Local two-sample testing: a new tool for analysing high-dimensional astronomical data

P. E. Freeman,† I. Kim and A. B. Lee

Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Accepted 2017 July 14. Received 2017 July 14; in original form 2017 June 2

ABSTRACT
Modern surveys have provided the astronomical community with a flood of high-dimensional data, but analyses of these data often occur after their projection to lower dimensional spaces. In this work, we introduce a local two-sample hypothesis test framework that an analyst may directly apply to data in their native space. In this framework, the analyst defines two classes based on a response variable of interest (e.g. higher mass galaxies versus lower mass galaxies) and determines at arbitrary points in predictor space whether the local proportions of objects that belong to the two classes significantly differ from the global proportion. Our framework has a potential myriad of uses throughout astronomy; here, we demonstrate its efficacy by applying it to a sample of 2487 $i$-band-selected galaxies observed by the HST-ACS in four of the CANDELS programme fields. For each galaxy, we have seven morphological summary statistics along with an estimated stellar mass and star formation rate (SFR). We perform two studies: one in which we determine regions of the seven-dimensional space of morphological statistics where high-mass galaxies are significantly more numerous than low-mass galaxies, and vice versa, and another study where we use SFR in place of mass. We find that we are able to identify such regions, and show how high-mass/low-SFR regions are associated with concentrated and undisturbed galaxies, while galaxies in low-mass/high-SFR regions appear more extended and/or disturbed than their high-mass/low-SFR counterparts.

Key words: methods: data analysis – methods: statistical – galaxies: evolution – galaxies: high-redshift – galaxies: statistics – galaxies: structure.

1 INTRODUCTION
Modern astronomical data are intrinsically high dimensional; for any given object, we may have images and photometric magnitudes (and perhaps spectra), as well as estimates of mass, star formation rate (SFR), metallicity, etc. Astronomical data analysis, however, still often operates in low dimension, due less to a lack of will than to a lack of tools that astronomers can wield to effectively analyse data in their native spaces.

One specific area in which the ability to work with high-dimensional data is useful is in the analysis of galaxy morphology. Morphological studies are key to understanding the evolutionary histories of galaxies and to constraining theories of hierarchical structure formation. (For a recent review, see e.g. Conselice 2014.) A galaxy’s morphology indicates its current state: Is it undergoing a merger? Is it compact and quiescent? It may also contain information about its assembly history: Is it undergoing a post-merger burst of star formation? Does its central bulge indicate past mergers? We may think of a galaxy’s morphology as a continuous surface brightness function observed in three dimensions (two spatial, one wavelength) that is sampled from a distribution of such functions. Due to finite resolution, what we actually observe is a pixelated and discretized version of the sampled function. Discretization helps us by moving morphological analysis from the realm of infinite dimensionality1 to the finite realm, but the dimensionality (i.e. the number of image pixels times the number of wavelengths at which image data are collected) is still very large, and thus analyses are still subject to the ‘curse of dimensionality’. To make analyses tractable, one conventionally reduces the dimensionality further by, for example, computing summary statistics, which may be either parametric (e.g. the S´ersic index $n$; S´ersic 1963) or non-parametric (e.g. the Gini coefficient $G$; Abraham, van den Bergh & Nair 2003; Lotz, Primack & Madau 2004).

Let $T$ represent a collection of morphological statistics. We may model an ensemble of galaxies by a sample from a distribution of moderate or high dimensionality:

$$f(T; z, \lambda_{\text{obs}}; M_*, S, \ldots | \theta),$$

(1)

