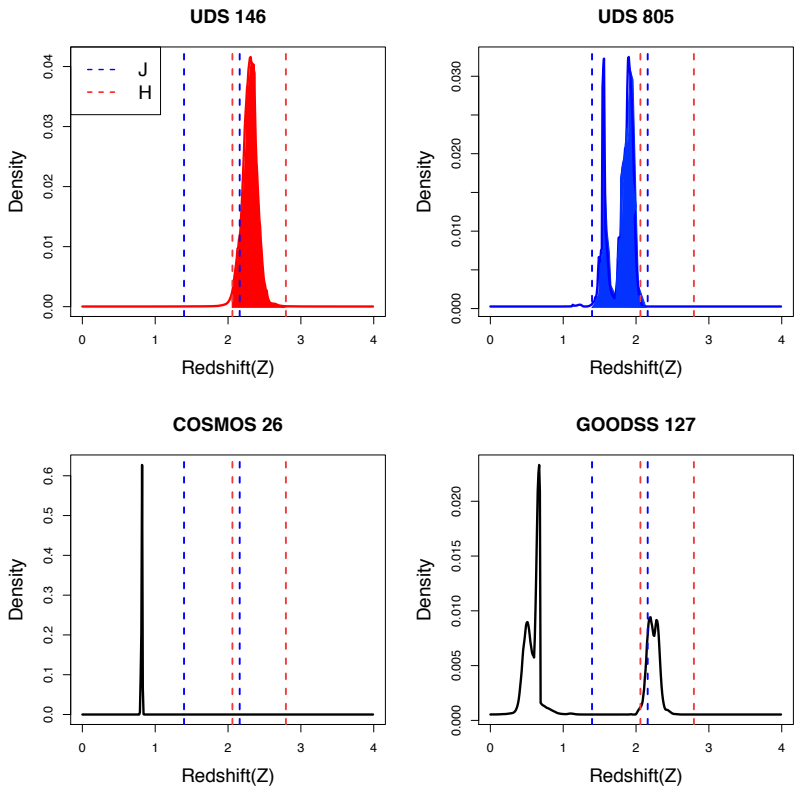


Figure 4: Illustration of the discrete-bin approach using the J- and H-band HST filters. This approach identifies UDS 146 as an "H-band galaxy" because its probability summed over the range of the H-band filter exceeds our adopted threshold of 0.8. Similarly, we assign UDS 805 to the J-band bin. On the other hand, the galaxies COSMOS 26 and GOODSS 127 are assigned to neither the J- nor H-band bins since neither has integrated area greater than 0.8 within either filter range.

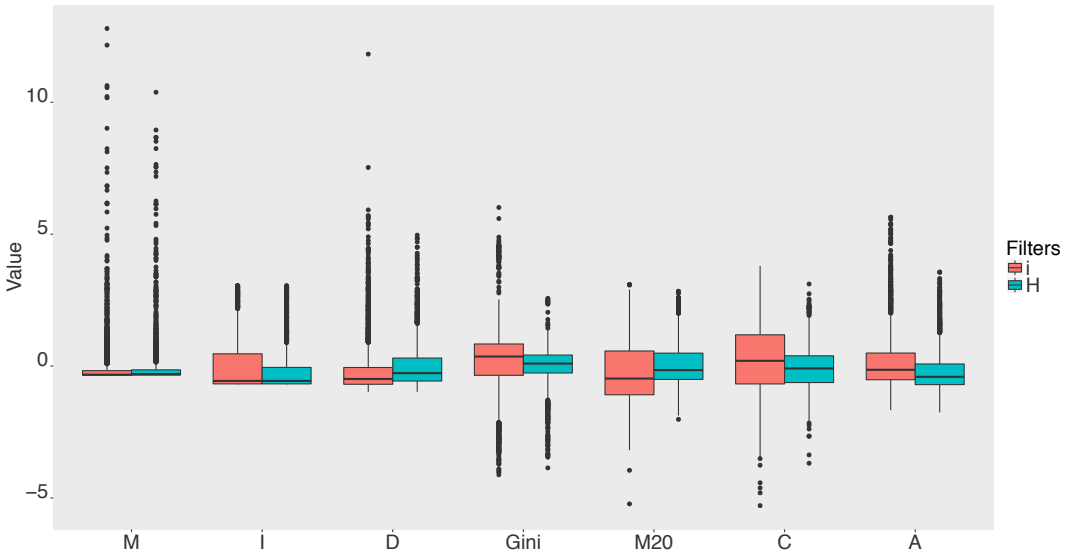


Figure 5: Boxplots of the seven morphology statistics for i- and H-band bin galaxies after standardization. The M, I, and D statistics have fairly skewed distributions, and every distribution exhibits clear outliers.

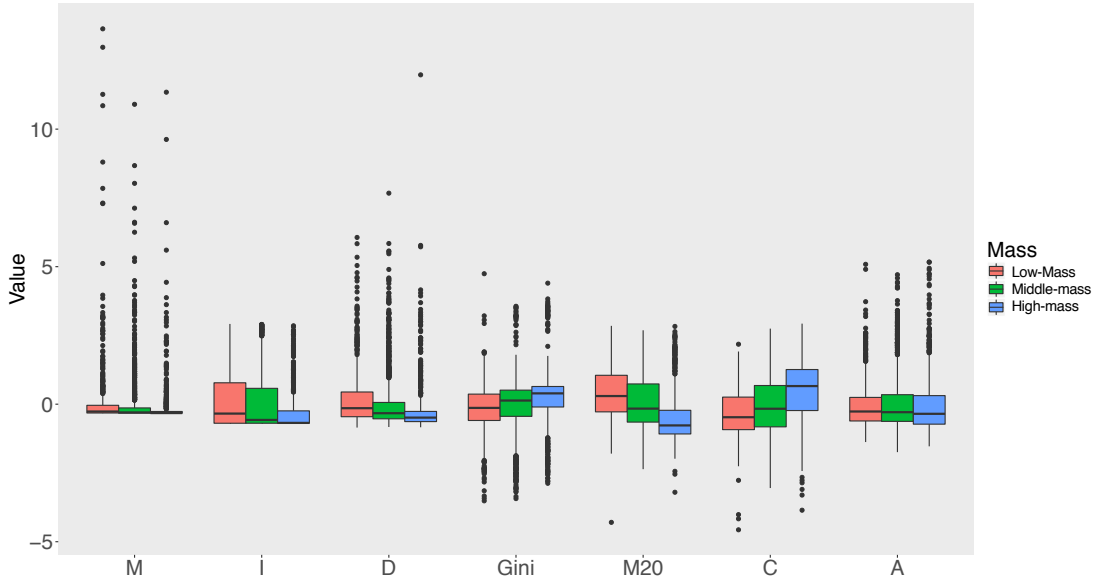


Figure 6: Boxplots of the seven morphology statistics from the different mass groups defined for i-band bin data, after standardization. The three groups are separated at the 25th and 75th mass percentiles (see Section 4.1.)

3 Methods

The objective of a *two-sample test* (or a *test of homogeneity*) is to determine whether the distributions behind two sets of data are the same or not. More precisely, for two given independent samples $\{x_{i,0}\}_{i=1}^{n_0}$ and $\{x_{i,1}\}_{i=1}^{n_1}$ from d-dimensional distributions $P(\cdot) = \mathbb{P}(\cdot|Y = 0)$ and $Q(\cdot) = \mathbb{P}(\cdot|Y = 1)$, where the group labels $Y = 0$ and $Y = 1$ indicate the two samples, the goal is to test

$$\mathrm{H}_0 : P = Q \quad \text{against} \quad \mathrm{H}_1 : P \neq Q. \tag{2}$$

A classic method for testing homogeneity is the two-sample t-test, which compares the means of two Gaussian distributions. There are several nonparametric extensions of classical t-tests; for example, *Maximum Mean Discrepancy* (MMD) [Gretton et al., 2012a], which compares the means of two distributions in Reproducing Kernel Hilbert Spaces (RKHSs) and the *energy distance*, [Székely and Rizzo, 2004], which is a member of MMD with the RKHS induced by a positive definite kernel [Sejdinovic et al. (2013)].
Another means to two-sample testing is to estimate a divergence between the two probability distributions of interest. The *f-divergence* is the most popular measure of divergence. It is defined as

$$D_f(P||Q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x)$$

where $dP = pd\mu$, $dQ = qd\mu$ and f is a convex function. This family of divergences includes many common divergences; such as, the *Kullback-Leibler divergence*, the *Pearson divergence*, the *Hellinger distance* and the *Total variation distance*. Two-sample tests based on a divergence have been investigated by many authors; see, for example, Sugiyama et al. (2011a) and Kanamori et al. (2012).

The methods mentioned above, however, only provide us with a global binary result of the form "Reject the null hypothesis" or "Fail to reject the null hypothesis." This type of binary result can leave much of the local information concealed. For our applications, we would like to know *how* the two distributions are different. More specifically, we would like to detect *locally* significant differences in a multivariate feature space. This question can be reformulated in terms of estimating highest density difference or highest density ratio regions in a sample space.

3.1 Local Two-Sample Tests

From the Bayes' theorem, testing (2) is equivalent to testing

$$\mathrm{H}_0 : \mathbb{P}(Y = 1|\mathbf{C}_i) = \mathbb{P}(Y = 1) \quad \text{against} \quad \mathrm{H}_1 : \mathbb{P}(Y = 1|\mathbf{C}_i) \neq \mathbb{P}(Y = 1) \tag{3}$$

for all arbitrary partitions $\cup_{i\in\mathcal{I}}\mathbf{C}_i = \mathbb{R}^d$. Since we would like to identify regions where the two distributions are significantly different, we consider a set of *local* alternatives instead of the global alternative in (3); each local alternative is restricted within $\mathbf{C}_i$, which leads to multiple testing with

$$\begin{aligned}
&\mathrm{H}_{1,1} : \mathbb{P}(Y = 1|\mathbf{C}_1) \neq \mathbb{P}(Y = 1)\\
&\mathrm{H}_{1,2} : \mathbb{P}(Y = 1|\mathbf{C}_2) \neq \mathbb{P}(Y = 1)\\
&\qquad\vdots\\
&\mathrm{H}_{1,m} : \mathbb{P}(Y = 1|\mathbf{C}_m) \neq \mathbb{P}(Y = 1)
\end{aligned}$$

where m can be infinite.

In the flow cytometry literature, Roederer and Hardy (2001) addressed the question of finding differences between two samples in a multi-dimensional space. Their method partitions the space into box-shaped sub-regions and then relies on chi-squared tests to identify local significant differences. This approach can be generalized to moderately high dimensions but, due to the rectangular shaped partitions, there is room for improvement, especially when the underlying differences between the two distributions vary smoothly in sample space. Moreover, to capture detailed local structures, it is natural to shrink the volume of each region $\mathbf{C}_i$ as the sample size increases, eventually approaching a point-wise test in the limit of large sample sizes.

Hence, building on the work of Roederer and Hardy (2001), we instead propose a point-wise test for differences at grid points $(x_1, \ldots, x_m)$ in sample space. More precisely, we test against m local alternatives,

$$\mathrm{H}_{\mathrm{i},0} : \mathbb{P}(Y = 1|X = x_i) = \mathbb{P}(Y = 1) \quad \text{vs.} \quad \mathrm{H}_{\mathrm{i},1} : \mathbb{P}(Y = 1|X = x_i) \neq \mathbb{P}(Y = 1)$$

for $x_i \in \mathbb{R}^d$ and i=$1, \ldots, m$.

Indeed, the point-wise local alternatives are also studied by Duong (2013) who use kernel density estimates (KDEs) to find locally significant differences between two samples. Our contribution here is to combine point-wise testing with a supervised learning method (such as regression) that does not rely on estimating densities.

As a test statistic, we propose

$$T_i = \widehat{\mathbb{P}}(Y = 1|X = x_i) - \widehat{\mathbb{P}}(Y = 1), \tag{4}$$

where $\widehat{\mathbb{P}}(Y = 1)$ is the number of times the label $Y = 1$ occurs for a sample of size n. The main challenge is to estimate the "class posteriors" $\mathbb{P}(Y = 1|X = x_i)$ at the m different

testing points. An advantage of our local two-sample test is that we can take advantage of the many already existing regression methods for multi- or high-dimensional data. By choosing a suitable regression method, we can adapt to different types of structure in the data as well as different types of data, potentially achieving a high power for local tests.

3.1.1 Permutation Test

The sampling distribution of T_i varies by test statistic. In this section, we first present a general framework for carrying out local two-sample tests by using the permutation test. Then, later in Section 3.1.2, we describe a local significance test based on the asymptotic normality of random forest regression estimators.

The permutation test is attractive because it places no assumptions on the data, other than that the observations are mutually exchangeable under the null hypothesis. The permutation test is also known in many cases to have similar power as tests based on large-sample theory. Due to computational cost considerations, the permutation p-value is usually estimated (instead of computed exactly) by $\widehat{p} = m/B$, where m is the number of times the permutation test statistic is greater than or equal to the observed test statistic, and B is the total number of random permutations of the data. The estimator $\widehat{p}$ is unbiased but tends to underestimate the type I error rate. Hence, in our work we instead use the biased estimator $\widehat{p}^* = (m+1)/(B+1)$, which has the same computational cost as $\widehat{p}$ but is more suitable for multiple comparisons adjustments (Ernst, 2004 and Phipson and Smyth, 2011).

Note that many regression methods for estimating $\mathbb{P}(Y = 1|X = x)$ involve one or more tuning parameters. Examples include the number of nearest neighbors k in the k-Nearest Neighbors (k-NN) method, and the maximum depth of a tree in decision tree models. In Appendix D, we present a general procedure for choosing tuning parameters by minimizing an estimated mean-squared error loss.

Below we summarize the main steps of the local two-sample test at $\{x_i\}_{i=1}^m$ testing points.

