Eigengalaxies: describing galaxy morphology using principal components in image space

Emir Uzeirbegovic [1,2]★ James E. Geach [1,2] and Sugata Kaviraj [1,2]

[1]*Centre for Astrophysics Research, School of Physics, Astronomy and Mathematics, University of Hertfordshire, Hatfield AL10 9AB, UK*
[2]*Centre of Data Innovation Research, School of Physics, Astronomy and Mathematics, University of Hertfordshire, Hatfield AL10 9AB, UK*

ABSTRACT

We demonstrate how galaxy morphologies can be represented by weighted sums of 'eigengalaxies' and how eigengalaxies can be used in a probabilistic framework to enable principled and simplified approaches in a variety of applications. Eigengalaxies can be derived from a Principal Component Analysis (PCA) of sets of single- or multiband images. They encode the image space equivalent of basis vectors that can be combined to describe the structural properties of large samples of galaxies in a massively reduced manner. As an illustration, we show how a sample of 10243 galaxies in the *Hubble Space Telescope* CANDELS survey can be represented by just 12 eigengalaxies. We show in some detail how this image space may be derived and tested. We also describe a probabilistic extension to PCA (PPCA) which enables the eigengalaxy framework to assign probabilities to galaxies. We present four practical applications of the probabilistic eigengalaxy framework that are particularly relevant for the next generation of large imaging surveys: we (i) show how low likelihood galaxies make for natural candidates for outlier detection; (ii) demonstrate how missing data can be predicted; (iii) show how a similarity search can be performed on exemplars; (iv) demonstrate how unsupervised clustering of objects can be implemented.

Key words: methods: data analysis – methods: statistical – techniques: image processing – galaxies: structure.

1 INTRODUCTION

The distribution of light in galaxies, commonly referred to as galaxy 'morphology', is a fundamental observable property. Morphology strongly correlates with the physical properties of a galaxy, such as its stellar mass (e.g. Bundy, Ellis & Conselice 2005), star formation rate (e.g. Ryan et al. 2012; Bluck et al. 2014; Smethurst et al. 2015; Willett et al. 2015), surface brightness (e.g. Martin et al. 2019; Jackson et al. 2020b), rest-frame colour (e.g. Strateva et al. 2001; Bamford et al. 2009; Skibba et al. 2009) and local environment (e.g. Dressler et al. 1997; Postman et al. 2005) and reveals key information about the processes that have shaped its evolution over cosmic time (e.g. Martin et al. 2018b; Jackson et al. 2020a). For example, the smooth light distributions of elliptical galaxies, which are a result of the largely random orbits of their stars (e.g. Cappellari et al. 2011), are signposts of a merger-rich evolutionary history (e.g. Conselice 2006). On the other hand, the presence of a disc indicates a relatively quiescent formation history, in which the galaxy has grown primarily through accretion of gas from the cosmic web (Codis et al. 2012; Martin et al. 2018a). In a similar vein, morphological details such as extended tidal features suggest recent mergers and/or interactions (e.g. Kaviraj 2014; Jackson et al. 2019; Kaviraj, Martin & Silk 2019), with the surface brightness of these tidal features typically scaling with the mass ratios of the mergers in question (e.g. Kaviraj 2010; Peirani et al. 2010).

Apart from being a fundamental component of galaxy evolution studies, morphological information has a wide range of applications across astrophysical science. For example, it can be a key prior in photometric redshift pipelines (e.g. Menou 2018; Soo et al. 2018) which underpin much of observational cosmology and weak lensing studies, is used as contextual data in the classification of transient light curves (e.g. Djorgovski et al. 2012; Wollaeger et al. 2018) and is an essential ingredient in the study of the processes that drive active galactic nuclei (e.g. Schawinski et al. 2014; Kaviraj et al. 2015). The morphological analysis of galaxy populations, especially in the large surveys that underpin our statistical understanding of galaxy evolution, is therefore of fundamental importance.

A vast literature exists on methods for measuring galaxy morphology. Popular techniques range from those that describe a galaxy's light distribution using a small number of parameters (e.g. de Vaucouleurs 1948; Sérsic 1963; Odewahn et al. 2002; Simard et al. 2002; Lackner & Gunn 2012) to non-parametric approaches such as 'CAS' (e.g. Abraham et al. 1994; Conselice 2003; Menanteau et al. 2006; Mager et al. 2018) or Gini-M_{20} (e.g. Lotz, Primack & Madau 2004; Scarlata et al. 2007; Peth et al. 2016), where the light distribution is reduced to a single value. The convergence of large observational surveys and rapidly increasing computing power has recently brought machine learning to the fore in morphological studies. While the use of machine learning can be traced back at least as far as the 1990s (e.g. Lahav et al. 1995), there has been a recent explosion of studies that apply such techniques to the exploration of galaxy morphology, particularly in large survey data sets (e.g. Huertas-Company et al. 2015; Ostrovski et al. 2017; Schawinski et al. 2017; Goulding et al. 2018; Hocking et al. 2018; Cheng et al. 2019; Martin et al. 2020).

Automated techniques lend themselves particularly well to the analysis of large surveys, but the most accurate method of mor-

★ E-mail: e.uzeirbegovic@herts.ac.uk

phological classification is arguably visual inspection (e.g. Kaviraj et al. 2007, 2019). Indeed the genesis of this subject can be traced back to the visual 'tuning-fork' classifications by Hubble (1926), where galaxies were classified into a so-called sequence of ellipticals and spirals. It is remarkable that this classification system still underpins the broad morphological classes into which galaxies are split in modern studies of galaxy evolution. Although visual inspection of large observational surveys is time-intensive, the advent of massively distributed systems like Galaxy Zoo has revolutionized its use on survey data sets (e.g. Lintott et al. 2011; Simmons et al. 2014, 2017; Willett et al. 2017). Galaxy Zoo has used more than a million citizen-science volunteers to classify large contemporary surveys, like the Sloan Digital Sky Survey (SDSS) and the *Hubble Space Telescope* (*HST*) Legacy Surveys and has provided the benchmark against which the accuracy of automated techniques have been routinely compared (e.g. Dieleman, Willett & Dambre 2015; Huertas-Company et al. 2015; Beck et al. 2018; Ma et al. 2019; Walmsley et al. 2019).

In contrast to classification based approaches, this paper describes how morphology may be interpreted as the 'layout' of an image space which is survey specific, linear, and has a probabilistic interpretation. We show how these properties enable the same image space to serve as a multi-application framework. The paper is structured as follows. In Section 2 we describe both the sample and our methodology, we discuss the construction of the image space, its validity, and its relevance to the interpretation of galaxy morphology. We also show how this result is extended to include a probabilistic interpretation. In Section 3 we show how this framework leads to principled approaches in various applications. In particular, we (i) show how low likelihood galaxies make for natural candidates for outlier detection; (ii) show how missing data can be predicted; (iii) demonstrate how similarity searches for images can be implemented given an exemplar; (iv) demonstrate how a natural unsupervised clustering of objects can be produced. We then outline how the methods may be used together in more advanced applications and how it is relevant to big data sets such as the upcoming Large Synoptic Survey Telescope (LSST) project (Robertson et al. 2017). Section 4 concludes and summarizes our findings, and provides a link to the codes developed for this work.

2 EIGENGALAXIES

2.1 Data from *HST* CANDELS

HST CANDELS (Grogin et al. 2011; Koekemoer et al. 2011) offers a high-resolution probe of galaxy evolution. The survey consists of optical and near-infrared (WFC3/UVIS/IR) images from the Wide Field Camera 3 (WFC3) and optical images from the Advanced Camera for Surveys (ACS) in five well-studied extragalactic survey fields. Here, we focus on GOODS-S, one of the deep tier (at least four-orbit effective depth) fields. We select a sample of 10 243 galaxies present in the 'Galaxy Zoo: CANDELS' (GZ-CANDELS) GOODS-S catalogue (Simmons et al. 2016), that fall within the region jointly covered by the F814W, F125W, and F160W bands. The majority of objects are at $z < 3$ (Simmons et al. 2016).