1 In practice, we would need an infinite number of parameters to fully model surface brightness; S´ersic profiles (S´ersic 1963), for instance, are insufficient.
where $z$ and $\lambda_{\text{obs}}$ are the redshift and observed wavelength, $M_*$ and $S$ are the galaxy stellar mass and SFR, and $\theta$ collectively represents the cosmological and astrophysical parameters that govern structure formation. A goal that is potentially realizable in the near future is to statistically infer $\theta$, by comparing samples from estimates $\hat{f}$ derived from astronomical surveys with samples from simulation models of $f(T; z, \lambda_{\text{obs}}; M_*, S, \ldots | \theta)$ (e.g. from the Illustris and Eagle projects; Vogelsberger et al. 2014, Schaye et al. 2015). For example, likelihood-free methods, such as the Approximate Bayesian Computation (ABC, e.g. Weyant, Schafer & Wood-Vasey 2013, and references therein), currently rely on comparing a few derived summary statistics instead of comparing two samples directly in higher dimensions. In the meantime, many authors attempt to infer the relationship between parametric and/or non-parametric structure statistics and other statistics of interest: $M_*$ and $S$, and, in particular, the ‘main sequence’ in the $M_* - S$ diagram (e.g. Elbaz et al. 2011; Wuyts et al. 2011; Salmi et al. 2012; Barro et al. 2014; Brennan et al. 2017; see also Snyder et al. 2015 for a similar analysis of simulated Illustris galaxies); the fraction of quenched galaxies (e.g. Bluck et al. 2014; Lang et al. 2014; Woo et al. 2015; Bluck et al.; Peth et al. 2016; Teimoorinia, Bluck & Ellison 2016; see also Huertas-Company et al. 2016); rest-frame colour (e.g. Wake, van Dokkum & Franx 2012); and local environment (e.g. Lackner & Gunn 2013).\(^2\) (See also Bell et al. 2012, who compare morphological statistics of star-forming and quiescent galaxies over cosmic time, and Fang et al. 2015, who apply unsupervised learning methods to structure statistics.) However, save one exception that we mention below, all of these authors work with one or two morphological statistics at a time instead of working with an entire ensemble (which may include the effective radius, the axis ratio and the Sérsic index, in addition to statistics introduced via bulge–disc decompositions, the Gini and $M_{20}$ statistics, etc.).\(^3\) By concentrating their efforts on projections of the ensemble rather than the full ensemble itself, the authors cannot truly map out dependences between variables.

In this work, we present a new statistical framework that utilizes local two-sample hypothesis tests. Astronomers will find this framework useful for detecting and quantifying dependences within statistical spaces of moderate or high dimension. In particular, local two-sample tests can identify objects that lie in regions of predictor space where the estimated proportion of a particular defined class of objects (e.g. galaxies of high mass or of low metallicity, etc.) differs significantly from the global proportion. There are a myriad of applications for this framework; we demonstrate it by exploring the relationship between non-parametric structure statistics — namely the seven image summary statistics $M$, $I$, $D$ (Freeman et al. 2013), $G$, $M_{20}$ (Abraham et al. 1994, 1996a,b; Conselice 2003) — and estimated stellar mass $M_*$ and SFR $S$. We note that our work has superficial similarity to that done by Peth et al. (2016), who study the relationship between the same seven image statistics and stellar quenching. However, their work utilizes clustering — the authors first determine principal components for the seven statistics, then use agglomerative hierarchical clustering to identify 10 galaxy groups (plus outliers) within principal component space. We, on the other hand, divide individual galaxies into groups based on defined response variables (estimated stellar mass, etc.), and then identify locally significant differences between the two populations without pre-clustering the data.

In Section 2, we outline our local two-sample hypothesis test framework.\(^4\) (For more details, see Kim & Lee, in preparation.) In Section 3, we demonstrate its efficacy by applying it to a sample of 2487 $i$-band-selected (rest-frame wavelength $\approx 4,500$ Å) galaxies observed by the Hubble Space Telescope’s (HST) Advanced Camera for Surveys (ACS). We perform two studies: a morphology–mass study, where we identify galaxies that lie in predominantly high-mass or low-mass regions (or neither), and the other where the division into two samples is based on SFR rather than mass. In Section 4, we summarize our findings.

2 LOCAL TWO-SAMPLE HYPOTHESIS TESTING

Suppose that we are given a set of $n$ measurements of an astronomical object, and that our interest lies in determining those regions of the $n$-dimensional space of statistics where objects of particular class, class $y_0$ (perhaps defined as belonging to a particular redshift bin, or a particular range of masses, etc.), are observed in greater or lesser proportions than the class’s global proportion. One way to identify these regions is via the use of two-sample, or homogeneity, tests. Let $P_0^y$ and $P_1^y$ represent distributions from which the independent samples $x_{0,1}^y, \ldots, x_{n,1}^y$ and $x_{1,1}^y, \ldots, x_{1,n}^y$ are drawn, respectively.

In order to define regions where class proportions differ from their global values, one needs to utilize a local two-sample test. In a global two-sample test, we would compare the hypotheses

\[
H_0 : P_0^y = P_1^y \quad \text{against} \quad H_1 : P_0^y \neq P_1^y.
\]

Examples of such tests include the maximum mean discrepancy test (Gretton et al. 2012) and the energy distance test (Székely & Rizzo 2004), both being non-parametric extensions of classical $t$ tests. However, such tests can provide only binary results: ‘reject the null hypothesis’ or ‘fail to reject the null hypothesis’. In the current context, rejection of the null hypothesis is typically uninteresting; what is interesting is determining where in the space of galaxy morphological statistics the distributions $P_0^y$ and $P_1^y$ significantly differ, hence the need for local testing.