Algorithm 1: Local two-sample test via permutation

---

(1) Given *i.i.d.* samples $\{x_{j,0}\}_{j=1}^{n_0}$ and $\{x_{j,1}\}_{j=1}^{n_1}$, calculate the test statistic $\{T_i\}_{i=1}^m$ at the m testing points.

(2) Sample without replacement from the combined pool of $\{x_{j,0}\}_{j=1}^{n_0}$ and $\{x_{j,1}\}_{j=1}^{n_1}$ to create two permuted samples of sizes n_0 and n_1, respectively. Compute the test statistic again using the permuted data.

(3) Repeat step (2) B times to obtain $\{T_i^{(1)}\}_{i=1}^m, \cdots, \{T_i^{(B)}\}_{i=1}^m$. Then approximate the permutation p-value at each testing point i by

$$P_i = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathrm{I}_{\{|T_i^{(b)}|>|T_i|\}}\right).$$

(4) Apply the Benjamini-Hochberg (BH) method to adjust the m local hypothesis tests. Start by sorting the p-values in ascending order $p_{(1)}, \ldots, p_{(m)}$. Define $l_i = \frac{i\alpha}{C_m m}$ and $R = \max\{i : p_{(i)} < l_i\}$ where $C_m = \sum_{i=1}^m (1/i)$ and α is a given significance level. Reject the null in favor of the alternative hypotheses $\mathrm{H}_{(1),1}, \ldots, \mathrm{H}_{(R),1}$ for which $p_{(i)} \leq p_{(R)}$. For the local points $x_{(1)}, \ldots, x_{(R)}$ in the rejection region,

   (a) if $T_i > 0$, then decide that $\mathbb{P}(Y = 1|x_i) > \mathbb{P}(Y = 1)$ or $f(x_i|Y = 1) > f(x_i|Y = 0)$;

   (b) if $T_i < 0$, then decide that $\mathbb{P}(Y = 1|x_i) < \mathbb{P}(Y = 1)$ or $f(x_i|Y = 1) < f(x_i|Y = 0)$.

In Figure 7, we illustrate the local two-sample test for a simple toy example with two Gaussian mixture distributions. There are two groups of observations with $n_1 = 500$ and $n_2 = 500$ data points sampled from the density functions $\frac{1}{2}\mathrm{N}(-3, 0.5^2)+\frac{1}{2}\mathrm{N}(1, 0.5)^2$ and $\frac{1}{2}\mathrm{N}(-1, 0.5^2)+\frac{1}{2}\mathrm{N}(3, 0.5^2)$, respectively. We perform tests for a fixed grid of evenly spaced points between -7 to 7 that are 0.02 apart. The red and blue points in the final decision area indicate the locally significant regions, whereas gray points represent the regions where there are no significant differences in distribution. We used the k-NN regressor to calculate the test statistic with FDR= 0.05. The k-NN regressor is defined as

$$\widehat{\mathbb{P}}(Y = 1|x) = \frac{1}{k}\sum_{j\in J} \mathrm{I}(Y_j = 1)$$

where $J = \{i : X_i$ is one of the k observations nearest to $x\}$, and k is chosen according to Appendix D.

3.1.2 Test based on Asymptotic Normality

As mentioned before, the advantage of the permutation method is that we can control the exact type I error as well as easily use different types of estimators to perform a valid test. On the downside, the null hypothesis of the permutation test corresponds to a global null hypothesis of the form $\mathrm{H}_0 : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1)$ *for all* $x \in \mathbb{R}^d$, rather than the local null hypothesis: $\mathrm{H}_0 : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1)$ *for a given* $x \in \mathbb{R}^d$. To offer a truly

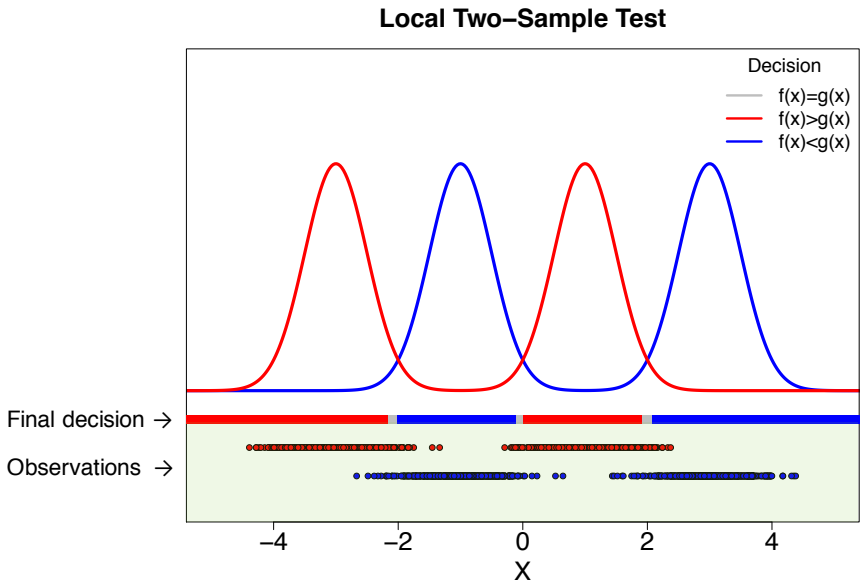


Figure 7: Graphical illustration of the local two-sample test (Algorithm 1) for two Gaussian mixtures in $\mathbb{R}^1$. See Section 3.1.1 for details.

point-wise test, we here provide an alternative way to test for locally significant differences using the large-sample limit of the test statistic.

Since our application contain data of mixed types, we choose to use random forests to estimate the class posteriors in eq. 4. In a recent paper, Wager and Athey (2015) describe a variant of random forests with predictions that are both asymptotically unbiased and Gaussian. To satisfy a condition they call "honesty," the base tree is grown using one subsample, while the predictions are estimated using a different subsample. We use the result from Wager and Athey (2015) to construct asymptotic confidence intervals that are centered at the unknown function $\mathbb{P}(Y = 1|X = x)$. To be precise, we define a test statistic

$$T_n(x) = \frac{\widehat{\mathbb{P}}(Y = 1|x) - \widehat{\mathbb{P}}(Y = 1)}{\sqrt{\widehat{V}_{IJ}(x)}} \xrightarrow{d} N\left(\mathbb{P}(Y = 1|x) - \mathbb{P}(Y = 1), 1\right), \tag{5}$$

where $\widehat{\mathbb{P}}(Y = 1|x)$ is estimated by a forest of double-sample regression trees with binary outcomes Y, and $\widehat{V}_{IJ}(x)$ is a consistent estimator of the variance of the numerator based on the infinitesimal jackknife (Wager et al. 2014).

Below we summarize the local two-sample test based on asymptotic normality of the test statistic. As in Algorithm 1, we use the Benjamini-Hochbergh procedure to control the false

discovery rate (FDR).

Algorithm 2: Local two-sample test via asymptotic normality

---

(1) Given *i.i.d.* samples $\{x_{j,0}\}_{j=1}^{n_0}$ and $\{x_{j,1}\}_{j=1}^{n_1}$, calculate the test statistic $\{T_i\}_{i=1}^{m}$ at the m testing points.

(2) Calculate the p-value of each T_i based on the normal approximation. Namely, the p-value of the test statistic evaluated at ith grid point is

$$p_i = 2\Phi(-|T_i|)$$

where Φ is the distribution function of the standard normal random variable.

(3) Apply the Benjamini-Hochberg (BH) method to adjust the m local hypothesis tests. Start by sorting the p-values in ascending order $p_{(1)}, \ldots, p_{(m)}$. Define $l_i = \frac{i\alpha}{C_m m}$ and $R = \max\{i : p_{(i)} < l_i\}$ where $C_m = \sum_{i=1}^{m}(1/i)$ and a given significance level α. Reject the null in favor of the alternative hypotheses $\mathrm{H}_{(1),1}, \ldots, \mathrm{H}_{(R),1}$ for which $p_{(i)} \leq p_{(R)}$. For the local points $x_{(1)}, \ldots, x_{(R)}$ in the rejection region,

   (a) if $T_i > 0$, then decide that $\mathbb{P}(Y = 1|x_i) > \mathbb{P}(Y = 1)$ or $f(x_i|Y = 1) > f(x_i|Y = 0)$;

   (b) if $T_i < 0$, then decide that $\mathbb{P}(Y = 1|x_i) < \mathbb{P}(Y = 1)$ or $f(x_i|Y = 1) < f(x_i|Y = 0)$.

3.2 Diffusion maps

Dimensionality reduction methods can be useful for visualizing and describing low-dimensional structures that are embedded in a high-dimensional space. In this work, we use *diffusion maps* (Coifman and Lafon 2006) to visualize and inspect the results from the point-wise analysis. Diffusion maps is a nonlinear data reduction technique that aims to preserve the connectivity structure of the data, where "connectivity" is learnt by propagating local information through a diffusion process.

As a starting point for constructing the diffusion map, one first defines a weight that reflects the local similarity of two points x_i and x_j in $\mathcal{X} = \{x_1, \ldots, x_n\}$. A common choice is the Gaussian kernel

$$w(x_i, x_j) = \exp\left(-\frac{s(x_i, x_j)^2}{\epsilon}\right) \tag{6}$$

where $s(x_i, x_j)$ is some distance function such as the Euclidean distance. These weights are used to build a Markov random walk on the data with the transition probability from x_i to x_j defined as

$$p(x_i, x_j) = \frac{w(x_i, x_j)}{\sum_{k \in \Omega} w(x_i, x_k)}.$$

The one-step transition probabilities are stored in an $n \times n$ matrix denoted by $\mathbf{P}$, and then propagated by a t-step Markov random walk with transition probabilities $\mathbf{P}^t$. Instead of choosing a fixed time parameter t, however, we here combine diffusions at all times (Coifman et al. 2005) and define an averaged diffusion map according to

$$\Psi_{\text{av}} : x \mapsto \left[\left(\frac{\lambda_1}{1 - \lambda_1} \right) \psi_1(x), \left(\frac{\lambda_2}{1 - \lambda_2} \right) \psi_2(x), \ldots, \left(\frac{\lambda_m}{1 - \lambda_m} \right) \psi_m(x) \right],$$

where λ_i and ψ_i, respectively, represent the first mth eigenvalues and the corresponding right eigenvectors of $\mathbf{P}$.

In our application, we also use a generalization of the weight in (6) proposed by Zelnik-Manor and Perona (2005) for spectral clustering. In their paper, the authors show that a data-driven varying bandwidth leads to more meaningful clustering results for data with multiple scales and propose the weight

$$\widehat{w}(x_i, x_j) = \exp\left(-\frac{s(x_i, x_j)^2}{\sigma_i \sigma_j} \right),$$

where σ_i is some distance between x_i and the kth neighbor of x_i. For our visualization purposes, we choose $m = 2$ and $k = 30$, but there are other values that give similar results.

4 Results

Our ultimate goal is to identify the processes driving the structural evolution of galaxies, as well as understand the connection between morphology and galaxy properties such as stellar mass and star-formation rate. Here we show that one can use the statistical techniques in Sec. 3.1.2 to tease out the main morphological differences between two populations. We use an example where the two populations both fall within the same redshift bin (in this case, the i-band or $0.56 < z < 1.12$).

For visualization purposes, we will display all our results in a two-dimensional diffusion map computed from the seven morphological indicators defined in Sec. 2.3 and Appendix C. As Figure 8 shows, this diffusion map roughly organizes the galaxies according to their structure, with galaxies with similar morphology falling into nearby regions in diffusion space.

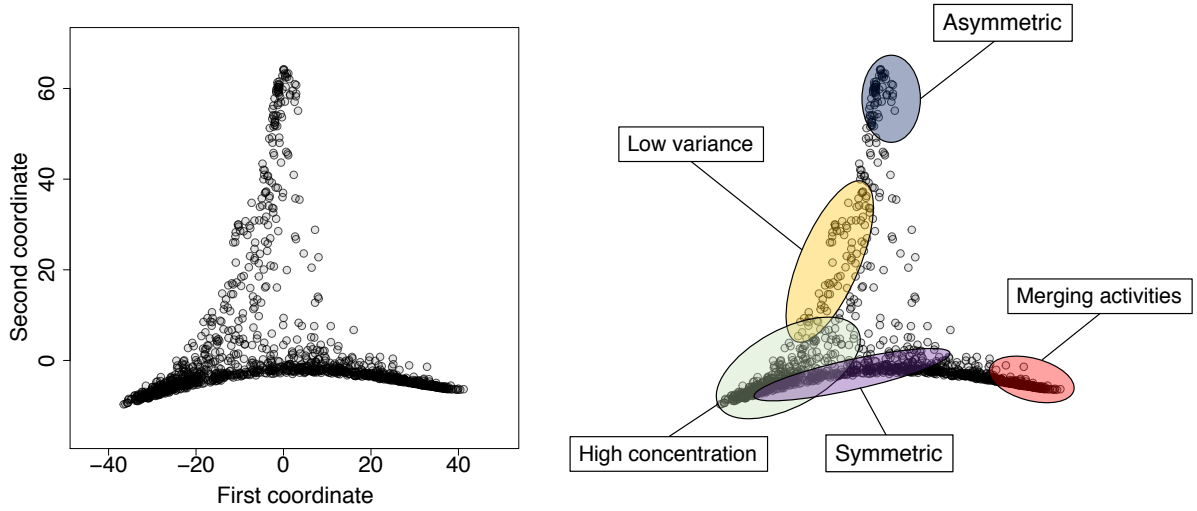


Figure 8: Diffusion map of galaxies observed in the i band constructed from seven summary statistics. The diagram on the right describes the characteristics of galaxies in different regions of the map (cf. Figures 30 and 31).