For each object in the catalogue, we take a 1.8 arcsec cut-out, using the catalogued sky coordinate as a centroid. ACS images are downsampled by a factor of two to match the pixel scale of the WFC3 images (0.06 arcsec pixel^{-1}), so that for each band we obtain a 30×30 pixel image at the position of each galaxy. The cut-out dimensions are selected with prior experimentation because they result in the least superfluous background for most targets. Individual

cut-outs are given as $30 \times 30 \times 3$ pixel array in which the dimensions index the height, width, and band, respectively.

2.2 Morphology as an image space

We may gain considerable flexibility by thinking about galaxies as points in a vector space and about morphology as the 'layout' of the space. In this section, we describe a simple vector space sufficient for our demonstrations in Section 3.

The first task is to construct a projection from the $30 \times 30 \times 3$ cut-outs to vectors in a *j*-dimensional space. This kind of embedding has been done in many ways in existing literature. Some examples from astronomical applications include Naim, Ratnatunga & Griffiths (1997) who pre-calculate features thought to be important in order to feed a self-organizing map (Kohonen 1990), and Hocking et al. (2018) who uses patches of pixels as a primitive and then calculate histograms of patches to reduce images to vectors. Here, we avoid techniques which presuppose meaningful summarizations from the outset and begin by making a naive projection by flattening cut-outs into vectors,[1] hence $j = 2700$. There are many possible constraints and coherences that may be desirable in an image space but we will be singularly interested in preserving the similarity of galaxy points which are 'near' each other. This generates clear testable implications which we can validate. The two main factors bearing on this correspondence are how the projection maps galaxies to vectors and how the distance between matrices is calculated. These will now be considered in turn.

A first necessary modification to the naive projection is to remove uninteresting sources of variation. In addition to the image centering, pixel scaling, and cropping described in Section 2.1, we identify range clipping and rotation to a common plane as the most important modifications. To achieve these, we truncate the pixel distribution at the 99th percentile to avoid extreme spikes in flux in some galaxies from making them seem excessively different to the rest. Next, we produce a temporary composite image of the array by using, for each pixel, its maximum magnitude across all bands. We then use the locations of all pixels in our composite with values exceeding the 75th percentile to create a matrix of the coordinates of bright pixels. The first principal component of the two column matrix of pixel coordinates is a vector of weights (x_0, y_0) which can be interpreted as the linear transformation of our 2D coordinates to the 1D space that preserves most variance (i.e. some line passing through the origin of the original 2D space). The angle between the original *x*-axis in our image and the new variance-preserving one is then given by $\theta = \arctan2(y_0, x_0)$. Finally, we transform the original array by rotating every band in turn by $\theta - \pi/2$ radians (in order to make vertical what would otherwise be a rotation to the horizontal plane) which results in vertically aligned brightness.

2.3 Low-rank approximation

Having fixed how images are to be represented in our projection, we can further improve the image space by noticing that flux density in image pixels are typically highly correlated. This implies that it may be possible to closely approximate a galaxy vector in fewer than *j* dimensions. This is very much desirable because a $j = 2700$ dimensional space will likely suffer from some of the degeneracies

---

[1]The order in which the array is flattened does not matter because we will only be interested in the pairwise relationships between pixels, which is not changed if they are presented in a different order.

sometimes termed 'the curse of dimensionality' which includes effects like the loss of interpretability of distance measures discussed further in Section 3.6.

Let $S = (\mathbf{v}_1, ..., \mathbf{v}_L)$ be a survey matrix in which the row vectors $\mathbf{v}_i$ with j components are the projected images. We are interested in $\hat{\mathbf{S}}$ with dimensions $L \times k$ where $k < j$. The associated minimization problem can be expressed as:

$$\min_{\hat{\mathbf{S}}} ||\mathbf{S} - \hat{\mathbf{S}}|| \text{ s.t. rank}(\hat{\mathbf{S}}) < l. \tag{1}$$

Here, $|| \cdot ||$ (the norm) governs how the distance between vectors is quantified. The Frobenius norm (Gloub & Van Loan 1996) is given by

$$||\mathbf{X}|| = \sqrt{\sum_{i=0}^{m} \sum_{j=0}^{n} |\mathbf{x}_{i,j}|^2}. \tag{2}$$

This is commonly used and can be thought of as a matrix generalization of the Euclidean distance. Given a Frobenius norm, the minimization problem in equation (1) has a globally optimal, unique, and analytical solution as a consequence of the Eckart–Young–Mirsky (EYM) theorem (Eckart & Young 1936). The singular value decomposition of S is such that $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U},\mathbf{V}$ are orthogonal matrices and $\mathbf{\Sigma}$ is a diagonal matrix with $(\sigma_1, ..., \sigma_k)$ singular values. The theorem shows that given a Frobenius norm, the optimal and unique solution to equation (1) for a matrix $\hat{\mathbf{S}}$ of rank k is given by $\hat{\mathbf{S}} = \sum_{i=1}^{k} \sigma_i \mathbf{u_i}\mathbf{v_i}^T$, where $k \leq l$, and $\mathbf{u}_i$, $\mathbf{v}_i$ are vectors from $\mathbf{U}$, $\mathbf{V}$, respectively.

Principal component analysis (PCA) is a recursive procedure in which a plane (principal component, PC) which minimizes the average squared distance from points to the PC is fitted to the residuals of the previous iteration. The process carries on until there are no further residuals. An efficient way to measure the contribution of each PC is to examine the covariance matrix of the data projected on to the new basis. Since each PC is required to be orthogonal, the covariance matrix is diagonal. Further, since residual difference is smaller on each iteration, subsequent PCs account for progressively less variance thus creating a natural ordering. The cumulative sum of the diagonal elements of the covariance matrix therefore provides a convenient way to measure the total fraction of variance accounted for by k PCs. This measure is often termed 'explained variance' (EV) and the fraction of variance explained by any given PC is termed the 'explained variance ratio' (EVR). If the data are centred, it turns out that PCs discovered in the iterative process above are equivalent to the eigenvectors of the covariance matrix $\mathbf{S}^T\mathbf{S}$ when ordered by their corresponding eigenvalues. It is further the case that the eigenvectors are equivalent to the right singular vectors in matrix V above and the eigenvalues are given by $\lambda_i = \sigma_i^2$. Thus, it can be shown by the EYM theorem that PCA is also an optimal solution to the minimization problem posed in equation (1).

PCA has a long precedence in astronomy applied in other ways to other things. For example, De La Calleja & Fuentes (2004) use PCA to project a set of 310 disparate images to a lower rank space to facilitate further classification steps. They referred to the basis eigenvectors as 'eigengalaxies' and used their weightings to test multiple machine learning methods against each other for the purpose of galaxy classification. Li et al. (2005) use PCA to decompose stellar spectra from the STELIB spectroscopic stellar library (Le Borgne et al. 2003) and Two Micron All Sky Survey (2MASS; Skrutskie et al. 2006) near-infrared photometry to derive 'eigenspectra' and then used them to fit the observed spectra of a selection of galaxies from the SDSS (Fukugita et al. 1996) Data Release 1. Anderson et al.

(2004) generate many galaxies from a parametric model, calculated the eigenvectors of the generated set, and then used the reduced eigenspace to find the nearest synthetic model. Wild et al. (2014) use PCA to concisely describe a large number of model spectral energy distributions (SEDs). They termed the retained eigenvectors as 'super-colours' and used them to impute SEDs from sparse samples in observed galaxies. Galaz & de Lapparent (1998) use PCA to perform spectral classification of selected galaxies from the ESO-Sculptor galaxy redshift survey data (de Lapparent et al. 1993).