In the flow cytometry literature, Roederer & Hardy (2001) address the problem of detecting differences between two samples in a multidimensional space. Their method partitions the space into hypercubes, and identifies those hypercubes where $P_0^y \neq P_1^y$. To capture detailed local structures, it is natural to shrink the volume of each hypercube as the overall sample size increases, eventually approaching a point-wise test in the limit of large sample sizes. We thus propose a point-wise tests for differences at specified points $(x_1, \ldots, x_p)$:

\[
H_{0,0} : \mathbb{P}(Y = y | X = x_i) = \mathbb{P}(Y = y) \quad \text{against} \quad H_{1,0} : \mathbb{P}(Y = y | X = x_i) \neq \mathbb{P}(Y = y).
\]

Such a test is equivalent to testing for differences in density at the specified points. Unlike Duong (2013), who uses kernel density estimation to find locally significant differences between two

---

\(^2\) Note that such inference stands in contrast to using structure statistics to predict classification; for example, Simmons et al. (2017), and references therein.

\(^3\) Teimoorinia et al. (2016) may also be considered an exception, in that they apply an artificial-neural-network algorithm that relates eight statistics to quenching fraction, but ultimately their interest lies in determining which subsets of two or three statistics have the greatest predictive power.

\(^4\) The interested reader may find $x$ functions for carrying out local two-sample tests at github.com/pefreeman/lst.
local two-sample testing via asymptotic normality

Algorithm 1: Local two-sample testing via asymptotic normality

**Input:** iid training samples \( \{x_i, 0\}_i^{n_1} \) and \( \{x_i, 1\}_i^{n_2} \).

Defined test points \( \{x_1, \ldots, x_t\} \).

1. **Step 1.** Fit random forest regression to training samples \( \{x_i, 0\}_i^{n_1} \) and \( \{x_i, 1\}_i^{n_2} \).

2. **Step 2.** Calculate the test statistic \( T(x_i) \) at the \( t \) test points.

3. **Step 3.** Compute the \( p \)-value at each test point: \( p_i = 2\Phi(-|T(x_i)|) \), where \( \Phi(\cdot) \) is the cumulative distribution function for the standard normal distribution.

4. **Step 4.** Apply the Benjamini–Hochberg procedure to correct for the number of tests (Benjamini & Hochberg 1995). Let \( \alpha \) be the false discovery rate (FDR) control parameter, typically \( 0.05 \) but may be adjusted. Let \( \alpha_{\text{FDR}} \) be the identification of one of the two discrete classes (e.g. high mass, in a high-mass and low-mass comparison). For each datum, conclude

\[
\begin{align*}
(1) & \quad \mathbb{P}(Y = y|x_i) > \mathbb{P}(Y = 1) \text{ if } T(x_i) > 0 \text{ and the adjusted } p \text{-value is } p_i < \alpha_{\text{FDR}}; \\
(2) & \quad \mathbb{P}(Y = y|x_i) < \mathbb{P}(Y = 1) \text{ if } T(x_i) < 0 \text{ and the adjusted } p \text{-value is } p_i < \alpha_{\text{FDR}}. \\
\end{align*}
\]

(The threshold \( p \)-value for rejecting the null hypothesis, \( \alpha_{\text{FDR}} \), is typically 0.05 but may be adjusted.)

For all other points, conclude \( \mathbb{P}(Y = y|x_i) = \mathbb{P}(Y = y) \).

\[ T(x_i) = \frac{\hat{\mathbb{P}}(Y = 1|x_i) - \hat{\mathbb{P}}(Y = 1)}{\sqrt{\hat{V}(x_i)}}, \]

which under the null hypothesis converges to a standard normal distribution. \( \hat{V}(x) \) is a consistent estimator of the variance of the random forest predictions based on the infinitesimal jackknife (Wager, Hastie & Efron 2014).

Algorithm 1 shows the steps that we follow in our analyses of galaxy data. Because the dimensionality of the predictor space precludes us from defining a dense rectangular grid of points at which to run local two-sample tests, we split our galaxy data into training and test sets. We use the former to grow the forest, and we compute two-sample test \( p \)-values using the latter. Given those \( p \)-values and a significance criterion that is adjusted via the Benjamini–Hochberg false discovery rate algorithm with \( \alpha = 0.05 \) (Benjamini & Hochberg 1995), we determine whether each point lies in a region where the proportion of class \( y \) galaxies is consistent with, or significantly different from, the global proportion. (We choose \( \alpha = 0.05 \) as this is the standard value in statistical analyses, but note that one can adjust this value as necessary.)