4.1 Mass Study

We start by comparing the structural differences between high-mass and low-mass galaxies in the i-band. We say that a galaxy belongs to the high-mass group if its mass is greater than the upper c-quantile of the mass distribution whereas the galaxy belongs to the low-mass group if its mass is less than the lower c-quantile of the mass distribution (see Figure 9). For our study, we choose the cut-off value $c = 0.25$, which corresponds to a total of n=868

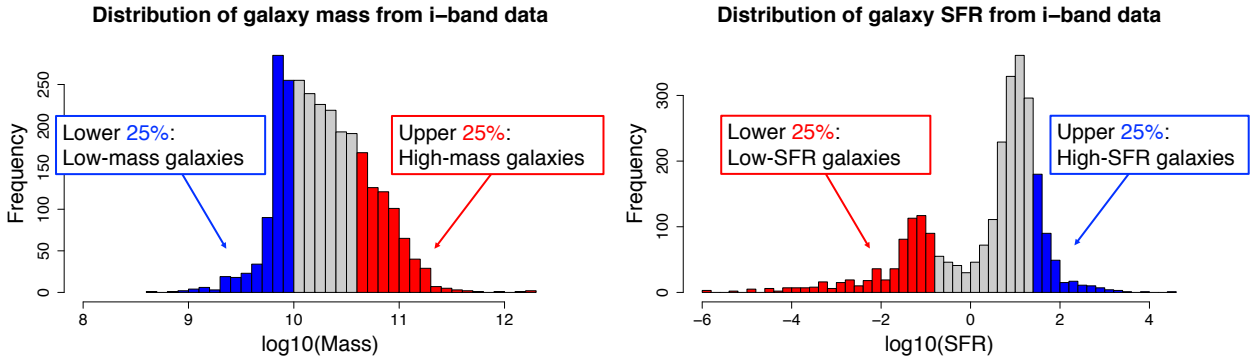


Figure 9: Distribution of $\log_{10}(Mass)$ *and* $\log_{10}(SFR)$ *for i-band-selected data.*

galaxy observations in the two mass groups. (A study of robustness with respect to changes in cutoff is provided in Appendix B.)

We use *Algorithm 2* and random forest regression to identify the regions where the two multivariate samples differ in density. Figure 10 shows the result of multiple testing under FDR control at the significance level $\alpha = 0.05$, where the test points are chosen according to the remark below and displayed in a two-dimensional diffusion map.

Remark 4.1. *With increasing dimension, it becomes computationally infeasible to apply Algorithm 2 to a fine grid of uniformly spaced test points in morphology space. Hence, we only test for differences at the location of the observed data. We split the data into training (70%) and test sets (30%) where the training set is used to define the test statistics in eq. 5 and the test sets are used as testing points for the local two-sample test.*

We note that the high-mass and low-mass dominated regions are well-separated in diffusion space, and hence also in the seven-dimensional morphology space. Specifically, high-mass dominated regions (red) tend to coincide with areas characterized by "high concentration" and "low variance," whereas low-mass dominated regions (blue) coincide with areas with a high ratio of "merging activities."

The parallel coordinate plot (Inselberg, 1997) in Figure 11 provides further insight on how the significant points are distributed with respect to the original morphology features. The data are clearly separated within the M_{20} and C coordinates, which implies that the variance measure (M_{20}) and the concentration measure (C) are key to distinguishing between the high-mass and low-mass groups. This result is also consistent with the variable importance measures from random forests, where Table 23 indicates that M_{20} and C are the two most important variables in estimating the test statistic in eq. (5).

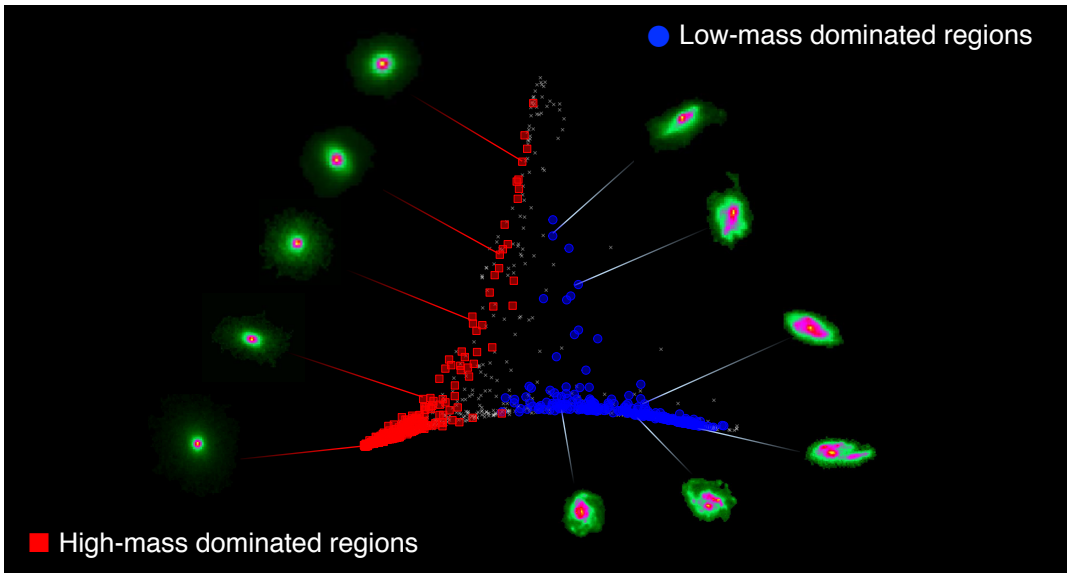


Figure 10: Result of local two-sample testing of differences between high- and low-mass galaxies in a seven-dimensional morphology space. The blue color indicates regions where the density of high-mass galaxies are significantly higher, and the red color indicates the regions that are dominated by low-mass galaxies. The test points are visualized in a two-dimensional diffusion map.

To facilitate a physical interpretation of the local test results, we finally divide the significant testing points into galaxy groups (classes) using an agglomerative hierarchical clustering. Figures 12 and 13 show the results from hierarchical clustering of the two difference regions using the first two diffusion coordinates and the Ward2 algorithm (Murtagh and Legendre, 2014) with complete linkage.

Table 14 shows a random subset of images from the five different groups, and Figure 15 shows boxplots of the morphology statistics of each group.

Note that the group LowMass-2 includes a very large number of multi-modal (high ratio of non-zero M, I, D), high variance (M_{20}) and low concentration (C) galaxies. The group LowMass-1 has similar characteristics with a slightly lower proportion of non-regular galaxies than LowMass-2.

We also note that the three high-mass clusters are all described by high concentration (C), low variance (low M_{20}, high G) and a low proportion of nonzero M, I, D statistics. The galaxies from HighMass-1 are especially concentrated compared to galaxies from the other high-mass clusters. The HighMass-3 group is characterized with high asymmetry (A).

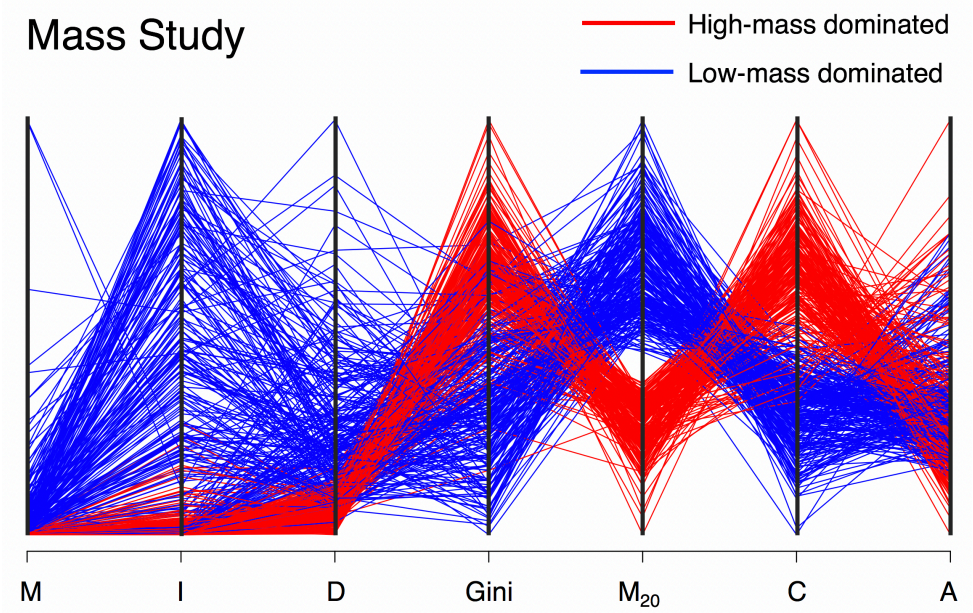


Figure 11: Parallel coordinate plot of the significant points in the comparison of high- and low-mass galaxies.

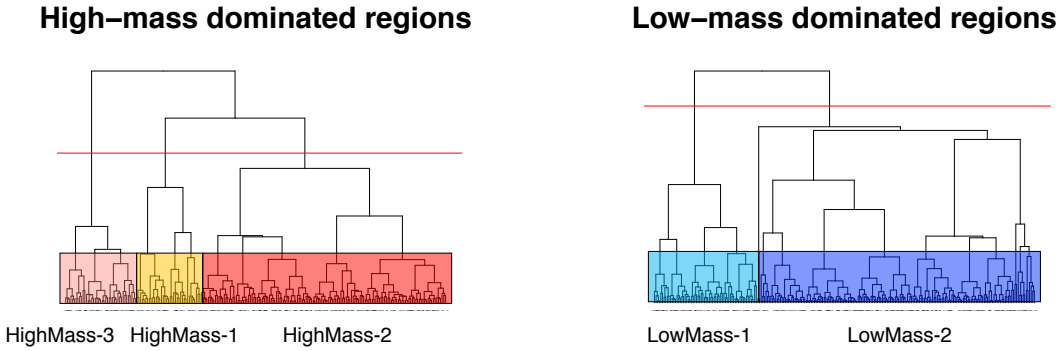


Figure 12: Hierarchical clustering of the two significant regions from the Mass study using the first two diffusion coordinates.

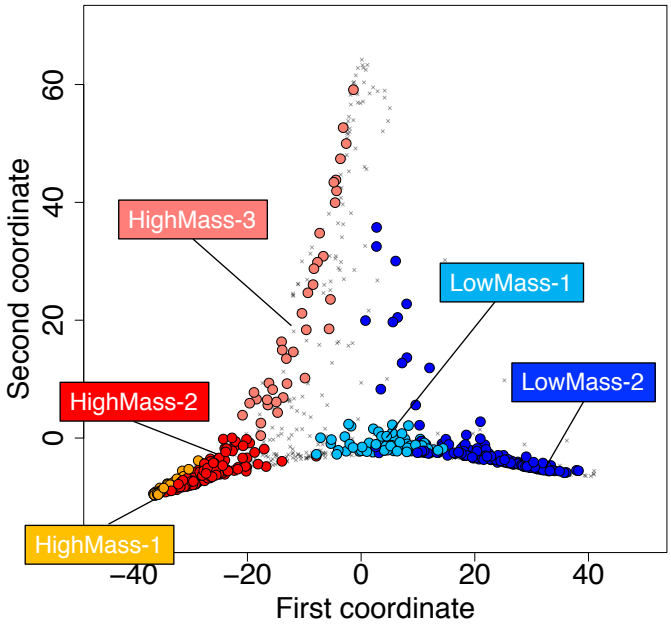
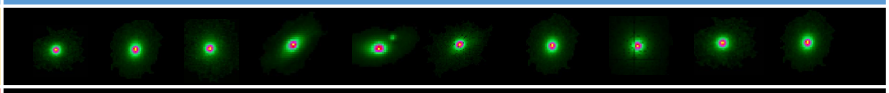
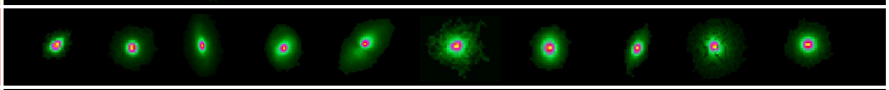
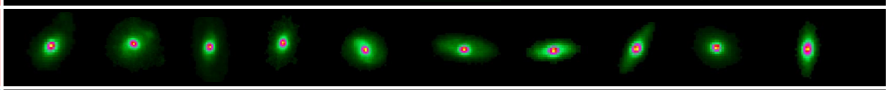
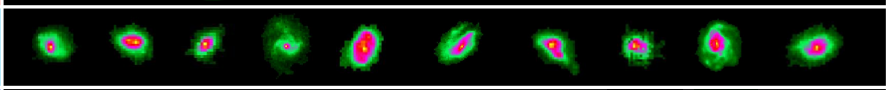
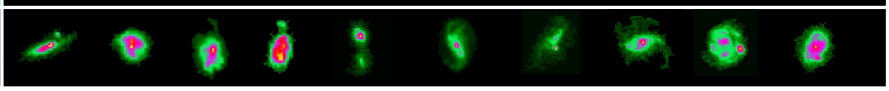
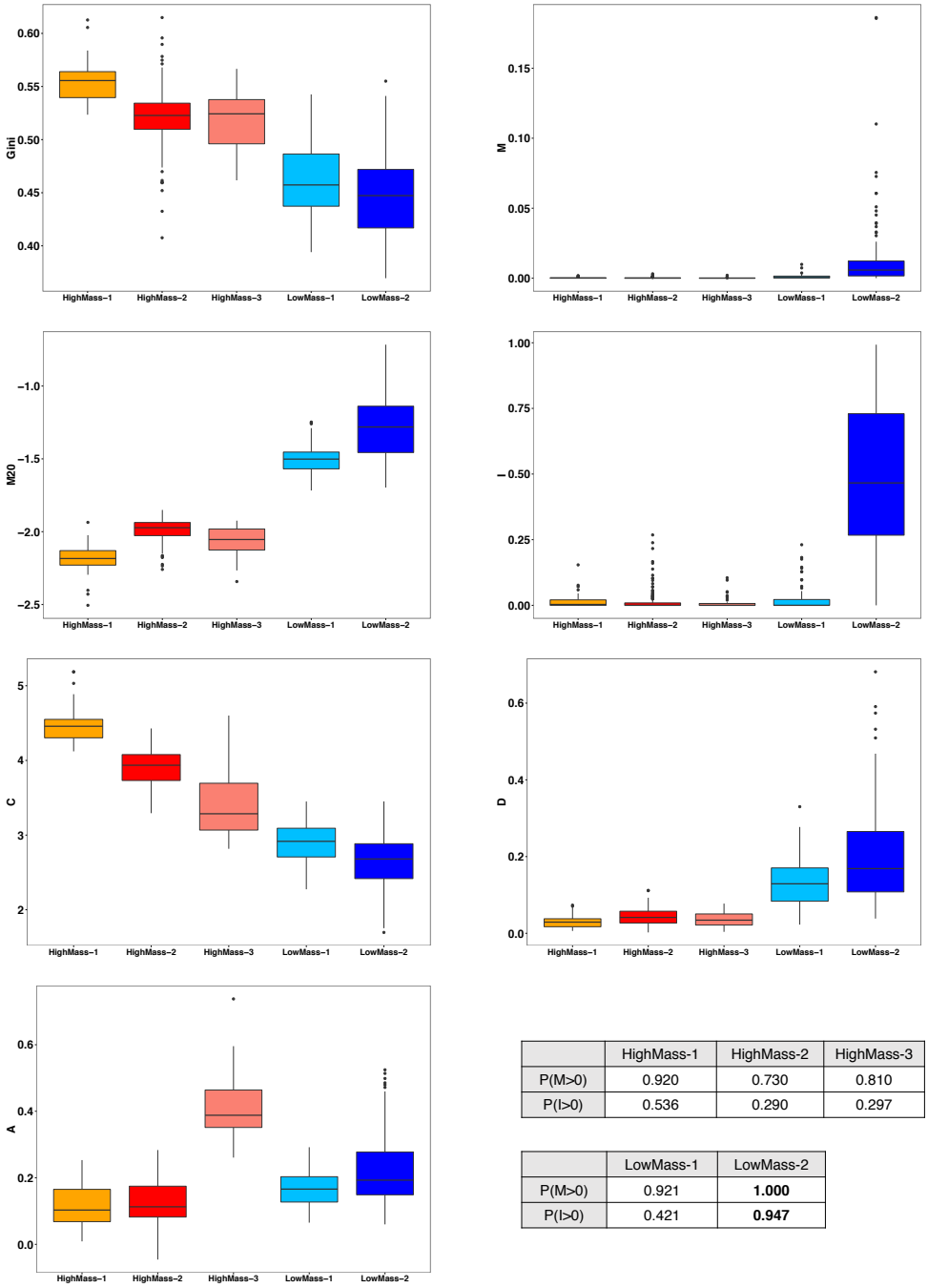