In the analysis that follows, we use the Frobenius norm and its vector equivalent, the Euclidean distance, as difference measures. We use PCA to optimally reduce the dimensionality of the image space. We term the eigenvectors as 'eigengalaxies' because they are j dimensional and can be reshaped back into $30 \times 30 \times 3$ cut-outs which map how they are weighted as later shown .

2.4 Results and interpretations

The projection and low-rank approximation enables us to create a low dimensional image space in which morphology is expressed in terms of the distances between galaxies. The same general methodology may be used to make image spaces for a broad range of surveys, and there are no strict limitations on image sizes, bands, or resolutions required. However, it is essential to test that the projection and the low rank approximation result in a space fit for purpose. Generally, PCA only fails when there are multiple identical eigenvalues which results in the associated eigenvectors not being unique, rendering the result meaningless. If PCA does not fail, it does not mean that it is appropriate to the problem at hand. However, there are a variety of heuristic means we can adopt to test the adequacy of the resultant space which we now cover.

Fig. 1 helps clarify what features of images the space preserves. The figure shows composite thumbnails of a random selection of galaxies followed by their reconstruction at different numbers of eigengalaxies. Since our space aims to preserve neighbourhood similarity, we can heuristically evaluate the suitability of the space for any given k by picking random galaxies and then looking to see what their nearest neighbours look like. Experimentation indicates that $k = 12$ is enough for good results. Fig. 2 illustrates 18 sets of galaxies. In each set, the galaxy in the top left is randomly selected, followed by eight of the nearest galaxies to it in image space, ordered by row. Even for noisy and complicated examples the nearest neighbours tend to be visually similar, suggesting that the image space does a good job of preserving visual similarity in local neighbourhoods.

Fig. 3 shows the cumulative variance explained by successive eigengalaxies. PCA successively maximizes the variance in each orthogonal eigengalaxy and each additional eigengalaxy therefore accounts for less variance. In this instance, only two eigengalaxies are required to account for ~85 percent of explained variance, but 12 eigengalaxies are required for reliable nearest neighbour similarity which accounts for ~96 percent of explained variance. Since eigenvectors are 2700 dimensional, we can investigate this front loaded distribution by reshaping them back into a $30 \times 30 \times 3$ images to get a sense of what each eigenvector emphasizes. Fig. 4 presents the 12 eigengalaxies as grey-scale images. It is evident that the first two eigengalaxies focus on central brightness whilst later eigengalaxies are successively more complicated, and therefore more likely associated to the details of spatial correlation. The EVR directly relates distances in the image space to the source of their variation, hence morphology is herein described mostly by relative flux densities across bands and coarse pixel correlation.
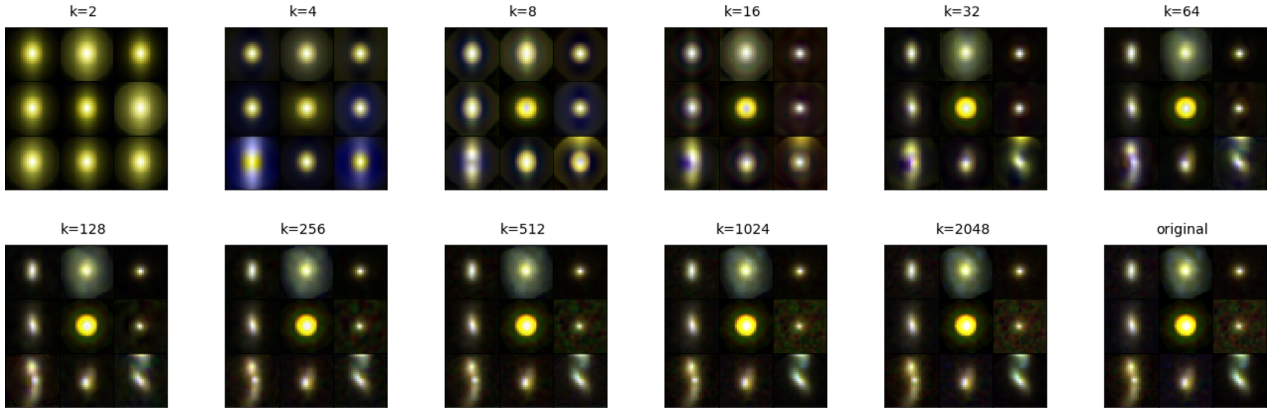
Figure 1. RGB composite (F160W, F125W, and F814W) thumbnails for a selection of galaxies reconstructed using different numbers of eigengalaxies (indicated by k). The number of eigengalaxies increases by powers of 2. The original image is shown in the bottom right. The fidelity of the reconstruction is directly related to the retention of eigengalaxies. Experimentation reveals that 12 eigengalaxies were enough to robustly preserve neighbourhood similarity.
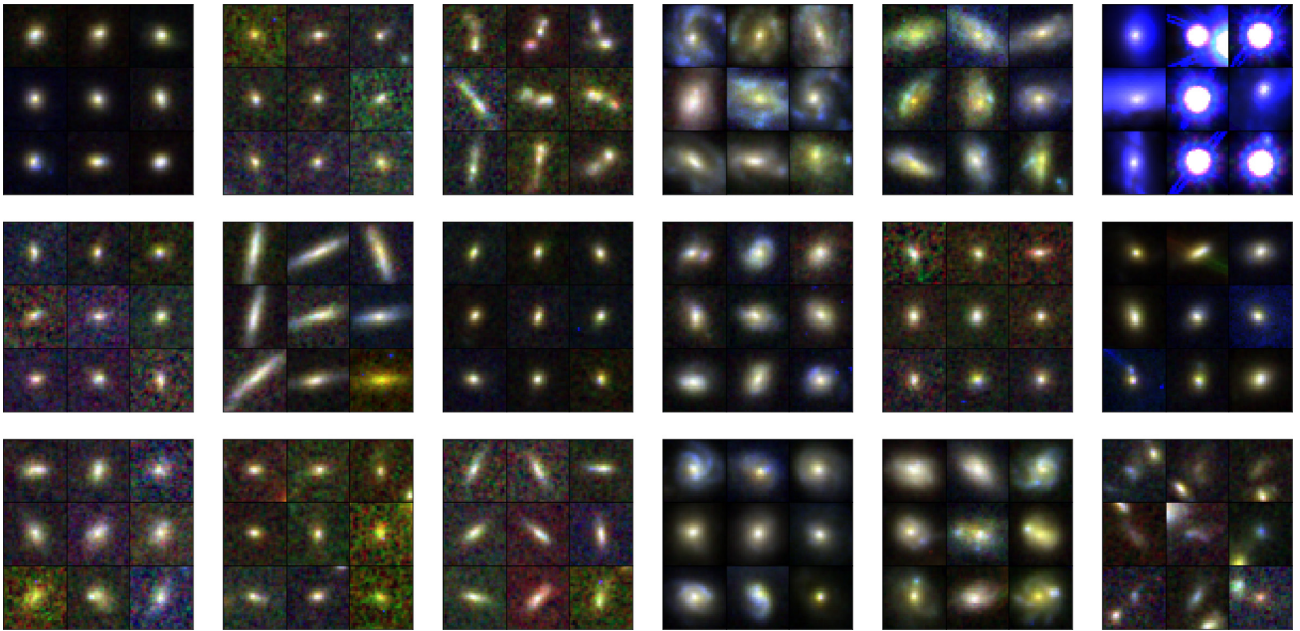

Figure 2. A collection of RGB composite (F160W, F125W, and F814W) thumbnail sets, in each of which the galaxy in the top left is randomly selected and is followed by eight of the nearest galaxies to it in image space, ordered by row. Even for noisy and complicated examples, the nearest neighbours tend to be visually similar suggesting that the image space does a good job of preserving visual similarity in local neighbourhoods.