### 3 APPLICATION TO GALAXY DATA

#### 3.1 Data

We demonstrate the efficacy of our framework by applying it to the analysis of HST-ACS i-band images from four fields observed as part of the Cosmic Assembly Near-IR Deep Extragalactic Legacy Survey (CANDELS; Grogin et al. 2011; Koekemoer et al. 2011).

To construct our data sample, we begin by defining a range of redshifts such that rest-frame 4500 Å is observed within the HST-ACS.
Figure 2. Distributions of masses (top panel) and SFRs (bottom panel) for the 1787 galaxies in the training set, with the lower and upper quartiles highlighted. Note that the bottom histogram does not include 79 galaxies for which \( \hat{S} = 0 \) (but which are included in the low-SFR group).

F814W filter (i.e. within the \( i \) band). We adopt the knee at \( \approx 9550 \) Å in the filter transmission curve as our upper wavelength bound, with matching lower bound at \( \approx 7020 \) Å; thus, \( z \in [(7020 - 4500)/4500, (9550 - 4500)/4500] = [0.560, 1.122] \). Next, we apply magnitude, mass and redshift cuts to the full four-field CANDELS galaxy sample. The CANDELS Team Release mass catalogues include estimates of \( H \)-band magnitudes and spectroscopic redshifts (if available) for each galaxy, as well as the median of a number of stellar mass estimates (\( \hat{M}_{\text{med}} \); see Mobasher et al. 2015 and Santini et al. 2015). For those galaxies lacking spectroscopic redshifts, we utilize a photometric redshift conditional density estimate \( \hat{p}(z) \) made using a hierarchical Bayesian technique that combines the output of five separate photometric redshift estimators summarized in Dahlen et al. (2013) (D. Kodra & J. Newman, private communication). To be included in our sample, a galaxy must have

(i) an \( H \)-band magnitude \( H < 25 \);
(ii) a spectroscopic redshift \( z_{\text{spec}} \in [0.560, 1.122] \), or, if the spectroscopic redshift is not available, an integrated conditional density estimate \( \int_{0.560}^{1.122} \hat{p}(z)dz \) that is \( \geq 0.8 \); and
(iii) a mass \( \hat{M}_{\text{med}} \geq 10^{9.84} M_\odot \).

We assume a mass threshold of \( 10^{10} M_\odot \) at \( z = 0 \) and use the algorithm of Behroozi, Wechsler & Conroy (2013) to adjust it downwards as \( z \) increases so as to maintain an approximately constant comoving galaxy number density (see Fig. 1). The average of the threshold curve over the range \( z = [0.560, 1.122] \) is \( 10^{9.84} M_\odot \). (We use the average of the curve rather than the curve itself so as to ensure that the distributions of redshifts in the low-mass and high-mass quartile samples that we analyse below are similar.)

Our final data sample consists of 2487 galaxies, of which 891 have measured spectroscopic redshifts (see Table 1). Image summary statistics – namely \( M, I, D \) (Freeman et al. 2013), \( G, M_{20} \) (Abraham, van den Bergh & Nair 2003; Lotz et al. 2004), and \( C, A \) (Abraham et al. 1994, 1996a,b; Conselice 2003) – are determined for each galaxy using our own R software suite.\(^5\) (Note that our current definition of the multimode statistic \( M \) differs slightly from that of Freeman et al. 2013: We divide the variable \( R \) shown in their

\(^5\) https://github.com/pefreeman/galmorph
Analysing data via local two-sample tests

Figure 4. Boxplots summarizing the distributions of each morphological statistic for the morphology–mass study (left-hand bank of panels; Section 3.2) and morphology–SFR study (right-hand bank of panels; Section 3.3). Within each bank, the leftmost boxplots (red) represent statistics for galaxies identified as being in HMD or LSD regions, the rightmost boxplots (blue) are for galaxies in LMD or HSD regions, and the centre boxplots (grey), labelled ‘NS’ for ‘not significant’, are for all other galaxies. The top left-hand plot in each bank of panels indicates the proportion of zero values for the $M$ and $I$ statistics, with the same colour scheme as above. We observe that galaxies in HMD/LSD regions are more concentrated and less disturbed than their LMD/HSD-region counterparts.

Table 1. Sample size by the CANDELS field.