Figure 13: Diffusion map with the hierarchical clustering result for the Mass study.

Table 14: Randomly chosen galaxy images from the significant clusters in the Mass study.

	Random galaxy images
HighMass-1	
HighMass-2	
HighMass-3	
LowMass-1	
LowMass-2	



	HighMass-1	HighMass-2	HighMass-3
P(M>0)	0.920	0.730	0.810
P(I>0)	0.536	0.290	0.297

	LowMass-1	LowMass-2
P(M>0)	0.921	**1.000**
P(I>0)	0.421	**0.947**

Figure 15: Boxplots of summary statistics for the significant clusters in the Mass study together with a table (bottom right) of the proportion of galaxies with $M > 0$ or $I > 0$.

4.2 SFR Study

We next study the connection between morphology and star-formation rate (SFR). Similar to the previous study, we divide the galaxies into two samples but now based on their star-formation rate instead of stellar mass. We choose the same cutoff value $c = 0.25$ as before, and use Algorithm 2 to identify local differences between the two samples (high- versus low-SFR galaxies) in our seven-dimensional morphology space.

Figure 16 shows the results of the local significance test. Note that the regions where the two samples are significantly different roughly coincide with the difference regions of the Mass study (Figure 10). To be specific, low-SFR dominated regions coincide in diffusion space with high-mass dominated regions, and high-SFR dominated regions occur roughly in the same parts of the diffusion space as the low-mass dominated regions.

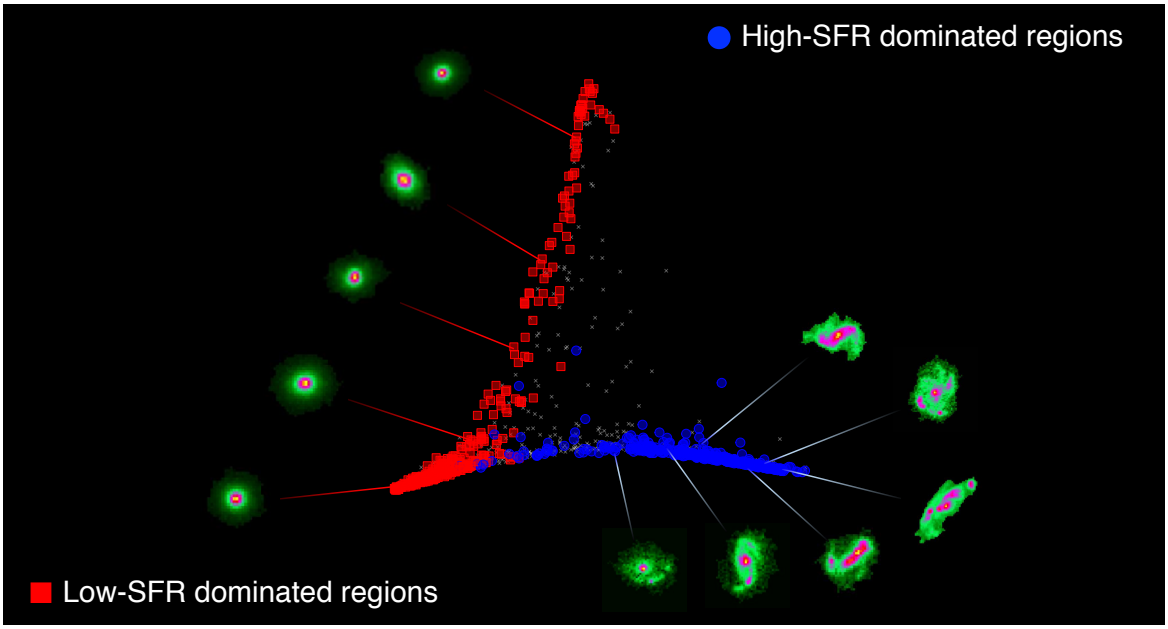


Figure 16: Result of local two-sample testing of differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red color indicates regions where the density of low-SFR galaxies are significantly higher, and the blue color indicates the regions that are dominated by high-SFR galaxies. The test points are visualized in a two-dimensional diffusion map.

To more easily interpret the local test results, we also divide the significant points into five galaxy groups or classes (Low-SFR 1, 2 and 3 and High-SFR 1, 2) using agglomerative hierarchical clustering; see Figures 17-18 and Table 19. Figure 20 is a parallel coordinate plot

of the significant points in the SFR study, and Figure 21 summarizes their characteristics with respect to the seven morphology indices after dividing the points into the five galaxy classes.

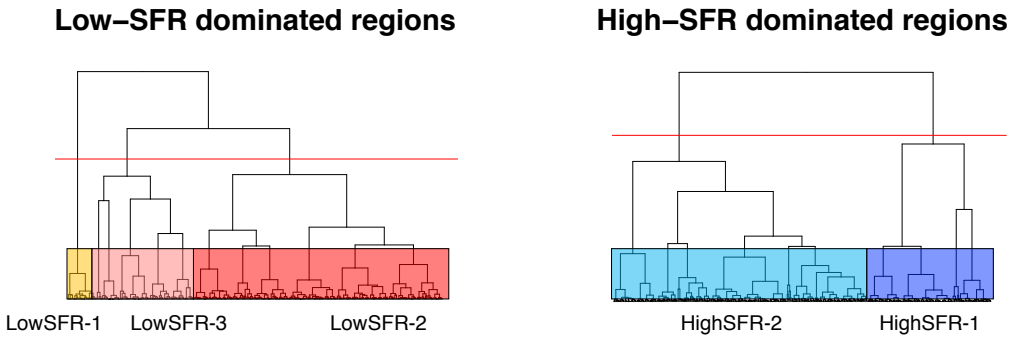


Figure 17: Hierarchical clustering of the two significant regions from the SFR study using the first two diffusion coordinates and the complete-linkage criterion.

The clustering results of the SFR study are similar to the results of the Mass study. Comparing Figure 13 and Figure 18, we see the following correspondence: (LowSFR-1 & LowSFR-2, HighMass-1 & HighMass-2) / (LowSFR-3, HighMass-3) / (HighSFR-1, LowMass-2) / (HighSFR-2, LowMass-1). In Section 4.3, we compare this result more systematically.

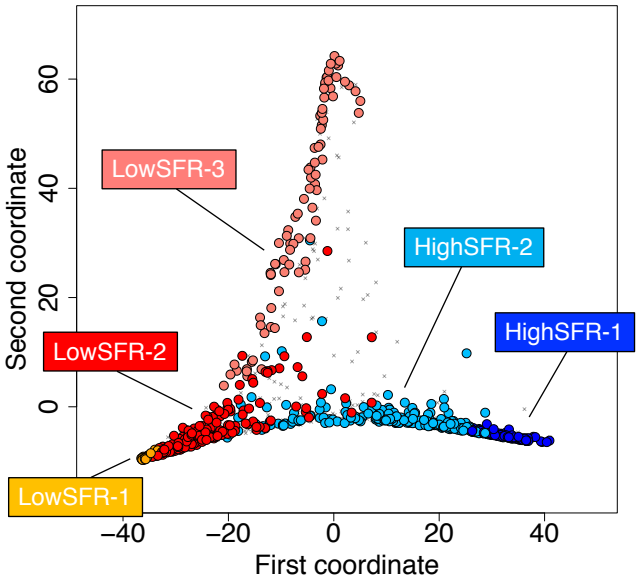
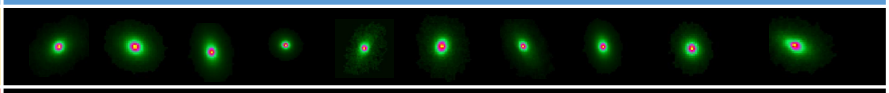
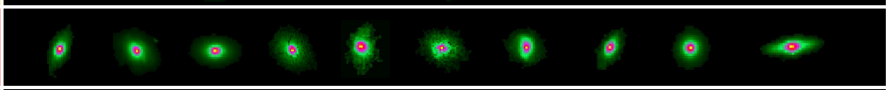
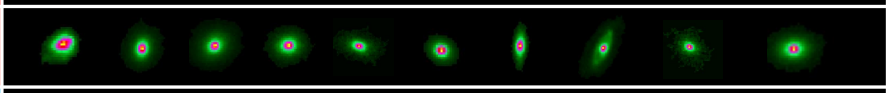
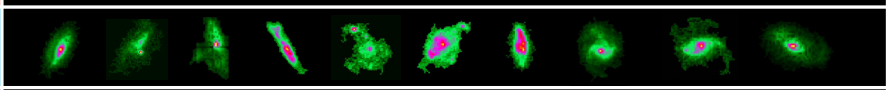
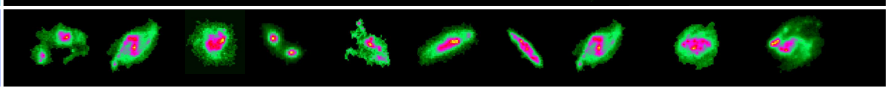


Figure 18: Diffusion map with the hierarchical clustering result for the SFR study.

Table 19: Randomly chosen galaxy images from the significant clusters in the SFR study.