While we have assumed that the eigengalaxies are equally representative of all galaxies, we can validate this assumption by graphically checking the distribution of the reconstruction error, which is calculated as the residual between the real image and its eigengalaxy based reconstruction. If it is unevenly distributed or bigger than 4 percent (i.e. 100-96 percent) then there may be something degenerate about the image space. Let F_0 be the collection of flux densities of the pixels in the original images and F_Δ be the corresponding differences in flux between the original pixels and those in images reconstructed using the eigengalaxies. Fig. 5 shows a histogram of $\sum_{x \in F_\Delta} x / \sum_{x \in F_0} x$ (i.e. the ratio of the sum of deltas and the sum of flux over all bands). It shows a symmetrical distribution centred at zero, 99 percent of objects lie between $|\sum_{x \in F_\Delta} x / \sum_{x \in F_0} x| < 2.5$ percent, hence it suggests that the eigengalaxies do represent most galaxies similarly.

We also assume that the eigengalaxies discovered are not an accident of the sample size. We can check this by making sure that the same number of eigengalaxies realize a similar explained variance at smaller sample sizes and that therefore there is a reason to believe that 12 eigengalaxies have an asymptotic EVR. Fig. 6 presents a contour plot of EVR corresponding to various combinations of sample size and eigengalaxy number. We see that at 12 eigengalaxies the contour has near zero or asymptotic growth as sample size increases, which suggests that the vector space is less likely to be an accident of the sample.

PCA is sensitive to outliers, and can perform poorly where data has very well separated clusters, and/or large non-linear sub-spaces. The most likely cause of outliers in our case would be extreme brightness against which we guard by range clipping. Further, there is no obvious reason to expect large non-linear sub-spaces and we would expect accidental well separated clusters to be highly unlikely in a 2700
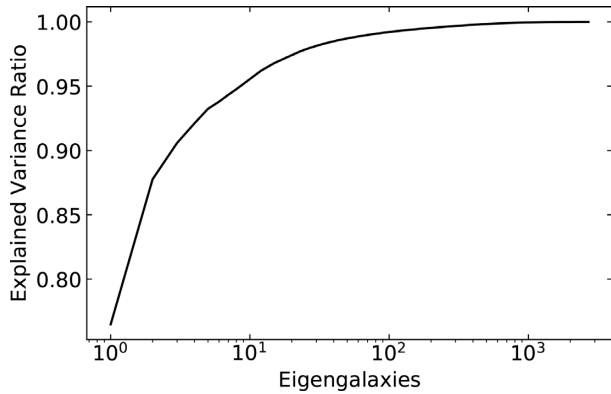
Figure 3. Cumulative explained variance illustrating how much variance each additional feature accounts for. In PCA successive components account for less variance than the previous component. In this instance, only two eigengalaxies are required to account for ∼85 percent of explained variance yet 12 eigengalaxies are required to account for 96 percent of explained variance.

dimensional space. At the time of writing, there are no standard well known tests which are able to diagnose these degenerate cases in high dimensional spaces but it should be possible to identify their effects at least heuristically by resampling the data, each time leaving out a large part of it, and then calculating the EVR for some fixed k.

We can pick $k = 12$ which does a good job as described above and $k = 1$, since the first eigenvector explains the most variance. If our data are summarized robustly by PCA, the variance of the EVRs given k components should be small. A large variance would give us reason to believe that the incidental presence/absence of outliers, clusters, non-linear sub-spaces, or some other degeneracy is causing substantial differences. It is important that the sample size taken on every iteration is large enough to make catching these effects likely. Fig. 7 illustrates a histogram of the EVR values achieved by performing PCA on 1000 random 70 percent samples of the data for $k = 12$ (top) and $k = 1$ (bottom), respectively. The figure shows distributions with tight ranges and thus provides some confidence that the data are appropriate for PCA as treated herein and does not suffer from the common degeneracies.

This image space is useful for general purposes such as exploring a new survey by filtering, clustering, ordering, and searching for morphologies of interest, as demonstrated in Section 3. All these analyses and many others can be performed on the same image space, which may result in a great deal of time saving and parsimony for the astronomer. We approximate the space to 12 dimensions because it is sufficient for our purposes, however in practice there is no need for the low rank approximation to be the same for all applications. For example, applications which depend on Euclidean distance benefit from lower dimensionality because the measure breaks down in high dimensions. Meanwhile, applications implicitly reliant on reconstruction fidelity which do not have issues with high dimensionality (such as missing data prediction) may benefit from approximating less by retaining more dimensions.

2.5 Probabilistic interpretation

There are many applications which require some way of taking the likelihood of galaxy points into account. This section describes a formal equivalence between fitting PCA (an optimal low-rank approximation) and fitting a Gaussian latent factor model under
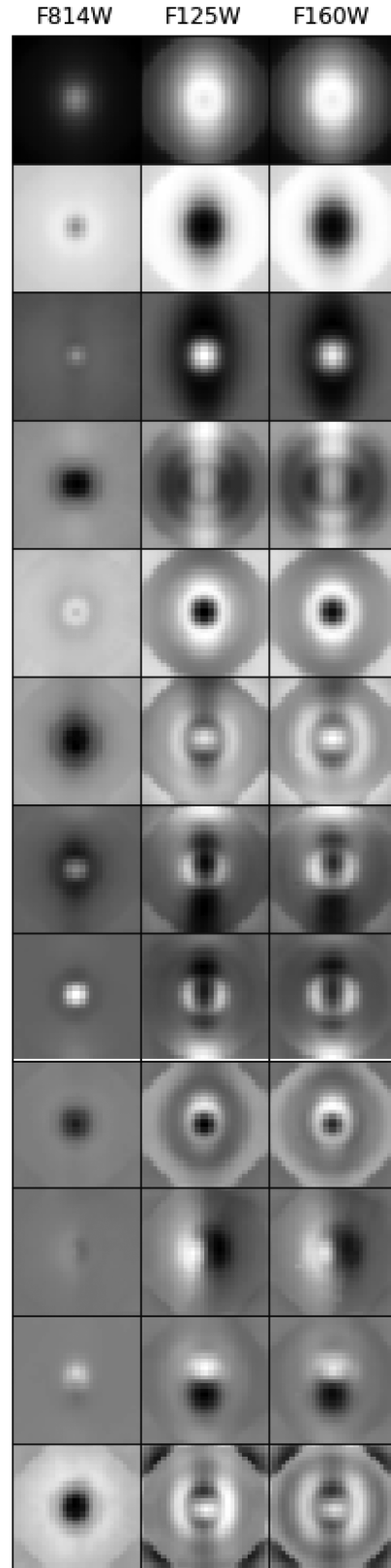

Figure 4. The 12 eigengalaxies accounting for 96 percent variance in the GZ-CANDELS GOODS-S sample. Each row is an eigengalaxy and each column is a band. All images are scaled identically, where white indicates relative emphasis and black indicates relative de-emphasis of the region. Each image is 30 × 30 pixels, or 1.8 × 1.8 arcsec^2.
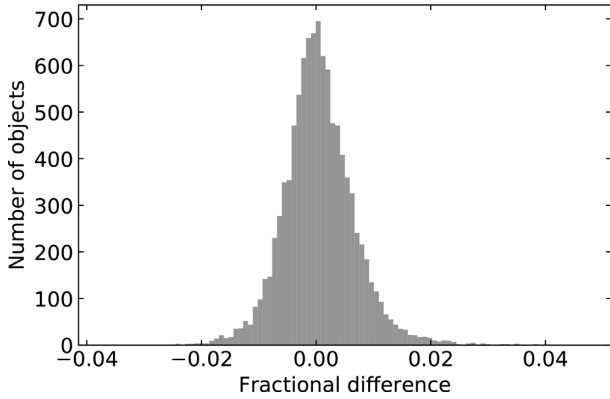
Figure 5. Histogram showing the frequency of the ratio of the sum of deltas (difference between original and reconstructed pixel flux over all bands) and the sum of the flux in the original images, for the PCA decomposition of the 10243 galaxies in the GZ-CANDELS GOODS–S sample. 99 percent of objects lie between ±2.5 percent. We account for ~96 percent of variance by design and the distribution above shows that the reconstructed images bear a good likeness (in terms of reconstructed flux) to the originals.
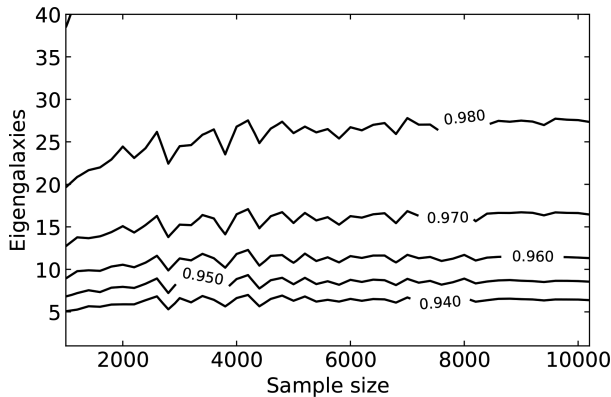