<table>
<thead>
<tr>
<th>Field</th>
<th>Total</th>
<th>F814W selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMOS</td>
<td>38 671</td>
<td>704</td>
</tr>
<tr>
<td>EGS</td>
<td>41 457</td>
<td>539</td>
</tr>
<tr>
<td>GOODS N</td>
<td>35 451</td>
<td>785</td>
</tr>
<tr>
<td>UDS</td>
<td>35 932</td>
<td>459</td>
</tr>
<tr>
<td>Total</td>
<td>186 441</td>
<td>2487</td>
</tr>
</tbody>
</table>

equation (1) by the number of pixels in the segmentation map, which places a hard upper limit of 0.5 on the area ratio $R_{l\text{, max}}$, achieved when $A_{l\text{, (1)}} = A_{l\text{, (2)}} = n_{\text{seg}}/2$. We adopt the catalogued SFRs estimated by A. Fontana (via ‘method 6.C’ described in Mobasher et al. 2015).

For the analyses below, we split the data into training and test sets of 1787 and 700 galaxies, respectively. We then assign the smallest and largest 25 per cent of mass (or SFR) values for the training set galaxies to the low-mass (or low-SFR) and high-mass (or high-SFR) groups, respectively.

3.2 Morphology–mass study

In this study, the predictor and response variables are, respectively, given as

(i) the morphological summary statistics $M$, $I$, $D$, $G$, $M_{20}$, $C$ and $A$; and
(ii) the estimated stellar mass $M_\star = \hat{M}_{\text{med}}$.

We sort the training set masses in ascending order and identify the upper and lower quartiles (see the top panel of Fig. 2). High- and low-mass galaxies are defined as those with $\log_{10}(M) > 10.639$ and $< 10.070$, respectively, with the global ratio of low- to high-mass galaxies being unity by definition.

The application of our local two-sample testing algorithm yields the following results for the 700-galaxy test set:

(i) A total of 108 galaxies are identified as lying in regions of the predictor space where the proportion of high-mass to low-mass galaxies is significantly larger than 1.

(ii) A total of 169 galaxies lie in primarily low-mass regions of predictor space.

In the left-hand panel of Fig. 3, we display variable importances determined by the random forest. These indicate that measures of light concentration ($M_{20}$, $C$, $G$) are more important than measures of disturbance ($D$, $M$, $A$, $I$) for estimating local proportions of high- to low-mass galaxies. This observation is consistent with the boxplots displayed in the left-hand panels of Fig. 4, which show the distributions of individual morphological statistics for galaxies in high-mass-dominated (HMD; red) and low-mass-dominated (LMD; blue) regions, as well as those for galaxies lying where neither class dominates (grey). We observe that galaxies in HMD regions are more concentrated and less disturbed than their counterparts in LMD regions, as one would expect, given the redshift range of our sample and previous results regarding ‘cosmic downsizing’ (e.g. Cowie et al. 1996; see also fig. 7 of Bundy, Ellis & Conselice 2005), which demonstrates at $z \sim 1$ that the fraction of

6 To be clear: A galaxy that is identified as lying in a region predominantly containing, e.g. high-mass galaxies is not necessarily itself a high-mass, upper-quartile galaxy.
galaxies classified as peculiar decreases as mass increases – less disturbance – while the fraction of galaxies classified as ellipticals increases with mass – greater concentration.

To further visualize how the morphologies of galaxies, and their statistics, change within HMD and LMD regions, we utilize the diffusion map algorithm (Coifman et al. 2005; Coifman & Lafon 2006; Lafon & Lee 2006; Lee & Freeman 2011). Diffusion map is useful for uncovering non-linear sparse structure in high-dimensional data, including submanifolds, clusters and high-density regions. We contrast this with, for example, principal components analysis, a linear method wherein data are projected on to hyperplanes. The interested reader can find further general details within the astronomical literature in, for example, Richards et al. (2009) and Freeman et al. (2009), and details on our specific application of diffusion map in Appendix A. The result of our application is to map galaxies from the original seven-dimensional space of morphological statistics to an easily visualized two-dimensional diffusion space. In Fig. 5, we display the first two diffusion coordinates for all 2487 galaxies.