	Random galaxy images
LowSFR-1	
LowSFR-2	
LowSFR-3	
HighSFR-2	
HighSFR-1	

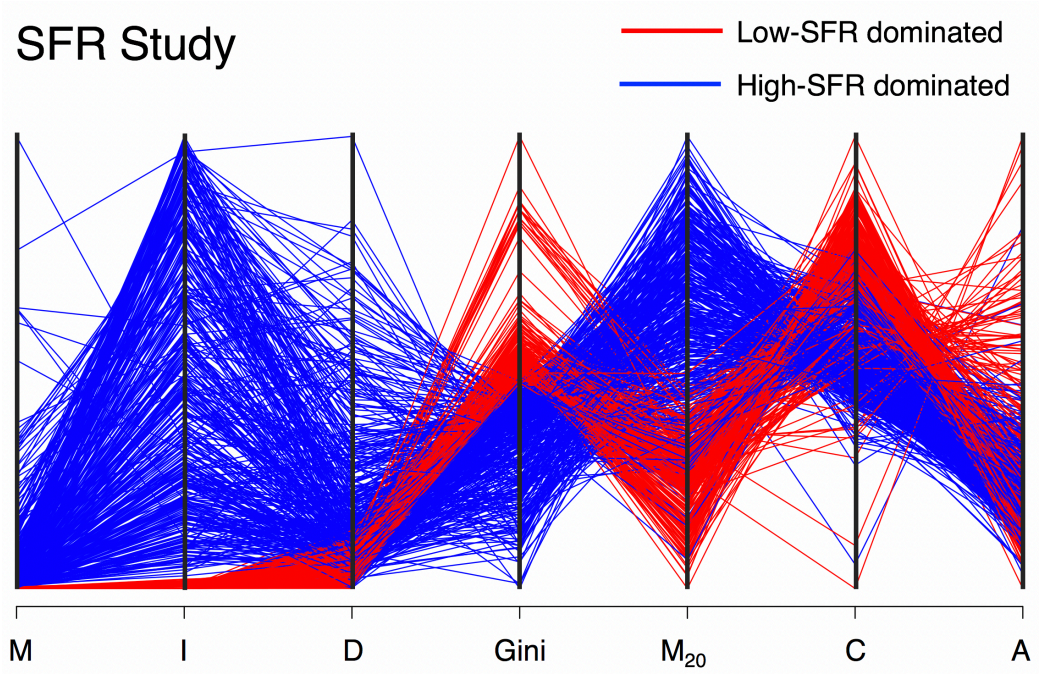
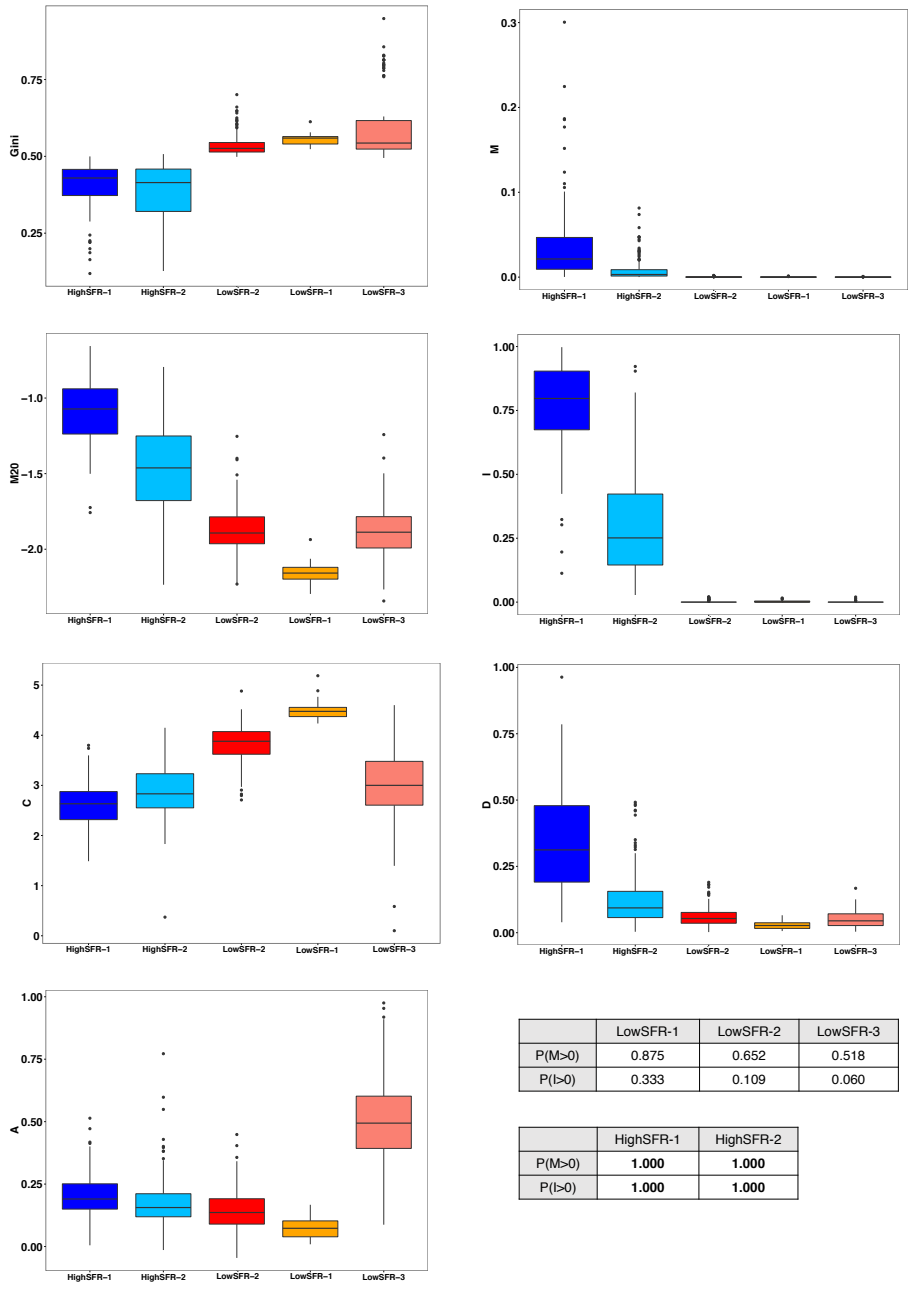


Figure 20: Parallel coordinate plot of the significant points in the SFR study.



	LowSFR-1	LowSFR-2	LowSFR-3
P(M>0)	0.875	0.652	0.518
P(I>0)	0.333	0.109	0.060

	HighSFR-1	HighSFR-2
P(M>0)	**1.000**	**1.000**
P(I>0)	**1.000**	**1.000**

Figure 21: Boxplots of summary statistics for the significant clusters in the SFR study together with a table (bottom right) of the proportion of galaxies with $M > 0$ or $I > 0$.

4.3 Merging the Mass and SFR studies

In this section, we merge the Mass and SFR studies (Figure 22), and investigate similarities and differences in the significance results at the sample points.

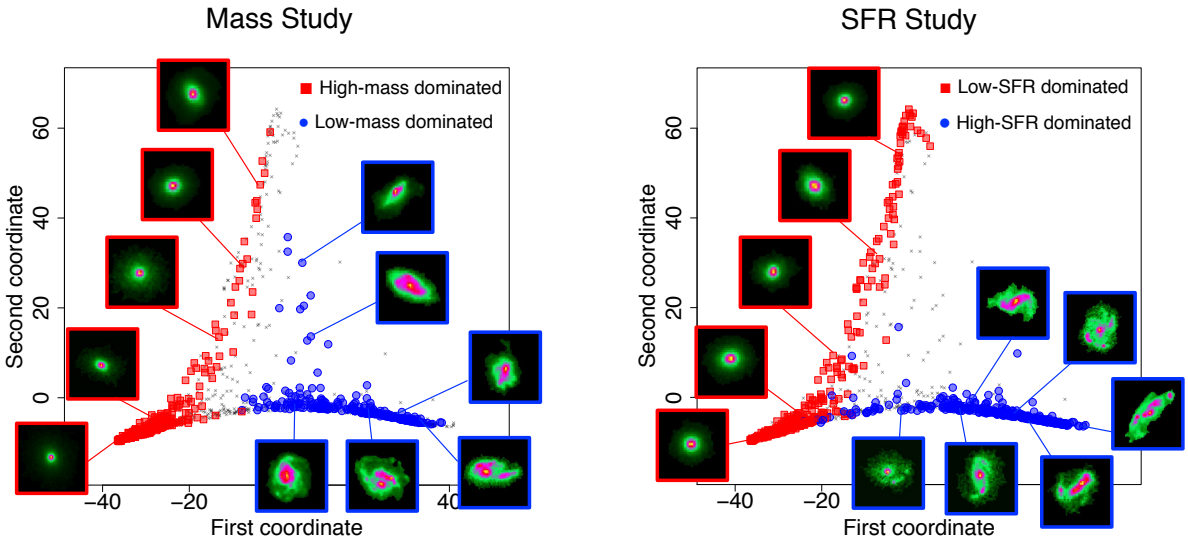


Figure 22: The results of the pointwise two-sample tests for the Mass study (left) and the SFR study (right) with example images attached.

Table 23 lists the variable importance measure in the random forest regression; that is, the table shows the relevance of each summary statistic for estimating the class posterior $\mathbb{P}(Y = 1|X = x)$ in eq. 5. The G statistic plays an important role in both studies. A noticeable difference between the two studies is in the (M, I) and (M_{20}, C) statistics: for the SFR study, both M and I are essential in distinguishing between high-SFR and low-SFR galaxies, whereas M_{20} and C are more important in distinguishing between high-mass and low-mass galaxies. These results are consistent with the previous findings of Sections 4.1 and 4.2. In fact, a closer look at the lower right edge of *LowMass-2* and *High-SFR 1* in Figures 13 and 18, reveals that there are more significant points from this part of diffusion space in the SFR study. This region is also where merger activities are prevalent (Figure 24 shows some randomly chosen images). Hence, our result indicates that the star-formation rate is more tightly linked to merger activities than the stellar mass is.

Figure 25 and Table 26 jointly summarize the significance results of the mass and SFR comparisons at the i-band sample points. As expected, there is a clear correspondence between

Table 23: Variable importance measure in the random forest regression

	Variable Importance Measure[1)]	
	Mass	SFR
M	3.72	**4.47**
I	2.26	**7.60**
D	4.04	2.67
Gini	**4.96**	**9.13**
M_{20}	**9.09**	4.40
C	**5.74**	3.52
A	2.96	2.50

[1)] mean decrease in node impurity measured by residual sum of squares

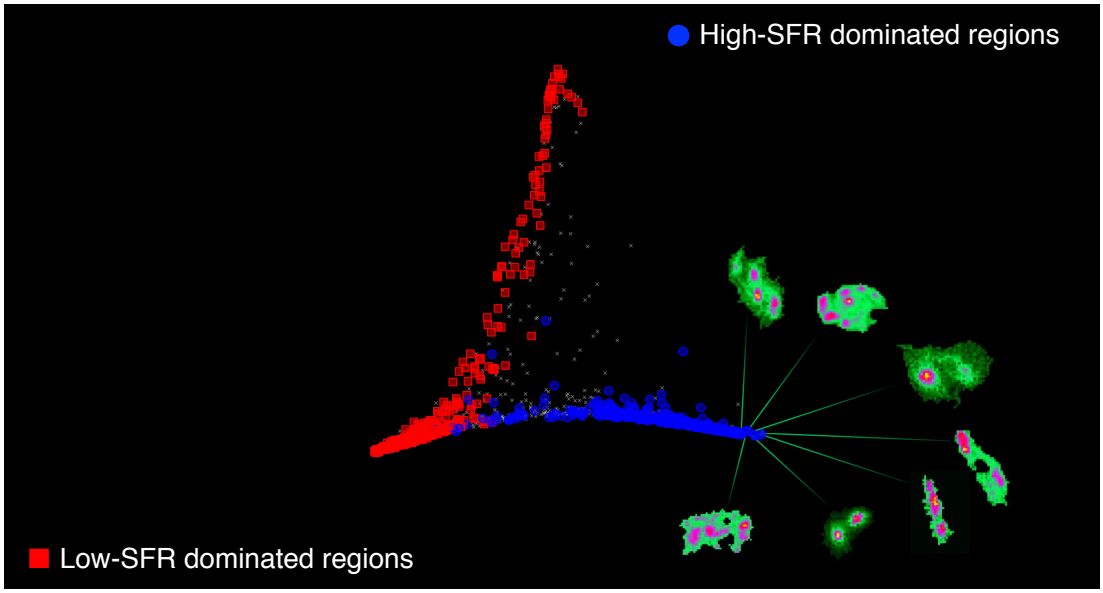


Figure 24: The lower right edge of the diffusion map has the most merger activities. Many of the sample points in this area are significant in the SFR study but not in the Mass study.

difference regions that are labeled as "high-mass dominated" versus "low-SFR dominated," and similarly correlations between "low-mass dominated" versus "high-SFR dominated" regions. We also see that there are more significant points in the SFR than mass comparison (683 versus 455 significant points, respectively, out of 1,000 test points total), but there are still many points (137 out of 1,000 test points total) where there are significant differences

between the two mass groups but not the two SFR groups. These outliers (as well as the galaxies that fall into the nine different entries of Table 26) could potentially be studied for further analysis.

To assess the consistency of our results with those in the astronomical literature, we construct G-M_{20} (Figure 27) and rest-frame UVJ (Figure 28) diagrams. (Our diagrams may be compared with, e.g., those in Figures 7 and 9 of Peth et al.)
In Figure 27, "merging," "bulge-dominated," and "disk-dominated" galaxy groups are defined by the line

$$G = -0.14M_{20} + 0.33$$

and below it the vertical line

$$M = -1.68\,.$$

Galaxies above the first line are the mergers, while those below the first line and to the left of the second are disk-dominated and those below and to the right are bulge-dominated. Our results are consistent with our earlier finding that M_{20} is more important than G for the Mass study in distinguishing the two (mass) populations, whereas G is more important than M_{20} for the SFR study in distinguishing the two (SFR) populations.
In Figure 28, "quenched" galaxies (as opposed to those that are "star-forming") are identified as lying above the locus defined by

$$U - V > \max[1.3, 0.88(V - J) + 0.59]$$

and to the left of the vertical line

$$V - J < 1.6\,.$$

The locus is defined by Williams et al. (2009), while the U, V, and J magnitudes are provided as part of the CANDELS catalog. We see that the results for our SFR study are consistent with expectation.

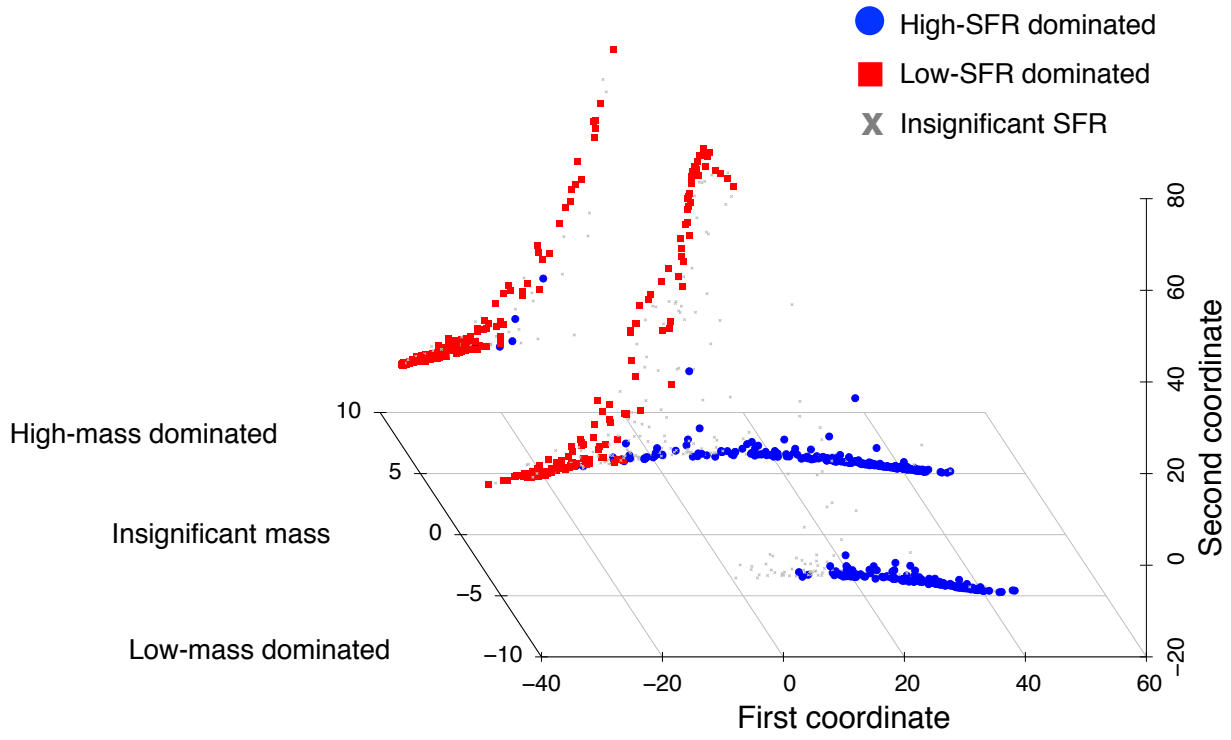


Figure 25: Summary of the significance results from the mass and SFR two-sample tests, wherein we artificially separate points at which the density of high-mass galaxies is significant, at which there are no significant differences, and at which the density of low-mass galaxies is significant.