Figure 6. Contour plot of explained variance corresponding to the required number of eigengalaxies versus sample size. Higher explained variances tend to require proportionally more eigengalaxies as sample size increases, however at 12 eigengalaxies explained variance is asymptotic to a constant.

certain conditions. This equivalence will enable us to interpret PCA probabilistically and therefore to assign likelihood to galaxy points.

Given data vectors from some d dimensional space, a linear latent factor model (LFM) aims to discover the basis for an optimal projection to a $q < d$ dimensional space, usually under Gaussian assumptions. In its simplest form the model can be written as follows:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \mu + \epsilon. \tag{3}$$

Here $\mathbf{t} \in \mathbb{R}^d$ is the data vector, $\mathbf{x} \in \mathbb{R}^q$ is the latent vector, W is a $d \times q$ projection matrix relating t to x where $q < d$, μ is an offset, and ϵ is a residual error. Usually it is given that $x \sim N(0, \mathbf{I})$, and that $\epsilon \sim N(0, \Psi)$ where the form of Ψ is to be defined. This given, the properties of the normal distribution imply that $t \sim N(\mu, \mathbf{W}\mathbf{W}^T + \Psi)$. Whittle (1952) showed that in the case that $\Psi = \sigma^2\mathbf{I}$ (i.e. the covariance matrix is diagonal and isotropic), and σ^2 is known, the maximum likelihood estimation of the matrix W is equivalent to the linear least-squares solution. Resultantly, $\mathbf{W}$ spans the same subspace as PCA and hence is also an optimal solution to the low rank approximation problem formalized in equation (1). However, the formulation in Whittle (1952) is highly limiting since it is unlikely that in real data
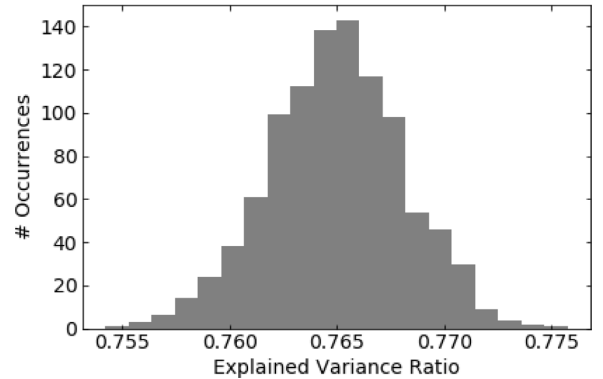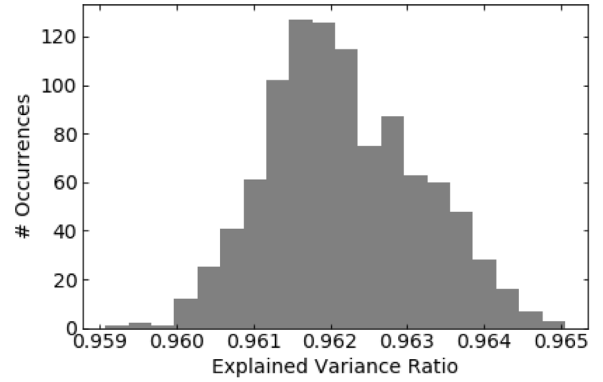

Figure 7. Both graphs show histograms of 1000 PCA fittings on 70 percent of the data randomly sampled on each iteration. The top image shows the EVR for $k = 12$. Note that the EVR at $k = 12$ when the whole data are retained is ~0.962. The bottom image shows the EVR for the first eigenvector. Note that the EVR at $k = 1$ when the whole data are retained is ~0.765. Both histograms show a very tight variance even at the extremes.

the covariance structure is entirely known or that the model and sample covariance are exactly the same. However, Tipping & Bishop (1999) show that maximum likelihood (ML) estimates for $\mathbf{W}$ and σ do exist without requiring the covariance to be known and, that the scaled principal eigenvectors make up the columns of $\mathbf{W}$ when the estimators are at their global maximum. They term this result probabilistic PCA (PPCA). The result (explained more formally below) is that given the PCA low rank approximation performed in the previous sections, we can directly write down an equivalent factor model which induces a multivariate Gaussian distribution over the image space and hence we are able to assign likelihoods in that space to every galaxy point. Other than the direct applications of this formulation covered in Section 3, the factor model allows us to compare image spaces by quantifying the implications of their structural differences on likelihood assignment to galaxy points, which we later leverage to create representative samples.

We will now signpost the crucial points in the derivation from Tipping & Bishop (1999), but the interested reader is encouraged to consult the paper directly. Given equation (3) and the assumption of diagonal and isotropic error $\epsilon \sim N(0, \sigma^2\mathbf{I})$, it follows that t conditional on x is given by $t|x \sim N(\mathbf{W}x + \mu, \sigma^2\mathbf{I})$. Since $x \sim N(0, \mathbf{I})$ it is easy to marginalize over x to obtain $t \sim N(\mu, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$. The corresponding log likelihood function is given by

$$\mathbf{L} = -\frac{N}{2}\left[d\ln(2\pi) + \ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\mathbf{S})\right], \tag{4}$$

where $\mathbf{S}$ is the sample covariance matrix and $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$. The ML estimator for μ is the sample mean. Meanwhile, globally optimal

estimates for σ, $\mathbf{W}$ can be obtained using iterative maximization algorithms such as those given in Rubin & Thayer (1982). Most importantly, what the authors show in Tipping & Bishop (1999) is that these parameters can be obtained analytically using the artefacts from PCA. The PCA equivalent of the σ^2 ML estimate is given by

$$\sigma_{\text{ML}}^2 = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j, \tag{5}$$

where λ_j are the excluded eigenvalues, hence it can be roughly interpreted as the variance lost averaged over the number of dimensions lost. The PCA equivalent of the $\mathbf{W}$ ML estimate is given by

$$\mathbf{W}_{\text{ML}} = \mathbf{U}_q (\Delta_q - \sigma^2 \mathbf{I})^{\frac{1}{2}} \mathbf{R}, \tag{6}$$

where Δ_q is a $q \times q$ diagonal matrix with the retained eigenvalues $\lambda_1, ..., \lambda_q$ on its diagonal. $\mathbf{R}$ is an arbitrary rotation matrix and can be dropped for our purposes by setting $\mathbf{R} = \mathbf{I}$. Thus, it is the case then that having calculated PCA, we can use $\hat{\mu}, \sigma_{ML}, \mathbf{W}_{ML}$ to immediately write down a multivariate Gaussian which induces a probability distribution over the image space.