In the top panel of Fig. 6, we retain the diffusion coordinates for the test-set data and mark those in HMD and LMD regions using red crosses and blue circles, respectively. We observe a clear separation between HMD- and LMD-region galaxies along the first diffusion coordinate (i.e. from the left- to right-hand side). Galaxies in the HMD region exhibit a consistent appearance – concentrated, symmetric and undisturbed – with general evidence of discs, while galaxies in the LMD region are generally less concentrated and exhibit a range of appearances due to increased disturbance that becomes more prevalent towards the right-hand side. To further quantify these results, we plot the individual statistics of each HMD- and LMD-region galaxy as a function of principal curve coordinates in the left-hand bank of panels in Fig. 7. Principal curves are smooth one-dimensional curves that provide a one-dimensional summary of multidimensional data (Hastie & Stuetzle 1989). In the top left-hand panel of Fig. 7, the principal curves that interpolate the HMD- and LMD-region galaxies are shown as solid red and blue lines (to the left- and right-hand side in each panel, respectively). The other panels in the left-hand bank indicate results consistent with those described above: Galaxies become progressively less concentrated and more disturbed as one moves from the left- to right-hand side within regions, and across regions.

To assess the consistency of our results with an a priori expectation, we construct UVJ and $M_*-UV$ diagrams (Figs 8 and 9). In the top panel of Fig. 8, quenched galaxies are identified as those that lie above the locus defined by $U - V = \max[1.3, 0.88(V - J) + 0.59]$ and the vertical line $V - J = 1.6$ (Williams et al. 2009). We observe that the majority of galaxies from HMD regions lie on the tight locus of quenched galaxies, as expected. In the top panel of Fig. 9, quenched galaxies lie towards the top (with $U - V$ values $\geq 1.3$). We observe a positive correlation between being in an HMD region and being quenched, as expected.
Figure 7. Scatter plots of each morphological statistic as a function of principal curve coordinate for the morphology–mass study (left-hand bank of panels; Section 3.2) and morphology–SFR study (right-hand bank of panels; Section 3.3). These scatter plots take advantage of the enhanced data visualization afforded by diffusion map to show more details about the distributions of statistics than the summaries provided by the boxplots in Fig. 4. The top left-hand panel in each bank shows the first two diffusion coordinates for galaxies identified as being in HMD or LSD regions (red crosses, towards the left-hand side in each panel), and LMD or HSD regions (blue circles, towards the right-hand side in each panel), with principal curves for each set of points overlaid. (Note that the red and blue points are the same red and blue points as displayed in Fig. 6; we have simply removed the grey points displayed in that figure and decreased the plotting range for the second coordinate to enhance visibility.) In the remaining panels of each bank, the width of each box corresponds to the totality of each principal curve. Note that for the $M$ and $I$ panels, zero values are not displayed (leading to, e.g. the absence of points in the middle box of the $I$ panel for the SFR study). The panels in both the left- and right-hand banks indicate that galaxies become progressively less concentrated and more disturbed as one moves from the left- to right-hand side through the identified regions, as well as across regions.

3.3 Morphology–SFR study

In this study, the predictor and response variables are, respectively, given by

(i) the morphological summary statistics $M$, $I$, $D$, $G$, $M_{20}$, $C$ and $A$; and

(ii) the estimated SFR $S$.

As is the case for the mass study, we sort the training set SFRs in ascending order and define low- and high-SFR galaxies as those with $\log_{10}(S) < -1.01$ and $>1.232$, respectively. (See the bottom panel of Fig. 2.) The application of our two-sample testing algorithm yields the following:

(i) A total of 313 out of 700 test-set galaxies lie in high-SFR-dominated (HSD) regions.

(ii) A total of 214 test-set galaxies lie in low-SFR-dominated (LSD) regions.

In the right-hand panel of Fig. 3, we see that the most important variables for estimating local proportions of high- to low-SFR galaxies are those that measure the concentration of light ($C$ and $G$), although the separation between concentration-related and disturbance-related statistics is not as clear-cut as it is for the mass study. As with the mass study, the boxplots displayed in the right-hand panels of Fig. 4 indicate clear differences between the distributions of summary statistics for each class of galaxies, with galaxies in LSD regions being more concentrated and less disturbed than their counterparts in HSD regions.\(^7\)

In the bottom panel of Fig. 6, we display the first two diffusion map coordinates for the galaxy test-set data, with galaxies in LSD and HSD regions marked as red crosses and blue circles, respectively. These regions are largely but not entirely coincident with the HMD and LMD regions shown in the top panel.\(^8\) As is the case for the HMD region, galaxies in the LSD region are concentrated, symmetric and undisturbed, but in addition to those galaxies that show visual evidence of discs, there are

---

7 We note here that we have also examined the specific star formation rate (SFR) in addition to the SFR; the results of our SSFR analysis are qualitatively similar and are not shown.