Table 26: Summary of the significance results for the mass and SFR studies.

		SFR Study			
		Low-SFR	High-SFR	Insignificance	Total
Mass Study	High-Mass	155	5	59	219
	Low-Mass	0	158	78	236
	Insignificance	142	223	180	545
	Total	297	386	317	1,000

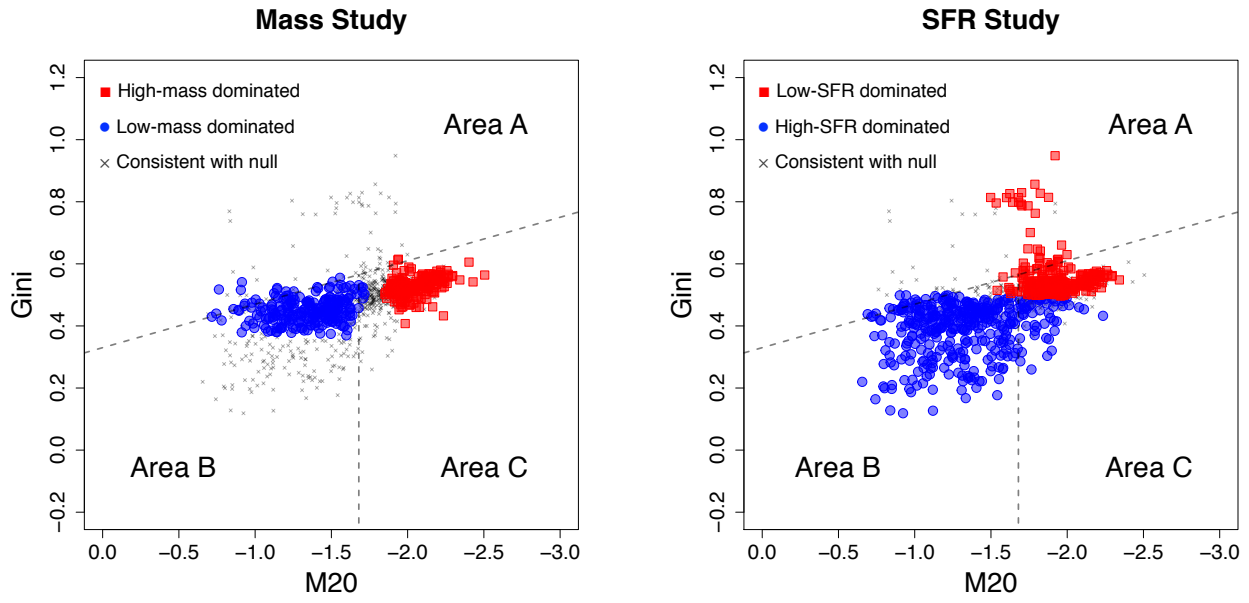


Figure 27: G-M_{20} for each significant group (cf. Figure 9 of Peth et al. 2016). M_{20} is more important than G for the Mass study in distinguishing the two populations, whereas G is more important than M_{20} for the SFR study. The dotted lines divide galaxy groups: mergers (Area A), disk-dominated (Area B), and bulge-dominated (Area C).

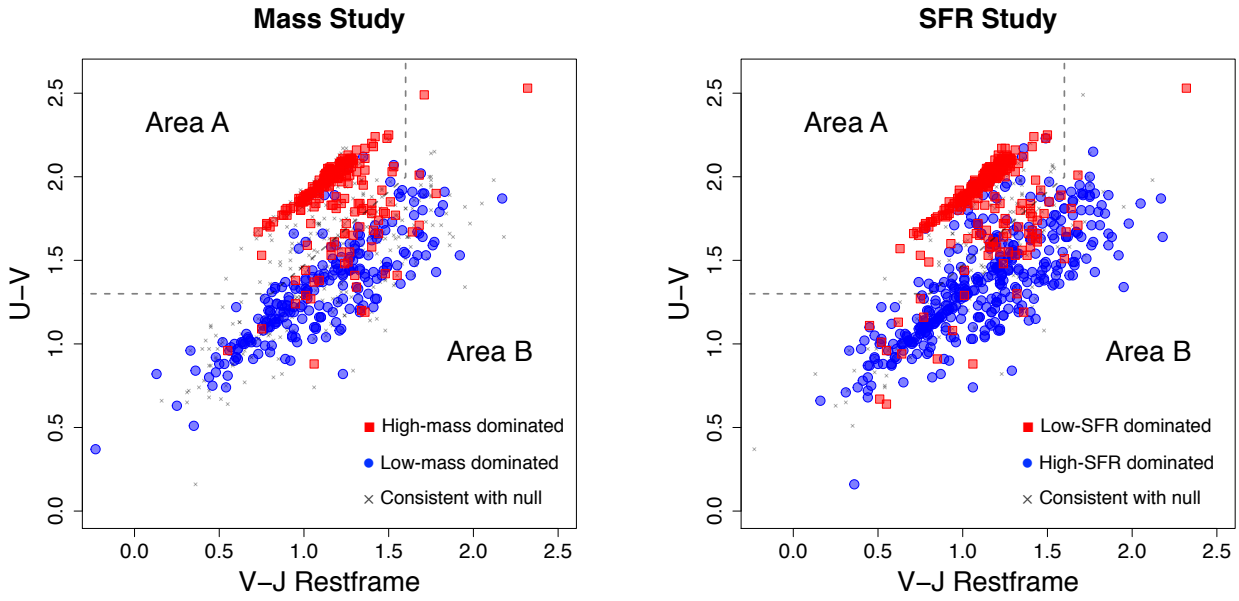


Figure 28: Rest-frame UVJ diagram for the Mass and the SFR studies (cf. Figure 7 of Peth et al. 2016). The dotted line separates galaxy groups: quenched (Area A) and star-forming galaxies (Area B).

5 Summary and Future Work

In this work, we investigate the connection between a galaxy's morphology and its physical properties−specifically, galaxy mass and star-formation rate−and show that we are able to identify and describe statistically significant differences between two defined populations of galaxies (e.g., those belonging to high- and low-mass quartiles).

As traditional morphology classes tend to oversimplify galaxy structures, we instead use what astronomers dub nonparametric morphology statistics to describe the appearance of a galaxy. We utilize seven oft-used statistics (M, I, D, G, M_{20}, C, A; Conselice 2003, Lotz et al. 2004, Freeman et al. 2013), while noting that it is an open question as to how much statistical information is preserved with these particular image summaries.

Our main results are that M_{20} and concentration measure (C) are the most important statistics for distinguishing between high- and low-mass galaxies, whereas the Gini coefficient (G) and the multi-mode and intensity statistics (M and I) turned out to be the most important ones when we look at star-formation rate instead (Table 23). The latter result is consistent with the fact that irregular and merging galaxies exhibit high rates of star-formation due to enhancement of gas density.

Furthermore, we found evidence that star-formation rate is more closely associated with the mentioned galaxy morphologies than galaxy mass is (Table 26). The results also show that star-formation rate and galaxy mass are negatively correlated to each other; that is, morphologies that are more common among high (low) mass galaxies tend to coincide with morphologies that are more common among low (high) SFR galaxies.

In future work, our method could be used to compare distributions of galaxy morphologies over different redshift ranges (after adjusting for those instrumental effects that differ between bands) as a test for evolution. Our method could also be used to test the validity of simulation methods by comparing simulated galaxy images to observed galaxy images.

References

[1] Abraham, R. G. and van den Bergh, S. (2001). The morphological evolution of galaxies. *Science*, **293**(5533), 1273–1278.

[2] Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(21), 7426–7431.

[3] Coifman, R. R., Lafon, S. (2006). Diffusion Maps. *Applied and Computational Harmonic Analysis*, **21**, 5–30.

[4] Conselice, C. J. (2003). The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, **147**, 1–28.

[5] Duong, T. (2013). Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, **25**, 635–645.

[6] Ernst, M. D. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, **19**(4), 676–685.

[7] Efron, B., and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

[8] Freeman, P. E., Newman, J. A., Lee, A. B., Richards, J. W. and Schafer, C. M. (2009). Photometric redshift estimation using spectral connectivity analysis. *Monthly Notices of the Royal Astronomical Society*, **398**, 2012–2021.

[9] Freeman, P. E., Izbicki, R., Lee, A. B., Newman, J. A., Conselice, C. J., Koekemoer, A. M., Lotz, J. M. and Mozena, M. (2013). New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, **434**, 282–295.

[10] Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, **13**, 723–773.

[11] Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K. and Sriperumbudur, B. (2012b). Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 1205–1213.

[12] Inselberg, A. (1997). Multidimensional detective. *Information Visualization, 1997. Proceedings., IEEE Symposium on*, 100–107.

[13] 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.

[14] Izbicki, R., Lee, A. B. and Schafer, C. M. (2015). Nonparametric Conditional Density Estimation in a High-Dimensional Regression Setting. *Journal of Computational and Graphical Statistics*, Accepted.

[15] Izbicki, R., Lee, A. B. and Freeman, P. E. (2016). Photo-z estimation: an example of nonparametric condtional density estimation under selection bias. URL https://arxiv.org/pdf/1604.01339v1.pdf

[16] Kanamori. T., Suzuki, T. and Sugiyama, M. (2012). f-Divergence Estimation and Two-Sample Homogeneity Test Under Semiparametric Density-Ratio Models. *IEEE Transactions on Information Theory*. **58**(2), 708–720.

[17] Kpotufe, S. (2011), k-NN regression adapts to local intrinsic dimension, *In Advances in Neural Information Processing Systems*, **24**, 729–737.

[18] Lafon, S. and Lee, A. B. (2004). Diffusion maps and coarse-graining : a unified framework for dimensionality reduction, graph partitioning and data set parameterization. *IEEE Transactions on pattern analysis and machine intelligence*, **28**, 1393–1403.

[19] Lotz, J. M., Primack, J. and Madau, P. (2004). A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, **128**, 163–182.

[20] Murtagh, F., and Legendre, P. (2014). Wards hierarchical agglomerative clustering method: which algorithms implement Wards criterion?. *Journal of Classification*, **31**(3), 274–295.

[21] Peth, M. A., Lotz, J. M., Freeman, P. E., McPartland, C., Mortazavi, S. A., Snyder, G. F., Barro, G., Grogin, N. A., Guo, Y., Hemmati, Y., Kartaltepe, J. S., Kocevski, D. D., Koekemoer, A. M., McIntosh, D. H., Nayyeri, H., Papovich, C., Primack, J. R., and Simons, R. C. (2016). Beyond Spheroids and Discs: Classifications of CANDELS Galaxy Structure at 1.4 ¡ z ¡ 2 via Principal Component Analysis. *Monthly Notices of the Royal Astronomical Society*, **458**, 963-987.

[22] Phipson, B. and Smyth, G. K. (2011). Permutation p-values should never be zero. *SAGMB*, **9**(1).

[23] Ramdas, A., Reddi, S. J., Poczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

[24] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, **41**(5), 2263–2291.

[25] Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H. and Okanohara, D. (2010). Conditional density estimation via least-squares density ratio estimation. *In International Conference on Artificial Intelligence and Statistics*, 781–788.

[26] Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011a). Least-Squares Two-Sample Test. *Neural Networks*, **24**, 735–751.

[27] Sugiyama, M., Yamada, M., von Bunau, P., Suzuki, T., Kanamori, T. and Kawanabe, M. (2011b). Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, **24**(2), 183–198.

[28] Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, **5**, 1–6.

[29] Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, **18**(4), 267–276.

[30] Wager, S., Hastie, T., and Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, **15**, 2014.

[31] Wager, S. and Athey, S. (2015). Estimation and inference of heterogeneous treatment effects using random forests. arXiv preprint arXiv:1510.04342.

[32] Zelnik-Manor, L. and Perona, P. (2005). Self-tuning spectral clustering. *Advances in Neural Information Processing system*, **17**, 1601–1608.

A Exploratory Data Analysis

Since the galaxy mass is continuous variable, we can visualize the galaxy mass as a function of morphologies via regression. Here, we use the local linear regression with 95% confidence interval.