It is unlikely that the image space is Gaussian, not least because it is unlikely to be symmetrical. However, this method for assigning likelihood is still very useful because it provides a consistent way to measure a galaxies distance from the mean average galaxy with respect to the covariance of the data in the low-rank approximated space. We can get some sense of the implications of this likelihood assignment by considering the least and most likely galaxies together. Fig. 8 illustrates the 25 least likely (top) and the 25 most likely (bottom) galaxies. The top image shows anomalous detections and artefacts along with real, but intrinsically rare types of objects. The bottom image shows mainly poorly resolved galaxies with flux magnitude concentrated in a pronounced bulge, likely corresponding to the first two eigengalaxies.

3 APPLICATIONS

In this section, we explore some applications of the probabilistic eigengalaxy framework described above.

3.1 Outlier detection

A key utility of outlier detection is to make discoverable rare phenomena buried in enormous data sets. This may include searches for exotic galaxies and rare objects but also the identification of anomalous detections and pipeline errors. The two-part challenge is first to define an 'outlier' in a way useful to astronomy, and the second is to scale the detection algorithm to the size of the data.

Dutta et al. (2007) implement a distributed version of PCA using random projection and sampling to approximate eigenvectors for the purpose of outlier detection and apply it to the 2MASS (Skrutskie et al. 2006) and SDSS (Fukugita et al. 1996) data sets. In defining an outlier they search for galaxies which are overrepresented by the last eigenvector. Other large scale PCA methods include incremental PCA (Ross et al. 2008), which approximates PCA by processing data in batches commensurate with the available random access memory. Baron & Poznanski (2017) utilize a random forest and fit to discriminate between real and synthetic data.[2] For every pair of objects, they count the number of trees in which each pair is labelled

---

[2]These authors use the flux at each wavelength as a feature set, generating synthetic data by sampling from the marginal distribution of each feature.
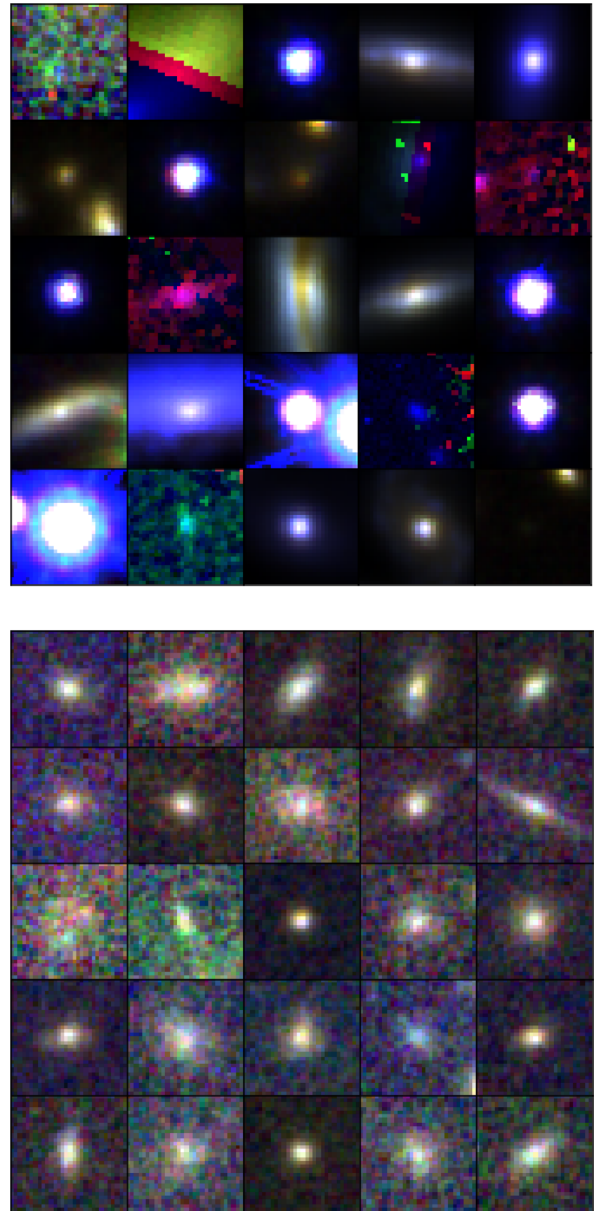

Figure 8. Each 1.8 arcsec thumbnail in both images is an RGB composite of the F160W, F125W, and F814W bands. The top image shows 25 of the least likely galaxies. The image shows anomalous detections and artefacts along with real, but intrinsically rare types of objects. The bottom image shows 25 of the most likely galaxies. The image shows mainly poorly resolved galaxies with flux magnitude concentrated in a pronounced bulge, likely well represented by the first two eigengalaxies.

'real' in the same leaf. The output is an $N \times N$ similarity matrix which is then searched for 'outliers', defined as objects with a large average distance from all other objects.

We focus on a simple and principled definition for an 'outlier', proceeding directly from our eigengalaxy framework. Given the probabilistic interpretation, a natural definition for an outlier is an object with a low likelihood assignment which implies that it is far from the mean given the covariance structure. Note that this same approach could be applied not only to imaging, but also to spectroscopy, light curves, catalogues, and other kinds of data. We define a formal outlier description within the eigengalaxy framework: given a generative factor model $N(\mathbf{W}x + \mu, \sigma^2\mathbf{I})$, an outlier x is an
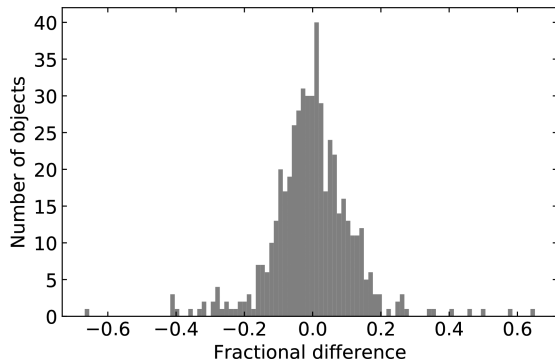
Figure 9. The frequency of the ratio of the sum of deltas (difference between the original and predicted pixel flux) and the sum of the flux of the original images for galaxies with missing data. The body of the distribution extends to ±20 percent error but ∼72 percent of objects are predicted within ±10 percent error, indicating that PPCA can predict missing data in a 2D sense with a high level of fidelity.

object such that $p(\boldsymbol{x}|\mathbf{W}\boldsymbol{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}) < T$ where T is a threshold likelihood, and can be set according to the purpose at hand.

To illustrate the concept with the GZ-CANDELS data set, we use our derived eigengalaxies to assign a log likelihood to every galaxy, using the *score_samples* method on the SKLEARN.DECOMPOSITION.PCA object which implements the Tipping & Bishop (1999) factor model representation to calculate likelihood. We sort the data by likelihood and present in the 25 objects with the lowest likelihood assignment. The result is identical to that given in the top image of Fig. 8 which shows not only anomalous detections and artefacts but also systems that are known to be rare, such as dust lanes which are signposts of recent minor mergers (see e.g. Kaviraj et al. 2012), ongoing mergers (see e.g. Darg et al. 2010), and edge-on spirals which appear to be accreting a blue companion. Outlier detection therefore offers an efficient way to identify examples of rare objects in large surveys.