8 See Table 2. We note that it (along with Fig. 6) indicates that, for example, to the left-hand side in the displayed diffusion map there are galaxies that are associated with both the LSD and HMD regions, with just one or the other, or with neither region. We defer a detailed study of the morphological differences between these four classes to a future work.
Figure 8. UVJ diagrams for the morphology–mass study (top panel; Section 3.2) and morphology–SFR study (bottom panel; Section 3.3). In the top panel, galaxies in HMD and LMD regions are shown as red crosses and blue circles, respectively; in the bottom panel, LSD- and HSD-region galaxies are similarly shown. These diagnostic diagrams show that, consistent with an a priori expectation, quenched galaxies largely inhabit a tight locus on the UVJ diagram that lies above the less-tight locus of star-forming galaxies.

Conversely, as is the case for the LMD region, galaxies in the HSD region are generally less concentrated and exhibit a range of disturbed morphologies, particularly, at the right-hand end of the panel, an HSD but not LMD region that is exclusively populated by highly disturbed galaxies that are presumably undergoing mergers. In the right-hand bank of panels in Fig. 7, we show from left to right the galaxy summary statistics for two principal curves fitted to LSD-region galaxies and one fitted to HSD-region galaxies. Due to the near-coincidence of the LSD/HSD and HMD/LMD regions, the observed statistic distributions are similar to those in the left-hand bank of panels, but we do note that the HSD region exhibits lower $G$, higher $M_{20}$, and higher $A$ values, etc., than the LMD region, due to the fact that it contains a higher number of disturbed galaxies.

In Figs 8 and 9, we show $UVJ$ and $M_*-UV$ diagrams with LSD and HSD regions highlighted. As was the case for the mass study, we observe that the majority of galaxies in LSD regions lie on the locus of quenched galaxies, and that there is a positive correlation between being in an LSD region and being quenched.

Figure 9. Same as Fig. 8, except that here we plot $U - V$ versus $M_*$. We observe that low-SFR (i.e. quenched) galaxies are associated with higher mass galaxies.

4 SUMMARY

In this paper, we provide the astronomical community with a local two-sample hypothesis test framework that one can use to more easily analyse data (e.g. morphological statistics, photometric magnitudes, mass and SFR estimates, etc.) in their native high-dimensional spaces. In this framework, one defines two classes based on a response variable of interest (e.g. the top and bottom 25 per cent of the stellar masses for a sample of galaxies), and uses regression to compute the class posterior estimates $\hat{P}(Y = y | x)$, given a predictor datum $x$ and where $y$ denotes one of two discrete classes (e.g. high mass in a comparison of high mass and low mass, etc.). We leverage the work of Wager et al. (2014) and Wager & Athey (2017) to convert these estimates into a asymptotically unbiased test statistic that under the null hypothesis converges to a standard normal distribution (equation 5). In our implementation, we split data into training and test sets, using the former to learn estimates $\hat{P}(Y = y | x)$ and generating test statistics at the latter. To mitigate the effect of multiple comparisons (i.e. the fact that the number of tests performed is greater than one), we apply the Benjamini–Hochberg procedure. More details are provided in Algorithm 1, and R-based software is available at github.com/pefreeman/lst.

Our testing framework has a potential myriad of uses, as it is suitable for use in any analysis situation in which one wishes to test
whether a locally estimated proportion of two classes of objects is significantly different from the global proportion. In this paper, we demonstrate the efficacy of our testing framework by applying it to a set of 2487 i-band-selected galaxies observed by the HST-ACS in the COSMOS, EGS, GOODS-North and UDS fields. For these galaxies, we compute seven morphological statistics (M, I, D, G, M200, C, A) and thus estimate \( P(Y = y|x) \) in this seven-dimensional space. (We note that because our estimation makes use of random forest regression, one can apply our framework to spaces of considerably higher dimensionality. For reference, the computation time for our analyses is ~1 CPU minute.) We perform two studies, one in which we determine the local proportion of high-mass (top 25 per cent of masses) to low-mass (bottom 25 per cent of masses) galaxies and another using SFR in place of mass. Both studies yield qualitatively similar results: Galaxies lying in identified high-mass or low-SFR regions exhibit a consistent appearance – concentrated, symmetric, undisturbed and generally with visual evidence of disc structure – while their counterparts in low-mass or high-SFR regions have less concentrated light and exhibit increasing levels of disturbance. We display these results first with boxplots (Fig. 4) but then show how one can further potentially visualize results at finer scales by transforming the predictor data into a lower dimensional space; here, we specifically apply diffusion map (Figs 6 and 7). We provide details on diffusion map in Appendix A and \( R^2 \)-based software that implements visualization via diffusion map at the address given above.