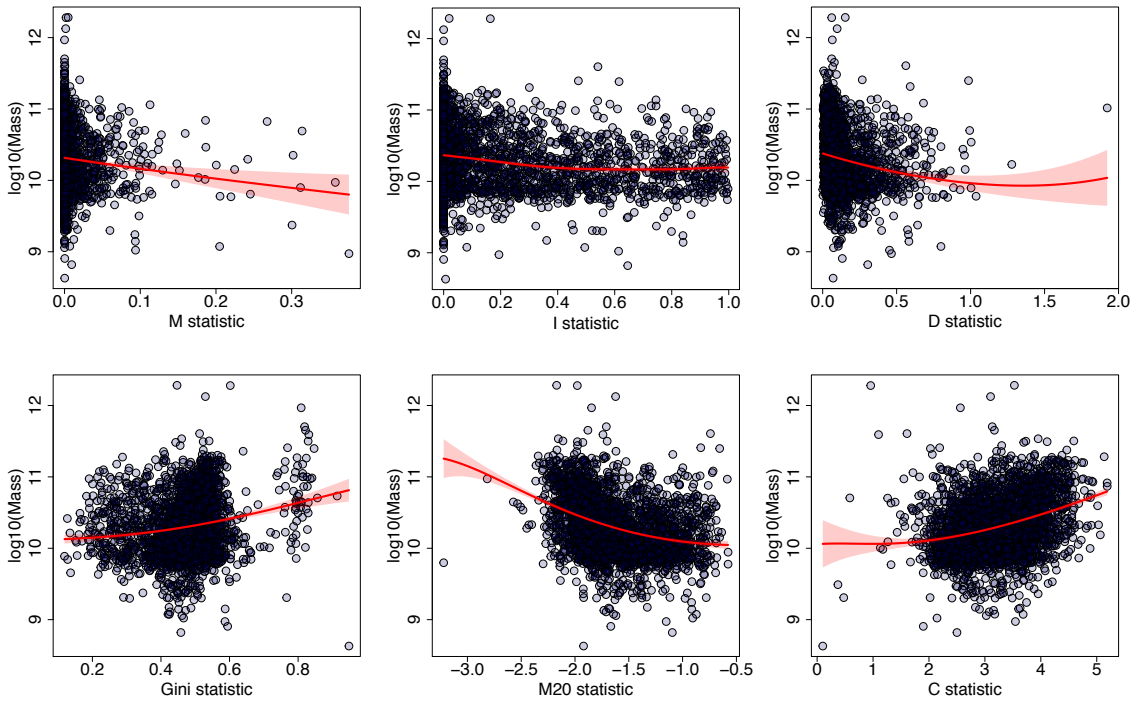


Figure 29: Regression analysis of the galaxy mass: the regression function is estimated via the local linear regression with 95% confidence intervals.

The following figures provide the information of how the individual summary statistic as well as the physical properties are distributed in diffusion space of morphology statistics.

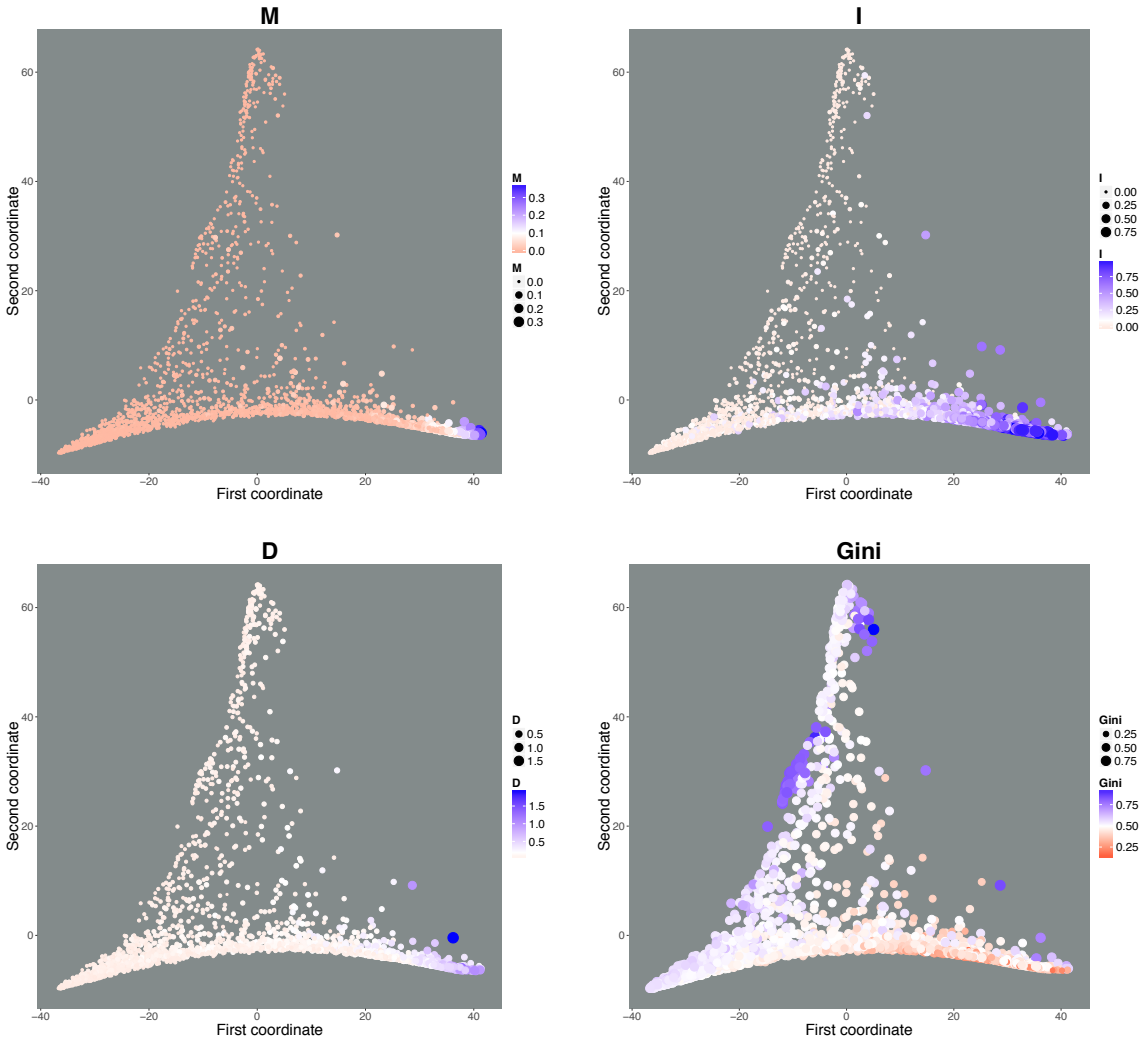


Figure 30: Diffusion map of summary statistics.

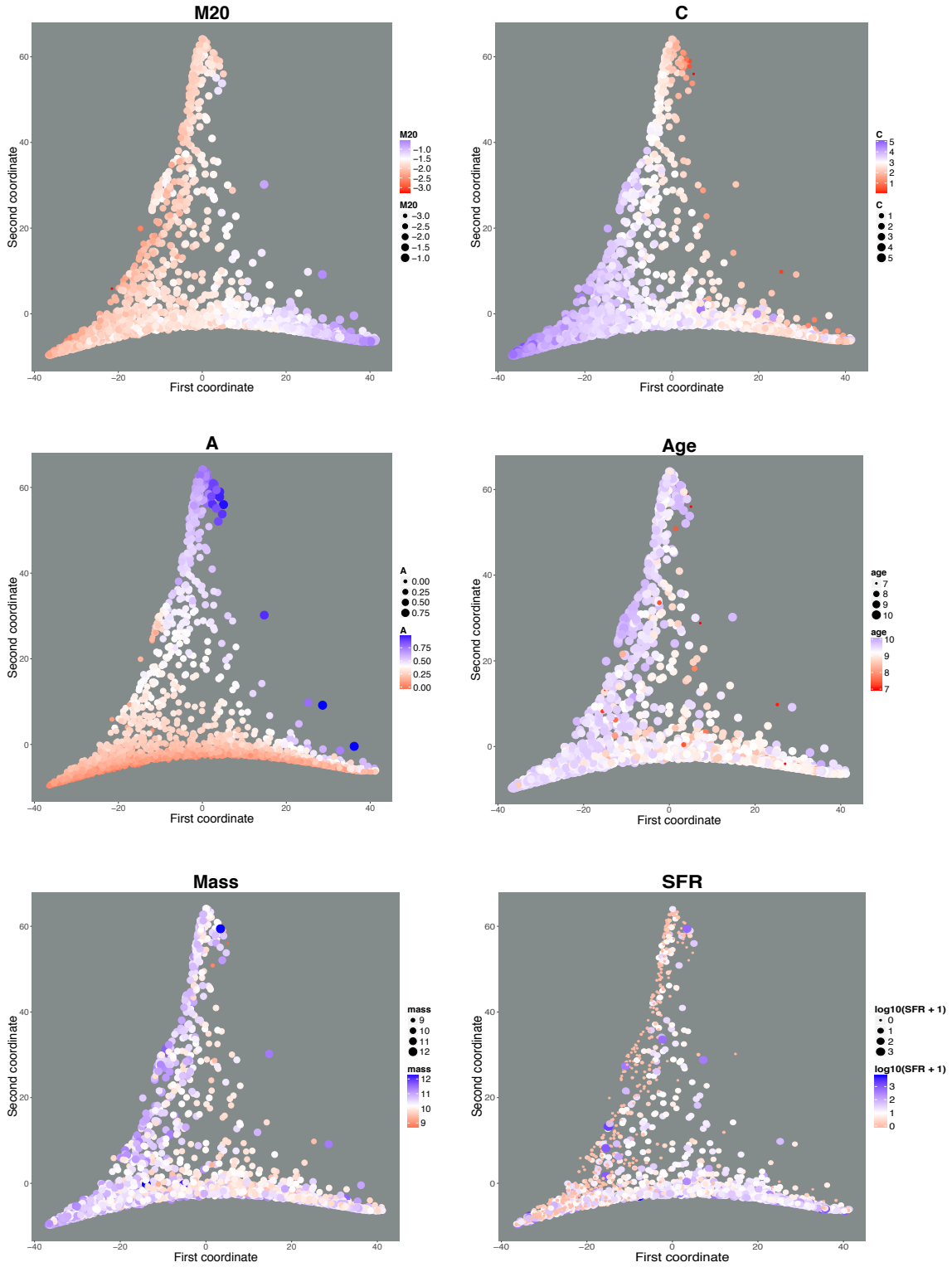


Figure 31: Diffusion map of summary statistics and physical properties.

A multivariate two-sample test can have higher statistical power to detect differences between two populations especially when their covariates are correlated to each other. (see Figure 32.) Figure 33 provides the Pearson's correlation coefficient plot between seven-dimensional morphology statistics. It shows that some of the statistics are stronlgy correlated to each other. Thus, we might be able to detect differences which cannot be found from an univariate test by conducting a test in the multi-dimensional space.

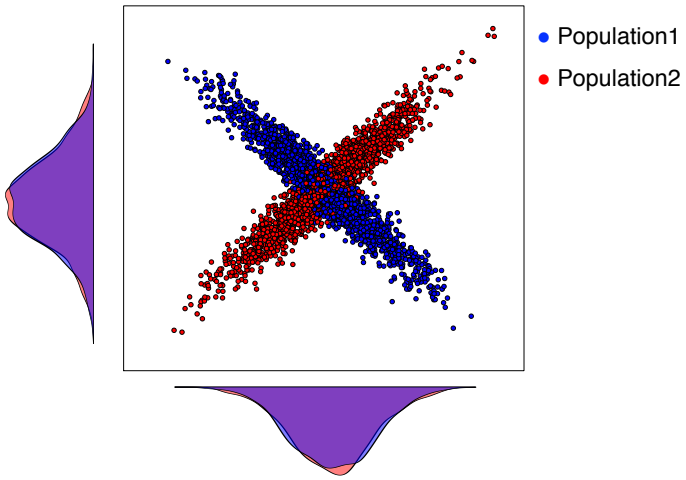


Figure 32: Toy example showing two populations that have the same marginal distributions; any univariate two-sample test has no power to detect the difference between these marginals.

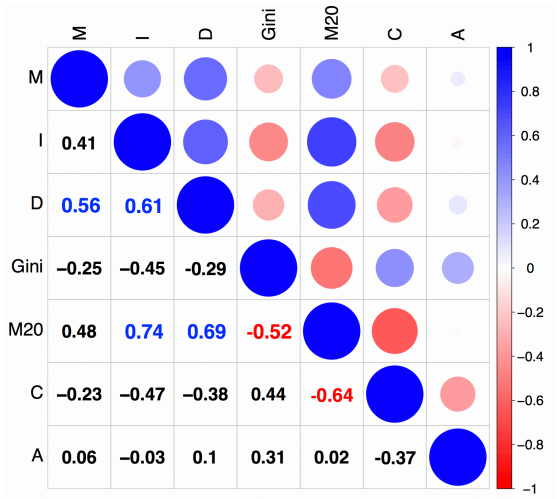


Figure 33: Correlation between summary statistics. Some of the statistics are correlated to each other.