3.2 Predicting missing data

In many situations one might be missing a particular band, for example due to bad data, partial coverage with a certain bandpass, obliteration by *Starlink* satellite trails, etc. In these cases we can consider how well we can predict the missing data using eigengalaxies. Tipping & Bishop (1999) define an expectation maximization (EM) algorithm for PPCA in the presence of missing data which works by treating missing values as jointly distributed with the latent variables and maximizing the expectation of the joint likelihood function. As a demonstration, we randomly omit a band with equal probability from our data set for 5 percent of rows chosen at random. We use the data as processed in Section 2.2 and an efficient variational EM version of the Tipping & Bishop (1999) algorithm[3] set out in Porta, Verbeek & Kröse (2005) to fit our PPCA model. We use the delta sum over flux sum ratio as in Fig. 5 for the reconstruction error to describe the prediction error. Fig. 9 illustrates the distribution of prediction error. The distribution is roughly symmetrical and is centred on zero. The body of the distribution extends to ±20 percent error but ∼72 percent of objects are accounted for within ±10 percent error showing that high fidelity prediction is possible for most objects. Fig. 10 provides

some examples of predicted data for different bands. It is noteworthy that by using information contained in the eigengalaxies, PPCA is successful in estimating total flux even when it bears no resemblance to that of the other bands. The ability to predict missing images offers a route to 'filling in' missing data, such as predicting photometric data points which are absent in the observations in order to reconstruct missing parts of a galaxy's spectral energy distribution.

3.3 Searching for galaxies similar to an exemplar

Given a survey with a large number of objects with diverse variety, and an interest in galaxies of a specific kind, it is useful to be able to present an exemplar galaxy and use it to quickly search for all other galaxies with similar features. The utility and suitability of a similarity search for any particular use case will depend primarily on how the objects are being described, and how the similarity between their descriptions is being calculated. For example, Protopapas et al. (2006) use cross-correlation as a proxy for similarity between light curves for the purpose of detecting outliers, whereby light curves with the lowest average similarity are defined as outliers. Sart et al. (2010) utilize dynamic time warps (Berndt & Clifford 1994) to measure the similarity between light curves. Hocking et al. (2018) compare various measures, including Euclidean distance and Pearson's correlation coefficient, and use cosine distance to measure similarity from a description generated by growing neural gas prior to hierarchical clustering.

In Section 2.3, we explained that the suitability of the image space can be inspected by considering the similarity of the nearest neighbours for any given galaxy, and we provided some examples in Fig. 2. 'Similarity search' works on precisely the same principle. For any given reference galaxy, the rest of the galaxies are ordered according to their Euclidean distance from it in the rank reduced space. One may then consider each galaxy in order of similarity. The method is principled because it is established in Section 2 that the dimensions of the space are not arbitrary, that they are contrived to be relevant to morphology and that the number of dimensions retained pertains directly to how well similarity neighbourhoods are preserved in the image space. The method is also straightforward to extend to the case where multiple reference objects are desirable, by calculating for each other object the median distance from the reference points and then ordering as before. Fig. 11 illustrates searches for spirals, edge-on spirals, ellipticals, and recent mergers. It shows very good likeness in the nearest neighbourhoods as may be expected given previous tests.

3.4 Unsupervised clustering

Unsupervised clustering is particularly useful for astronomy data sets because it provides a method for investigating very large collections of objects efficiently as long as the image space has been credibly constructed. If the clustering method is effective at grouping objects with similar features together, then one can study a data set by characterizing its morphological centres rather than examining each object separately. There is some precedence for this in astronomy. For example, Martin et al. (2020) follow Hocking et al. (2018) in using growing neural gas and hierarchical clustering directly on pixel data to identify structurally distinct clusters. Almeida et al. (2010) and Almeida & Prieto (2013) utilize k-means to classify spectra from SDSS into fewer base types. Valenzuela & Pichara (2018) use k-
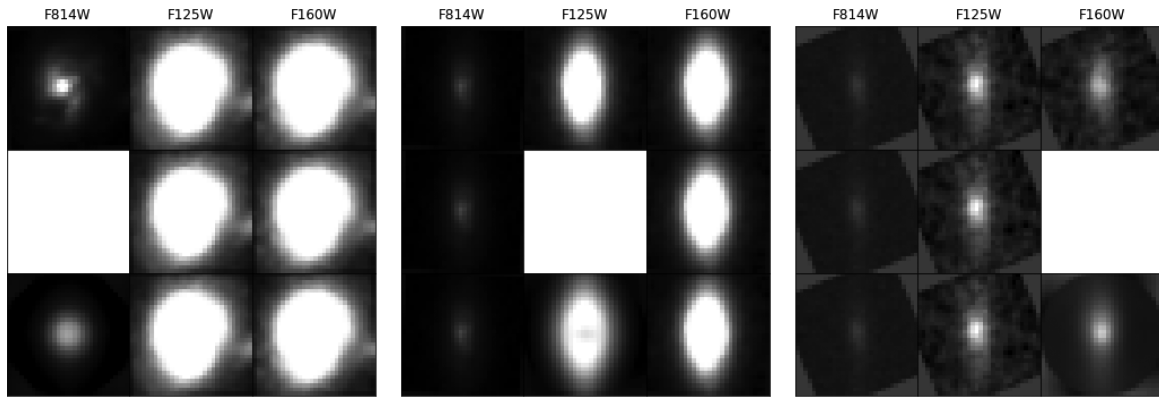
[3]Available in Python package PYPPCA.

Figure 10. An example of using PPCA for image imputation. In each figure, the first row shows the original galaxy across all three bands, the second row shows the same galaxy with a random band censored, the third row shows the censored galaxy predicted by PPCA. It is noteworthy that by using information contained in the eigengalaxies, PPCA is successful in estimating brightness correctly even when it bares no resemblance to that of the other bands, as illustrated in the left most image.
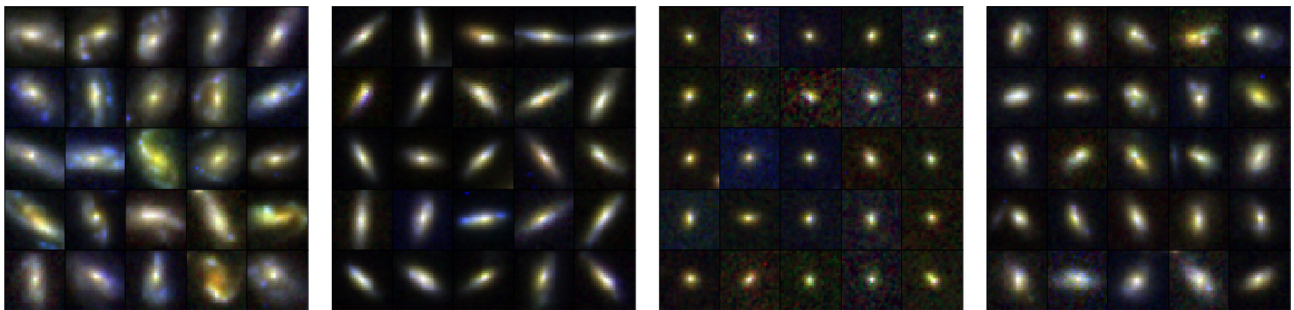

Figure 11. Examples of similarity searches. In each 1.8 arcsec thumbnail image (an RGB composite of the F160W, F125W, and F814W bands), the exemplar galaxy is given in the top left, followed by 24 of its nearest neighbours by Euclidean distance in the 12D eigengalaxy space. The image shows searches for spirals, edge-on spirals, ellipticals, and recent mergers from left to right, respectively.

medoids[4] to cluster and map sequences of light-curve segments to variational trees.