ACKNOWLEDGEMENTS

The authors would like to thank the members of the CANDELS collaboration for providing the data upon which this work is based. We would also like to thank Jeff Newman (University of Pittsburgh) for acting as IK’s external adviser for the project on which this paper is based, and Rafael Izbicki (Federal University of São Carlos) and Jen Lotz (Space Telescope Science Institute) for helpful discussions. This work was supported by NSF DMS-1520786 and NIMH R37MH057881. Our research has made use of SAOimage DS9. This work was supported by NSF DMS-1520786 and NIMH R37MH057881. We would also like to thank the members of the CANDELS collaboration for providing the data upon which this work is based. We would also like to thank Jeff Newman (University of Pittsburgh) for acting as IK’s external adviser for the project on which this paper is based, and Rafael Izbicki (Federal University of São Carlos) and Jen Lotz (Space Telescope Science Institute) for helpful discussions. This work was supported by NSF DMS-1520786 and NIMH R37MH057881. Our research has made use of SAOimage DS9, as well as the dmtools provided by the Chandra X-ray Center in the application package CIAO.

REFERENCES

Duhan T., 2013, JNS, 25, 635
Roedder M., Hardy R. R., 2001, Cytometry, 45, 56
S´ersic J. L., 1963, Boletín de la Asociacion Argentina de Astronomia, 6, 41
Székely G. J., Rizzo M. L., 2004, InterStat, 5, 1

APPENDIX A: DIFFUSION MAP

Dimensionality reduction methods are useful for visualizing low-dimensional structures embedded in higher dimensional spaces. One such method is diffusion map (Coifman et al. 2005; Lafon & Lee 2006),\(^9\) a non-linear method that seeks to preserve the connectivity structure of data within a high-dimensional space. (In practice, preservation means that the Euclidean distance between two points in diffusion space is approximately the same as the sum of all paths between the same two points in the original data space.) The connectivity structure is learned by modelling the traversal of the data space as a diffusion process.

As a starting point for constructing a diffusion map, one defines a weight that reflects the local similarity of two points, \( x_i \) and \( x_j \), in \( \mathcal{X} = \{x_1, \ldots, x_n\} \). In this work, we implement the weight estimator of Zelnik-Manor & Perona (2005):

\[
\tilde{w}(x_i, x_j) = \exp \left( -\frac{x_i - x_j}{\sigma_i \sigma_j} \right),
\]

\(^9\) Methods for computing diffusion coordinates, etc., are contained in the \( R \) package DIFFUSIONMAP.

Downloaded from https://academic.oup.com/mnras/article-abstract/471/3/3273/3979021 by guest on 08 October 2018
where \( s \) is the (e.g. Euclidean) distance between \( x_i \) and \( x_j \), and \( \sigma_i (\sigma_j) \) is the distance between \( x_i \) (\( x_j \)) and its \( k \)th nearest neighbour. (Note that we standardize the data of each predictor variable first, i.e. from each datum we subtract the sample mean and then divide the difference by the sample standard deviation.) We assume \( k = 30 \); other values give similar visualization results. (The appropriate value for \( k \) will, of course, differ from application to application.)

We use the weights \( \hat{w} \) to build a Markov random walk on the data with the transition probability from \( x_i \) to \( x_j \) defined as

\[
p(x_i, x_j) = \frac{\hat{w}(x_i, x_j)}{\sum_{k=1}^{n} \hat{w}(x_i, x_k)}.
\]

The one-step transition probabilities are stored in an \( n \times n \) matrix \( P \), and then propagated by a \( t \)-step Markov random walk with transition probabilities \( P^t \). Instead of choosing a fixed time parameter \( t \), however, we combine diffusions at all times and define an averaged diffusion map\(^{10}\) according to

\[
\Psi_{av} : x \rightarrow \left[ \left( \frac{\lambda_1}{1 - \lambda_1} \right) \psi_1(x), \ldots, \left( \frac{\lambda_m}{1 - \lambda_m} \right) \psi_m(x) \right].
\]

where \( \lambda_i \) and \( \psi_i \) represent the first \( m \) eigenvalues and right eigenvectors of \( P \). In this work, we fix \( m \) to 2, i.e. we visualize only the first two dimensions in diffusion space.

\(^{10}\) Note that this is the default way by which the DIFFUSIONMAP package function \texttt{diffuse()} constructs diffusion maps.