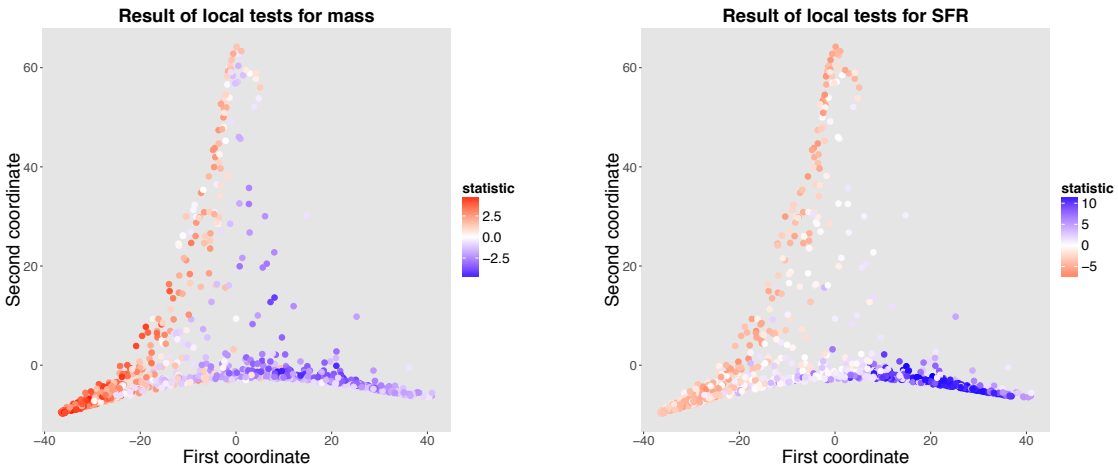


Figure 34: Distribution of test statistic $T_n(x) = \widehat{\mathbb{P}}(Y = 1|x) - \widehat{\mathbb{P}}(Y = 1)$ *in the diffusion map.*

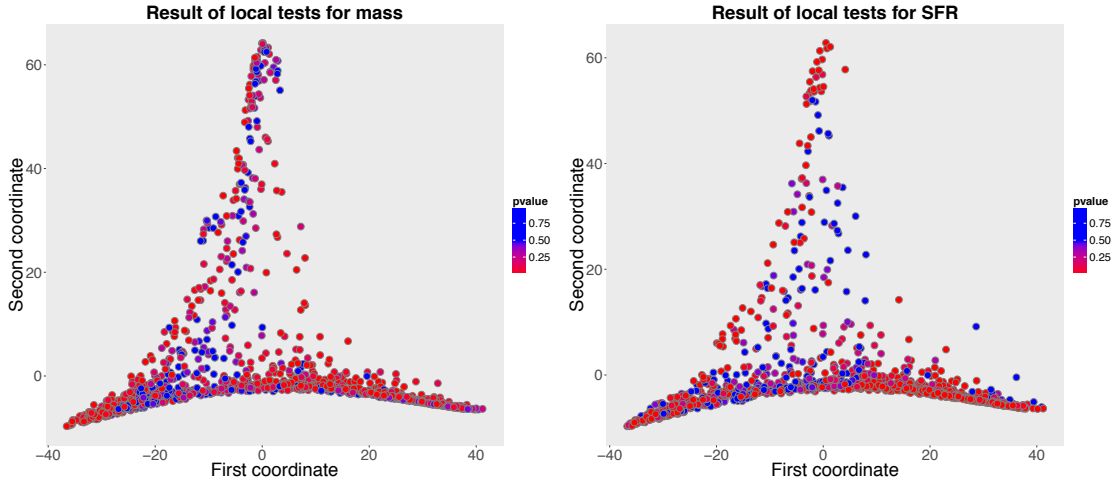


Figure 35: Distribution of p-values in the diffusion map.

B The change of cutoff value

By changing the cutoff value c for the division of two physical groups, we found that the main conclusion of the study remains the same. For example, high-mass galaxies are more likely to be compact and concentrated whereas low-mass galaxies tend to be more irregular. In fact, there is a trade-off between large and small values of c: A too large c leads to ambiguity of defining the physical group and thus less significant results. Whereas, a too small c would not allow us to have reasoable statistical power to find a difference due to the lack of the sample size.

The following table shows that 10% cutoff gives us more significant points but still there exist some points that are significant for 25% cutoff but not significant for 10% cutoff. We can notice that there is no point within the cells: [High-Mass(10%), Low-Mass(25%)] and [Low-Mass(10%), High-Mass(25%)], which indicates the robustness of the result depending on the choice of the cutoff value.

Table 36: Significant points by changing the cutoff value

		25% cutoff			
		High-Mass	Low-Mass	Insignificance	Total
10% cutoff	High-Mass	219	0	38	257
	Low-Mass	0	236	118	354
	Insignificance	0	0	389	389
	Total	219	236	545	1,000

C Summary Statistics

1. **Multimode** (M) statistic (Freeman et al. 2013)
   The M statistic identifies galaxies with disturbed morphologies. Consider an intensity quantile q_l such that a proportion l of the pixel intensities i_{mn} are smaller than q_l. (Here mn denotes pixel coordinates.) For a given q_l, define a new indicator variable j_{mn} such that

$$j_{mn} = \begin{cases} 1 & \text{if } i_{mn} \geq q_l \\ 0 & \text{otherwise} \end{cases} .$$

   Within the image j_{mn}, we obtain the areas of largest and second-largest groups of contiguous pixels, which we denote $A_{l,(1)}$ and $A_{l,(2)}$ respectively. We define the area ratio as

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

   where n_{seg} is the number of pixels in the segmentation map, i.e., the mask used to define the extent of the galaxy within the image. This formulation imposes a strict upper limit on R_1 of $1/2$ that is achieved if $A_{l,(1)} = A_{l,(2)} = n_{seg}/2$. The M statistic is the maximum observed value of R_l over all quantiles l:

$$M = \max_l R_l .$$

2. **Intensity** (I) statistic (Freeman et al. 2013)
   One of the shortcomings of the M statistic is that it does not consider the summed intensity within contiguous pixel groups. For instance, a contiguous group with a large number of pixels may have a smaller summed intensity than other, smaller groups of pixels. To mitigate this shortcoming we utilize the I statistic. We associate each pixel mn with a local maximum $mn_{\max}$ by following the maximum gradient ascent path. All pixels that are associated with a given local maximum $mn_{\max}$ are grouped together, and for each group, we sum the pixel intensities i_{mn}. (Note that in a data pre-processing step, we smooth the image data with a symmetric Gaussian kernel with $\sigma \sim 1$ pixel, to decrease the effect that pixel noise has on the construction of pixel groups.) We rank the summed intensities in descending order and use the first and second sorted values to compute the I statistic:

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

3. **Deviation** (D) statistic (Freeman et al. 2013)
   The deviation D statistic is used to capture evidence of galaxy asymmetry. It is the distance from the local maximum associated with $I_{(1)}$ to the galaxy's center of mass:

$$(m_{\rm cen}, n_{\rm cen}) = \left(\frac{1}{n_{seg}} \sum_m \sum_n m i_{mn}, \frac{1}{n_{seg}} \sum_m \sum_n n i_{mn} \right) , \tag{7}$$

   where the summation is over the n_{seg} pixels within the segmentation map. The D statistic is:

$$D = \sqrt{(m_{\rm cen} - m_{I_{(1)}})^2 + (n_{\rm cen} - n_{I_{(1)}})^2} / \sqrt{n_{seg}/\pi} \, .$$

   where the normalizing factor $\sqrt{n_{seg}/\pi}$ is a galaxy radius estimate achieved by assuming that the segmentation map is circular.

4. **Gini** (G) statistic (Lotz et al. 2004)
   The Gini coefficient measures the relative distribution of pixel intensities within the segmentation map: $G = 0$ means that the intensities are uniform across the galaxy, while $G = 1$ means that all of a galaxy's light falls into a single pixel. The G statistic is defined as

$$G = \frac{1}{\bar{i} n_{\rm seg}(n_{\rm seg} - 1)} \sum_k (2k - n_{\rm seg} - 1) i_{m_{(k)} n_{(k)}} \, ,$$

   where $\bar{i}$ is the sample mean of all intensities within the segmentation map and $m_{(k)} n_{(k)}$ denotes the coordinates of the pixel with the $k^{\rm th}$-smallest intensity value.

5. $\mathbf{M_{20}}$ statistic (Lotz et al. 2004)
   M_{20} describes the spatial distribution of pixel intensities. First, we compute a total second-order moment:

$$M_{\rm tot} = \sum_m \sum_n i_{mn} \left[(m - m_{\rm cen})^2 + (n - n_{\rm cen})^2 \right]$$

   where $m_{\rm cen}$ and $n_{\rm cen}$ are the coordinates of the galaxy's center of mass (equation 7) and the summation in done over all pixels mn within the segmentation map. We then repeat the summation done above using only the brightest 20% of the pixels; we call this sum $M_{\rm bright}$. Then M_{20} is

$$M_{20} = \log_{10} \left(\frac{M_{\rm bright}}{M_{\rm tot}} \right) \, .$$

6. **Concentration** (C) statistic (Conselice 2003)
   The concentration statistic encapsulates the area over which the bulk of a galaxy's summed intensity lies. Its calculation assumes circular symmetry. At a given radius r from the galaxy's center, we define two quantities: the summed intensity within the annulus defined by r and $r + dr$, and the overall average summed intensity:

$$\begin{aligned} \mu(r) &= \frac{\int_0^{2\pi} \int_{r-\delta r}^{r+\delta r} i(r', \theta) r' dr' d\theta}{\int_0^{2\pi} \int_{r-\delta r}^{r+\delta r} r' dr' d\theta} \\ \bar{\mu}(r) &= \frac{\int_0^{2\pi} \int_0^{r+\delta r} i(r', \theta) r' dr' d\theta}{\int_0^{2\pi} \int_0^{r+\delta r} r' dr' d\theta}. \end{aligned}$$

   (We show the calculations as integrals for conceptual clarity, but the actual calculations are done as sums over image pixels.) r is the solution of the equation $\mu(r)/\bar{\mu}(r) = \epsilon$, where ϵ is commonly chosen to be 0.2. We compute the total summed intensity within the radius r, then determine the smaller radii within which there are 20% and 80% of that total summed intensity. The C statistic is:

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

   The smaller $r_{20\%}$ is relative to $r_{80\%}$, the higher the value of C, as the galaxy will appear "more concentrated."

7. **Asymmetry** (A) statistic (Conselice 2003)
   The A statistic is a measure of how asymmetric a galaxy is after its image is rotated 180° the central pixel and then subtracted from the original image. For an asymmetric galaxy, the difference image will exhibit significant residual structures, leading the A statistic to differ significantly from zero. The A statistic is defined as

$$A = \frac{\sum_m \sum_n |i_{mn} - i_{180,mn}|}{\sum_m \sum_n |i_{mn}|} - B_{180},$$

   where i and i_{180} are the pixel intensities in the original and rotated images respectively and B_{180} is the average background asymmetry, defined using the intensities of pixels lying outside the segmentation map.

D Tuning of Parameters

For choosing tuning parameters, we consider the following loss function.

$$\begin{aligned} L(\widehat{\gamma}_h, \gamma) &= \int \left(\widehat{\gamma}_h(x) - \gamma(x)\right)^2 dP(x) \\ &= \int \left(\widehat{\mathbb{P}}_h(Y=1|x) - \mathbb{P}(Y=1|x)\right)^2 dP(x) + \left[\widehat{\mathbb{P}}(Y=1) - \mathbb{P}(Y=1)\right]^2 \\ &\quad - 2\left[\widehat{\mathbb{P}}(Y=1) - \mathbb{P}(Y=1)\right] \int \left(\widehat{\mathbb{P}}_h(Y=1|x) - \mathbb{P}(Y=1|x)\right) P(x) \end{aligned}$$

where $\widehat{\gamma}_h = \widehat{\mathbb{P}}(Y=1|x) - \widehat{\mathbb{P}}(Y=1)$ and $\gamma = \mathbb{P}(Y=1|x) - \mathbb{P}(Y=1)$. Using a typical estimator of $\mathbb{P}(Y=1)$ obtained by $\widehat{\mathbb{P}}(Y=1) = n_1/n$ where $n = n_0 + n_1$, the loss function can be expanded into

$$\begin{aligned} L(\widehat{\gamma}_h, \gamma) = \int \widehat{\mathbb{P}}_h^2(Y=1|x) dP(x) - 2\int \widehat{\mathbb{P}}_h(Y=1|x)\mathbb{P}(Y=1|x) dP(x) \\ - 2\left[\widehat{\mathbb{P}}(x) - \mathbb{P}(x)\right] \int \widehat{\mathbb{P}}_h(Y=1|x) dP(x) + K \end{aligned}$$

where K is a constant that does not depend on the tuning parameters. Since

$$\widehat{\mathbb{P}}(Y=1) - \mathbb{P}(Y=1) = o_p(1) \quad \text{and} \quad \int \widehat{\mathbb{P}}_h(Y=1|x) dP(x) \in [0,1]\,,$$

we estimate the loss function up to K based on cross-validation by

$$\widehat{L}(\widehat{\gamma}_h, \gamma) = \frac{1}{n'} \sum_{i=1}^{n'} \widehat{\mathbb{P}}_h^2(Y=1|x_i) - \frac{2}{n'} \sum_{j=1}^{n_1'} \widehat{\mathbb{P}}_h(Y=1|x_j) \tag{8}$$

where $n' = n_0' + n_1'$ is the number of a validation sample. Then, we choose the tuning parameters that minimize the estimated loss $\widehat{L}(\widehat{\gamma}_h, \gamma)$.

Let c_x be a critical value of the local test at x that satisfies

$$\mathbb{P}_{H_0}(|\widehat{\gamma}_h(x) - \gamma(x)| > c_x|x) \leq \alpha \tag{9}$$

where α is the level of significance. In general, the statistical power of a hypothesis test is closely related to the precision of confidence interval. So, we want to select the smallest c_x under the restriction of (9). By Chebyshev's inequality, we have

$$\mathbb{P}_{H_0}(|\widehat{\gamma}_h(x) - \gamma(x)| > c_x|x) \leq \frac{1}{c_x^2} \mathbb{E}\left[\left(\widehat{\gamma}_h(x) - \gamma(x)\right)^2 |x\right].$$

Therefore, choosing h with the smallest $\mathbb{E}\left[\left(\widehat{\gamma}_h(x)-\gamma(x)\right)^2|x\right]$ is an indirect way of selecting c_x as small as possible. Moreover, if we use the common h over x to reduce the computation time, the averaged mean squared error over x can be used as a surrogate loss function to optimize the power. Specifically, the following holds.

$$\begin{aligned}\mathbb{E}_x\left[\mathbb{P}_{H_0}(|\widehat{\gamma}_h(x)-\gamma(x)|>c_x|x)\right] &\leq \mathbb{E}_x\left[\frac{1}{c_x}\mathbb{E}\left[|\widehat{\gamma}_h(x)-\gamma(x)||x\right]\right] \\ &\leq \sqrt{\mathbb{E}\left[c_x^{-2}\right]}\sqrt{\mathbb{E}\left[L(\widehat{\gamma}_h,\gamma)\right]}.\end{aligned}$$

This fact motivates us to employ $\widehat{L}(\widehat{\gamma}_h,\gamma)$ as a loss function to choose the tuning parameters.