The eigengalaxy framework enables a simple unsupervised clustering treatment. We can create a distance matrix (an $N \times N$ matrix in which each cell indicates a distance between object j at row j and object i at row i) by calculating the Euclidean distance between every galaxy. The distance matrix provides an input for a broad range of unsupervised clustering algorithms. This provides us with an opportunity to define and discover groups of galaxies based on the similarity of their multiband morphologies as encoded by the eigengalaxies. Amongst the central problems in clustering is how the similarity between objects is defined and how many clusters there are. In our case, similarity is defined by the Euclidean distance between objects, and it turns out that there is a robust way to conceive the clustering problem such that the number of clusters is automatically decided. Let $\{d_{i,j}\}$ be an $N \times N$ distance matrix in which each cell indicates the Euclidean distance between object j at row j and object i at row i. The objective then is to choose a set of exemplar objects such that the sum of the distances between each object and its closest exemplar is minimized. Formally, for some set of galaxy points $s \in S$:

$$\min_{\{q_i\}_{i=1}^m} \left(\sum_{s \in S} \min_i d_{s,q_i} + \sum d_{q_i,q_i} \right), \qquad (7)$$

where $q_i, ..., q_m \in S$ are m examplars. The second summation acts as a regularizer barring trivial solutions (such as picking every point as its own exemplar), and regulating the number of exemplars that are chosen. The result is the selection of a set of objects which we may call 'cluster exemplars' and clustering is achieved by labelling objects according to their closest exemplar. The benefit of this formulation is that it is exact and will result in both the number of centres and their membership. The exact problem is NP-hard (i.e. cannot necessarily be solved in reasonable time) (Komodakis, Paragios & Tziritas 2009) but there are myriad relaxations and approximations that deliver near optimal solutions in empirical tests. One such approximation is known as affinity propagation (Dueck 2009) (AP) which minimizes a similar equation to equation (7) using message passing over factor graphs.

To demonstrate this, we produce a distance matrix for the GZ-CANDELS data set as outlined above. We use AP which produced 536 clusters from the total sample of 10243 objects, with clusters varying in size from 1 to 316 with a median size of 4. Affinity propagation has a tunable 'preference' parameter that allows the coarseness of clustering to be adjusted; here we use the default value (the median of the distance matrix). Fig. 12 illustrates samples of galaxies from four morphological clusters. It is noteworthy that affinity propagation is an exemplar-based algorithm, so that each cluster is actually characterized by an exemplar galaxy. Thus, in this instance the whole data set is summarized by 536 exemplars (5 percent of the sample) which could, in principle, be sorted by rarity (see Section 3.1) and then viewed as one large image. This
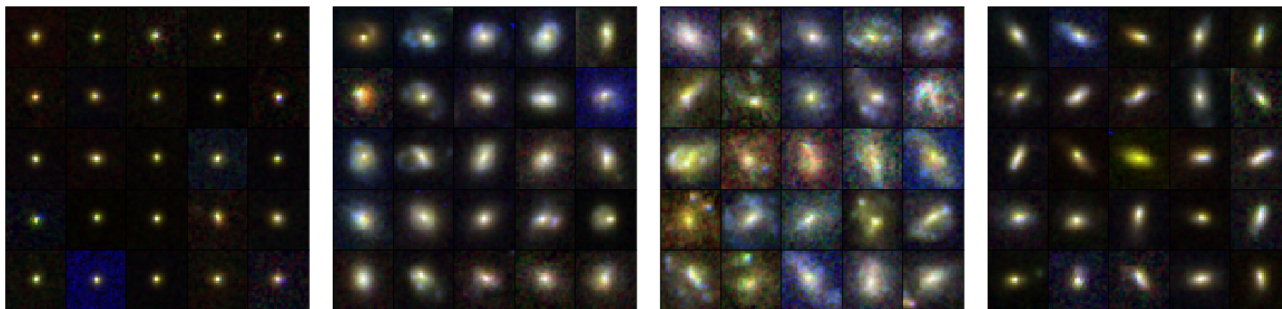
---

[4]In k-medoids, data points become cluster centres, unlike k-means where the cluster centre is not necessarily correspondent to a data point.

Figure 12. Composite image samples of four morphological clusters from a total of 536 created using affinity propagation clustering of a distance matrix defined by pairwise Euclidean distances in 12D eigengalaxy space. Each 1.8 arcsec thumbnail image is an RGB composite of the F160W, F125W, and F814W bands.

massive reduction in the scale of the full data set highlights the efficiency gains that can be made by using this method for exploring the extremely large imaging surveys of the future.

3.5 Combining methods

The methods described above utilize the same framework and can be used together to create additional capabilities simply based on the eigengalaxy weights in the score matrix. There are many possible combinations but we highlight three:

(i) Missing value prediction → all methods. Predicting missing band values, dead pixels, etc. can be used to 'complete' data so that it may be considered on par with the rest of the data. It may then be used with any of the other methods.

(ii) Outlier detection → similarity search / clustering / classification. We may discover an interesting outlier (e.g. a rare object such as a gravitationally lensed galaxy) of which we would like to find many more examples (similarity search). We may want to cluster the outliers to determine a self-similar morphological clusters. We may also want to label a training set of outliers, train a classifier, and then select only the outliers of interest.

(iii) Clustering → missing value prediction. To make the prediction of missing values more accurate, we may first generate eigengalaxies using the available bands and use it to create a clustering. This produces a set of self-similar classes to which missing values have more in common than the whole population. For each cluster, we could then separately derive eigengalaxies and predict missing values for objects within that cluster.

3.6 Big data

The GZ-CANDELS data set is relatively small and can be easily processed. In this section, we consider problems and adaptations which may be necessary to use these methods with extremely large data sets, such as LSST (Robertson et al. 2017), *Euclid* (Refregier et al. 2010), and the Square Kilometre Array (Weltman et al. 2020) which may contain billions of observed objects. We suggest that the eigengalaxy framework offers a novel solution to processing such big data by making it possible to find outliers, search for and cluster objects using only the reduced score matrix form of a data set. For example, if LSST was to provide a corresponding score matrix, eigengalaxies, and galaxy likelihoods, all the methods described here would be available based on a small fraction of the full data set.

An important consideration in making the methods presented here feasible for big data is the number of eigengalaxies required to achieve a practically useful level of explained variance. Aggarwal,

Hinneburg & Keim (2001) provide a deeper discussion on common issues for algorithms operating in high dimensional space. In our case, we would expect the following problems given too many eigengalaxies:

(i) The score matrix form of the data may itself be too large to be useful. In our case, as the number of eigengalaxies approaches 10243, the total size of the data converges on to the total size of the cut-out collection.

(ii) Euclidean distance loses interpretability and the power to distinguish objects as the number of dimensions grows.

(iii) Some techniques like fully calculating a distance matrix or calculating the similarity of every object with reference to an exemplar would not be computationally scalable.

Taking LSST as an example, we would expect that the number of eigengalaxies required to explain, say, 96 percent variance to be more than in our GZ-CANDELS experiments for two primary reasons: (i) it is a deeper survey and it therefore observes more low mass galaxies, the morphological mix of which is yet unknown (e.g. Martin et al. 2019), and (ii) we would expect bright galaxies to have more detail such as extended debris and tidal features (e.g. Duc et al. 2011) and therefore exhibit more variance. However, the additional sources of variance are limited, and we would expect a reasonable asymptotic number of eigengalaxies to emerge for a useful level of explained variance at relatively small sample sizes.

The kind of variance that needs to be captured and the sufficient ratio of it to retain ultimately depends on the intended application. If the number of eigengalaxies is too high, we could explore various ways to reduce the variance not required for the intended purpose in order to stay within an eigengalaxy quota at a useful explained variance ratio. Potentially applicable techniques include using smaller cut-outs, down-sampling pixels, using fewer bands, combining bands, stretching the range of the flux densities to make differences less subtle, and partitioning the data by some other variable (e.g. brightness, location, etc.) and then conducting the analysis per partition.

If the number of eigengalaxies was workable from a data size perspective but a problem for Euclidean distance then we could use other distance measures more robust in high dimensional space. Finally, in every case we would need to replace exhaustive similarity searches with more scalable methods such as the nearest neighbour algorithm (Beis & Lowe 1997), and affinity propagation would need to be replaced with a less expensive clustering method such as *k*-means (or *k*-medoids if Euclidean distance was unacceptable).

4 CONCLUSIONS AND SUMMARY

We have demonstrated how galaxy cut-outs can be projected to a linear similarity preserving image space. We used PCA for optimal low-rank approximation and termed the resultant orthogonal basis vectors 'eigengalaxies'. We further showed how a formal equivalence of PCA to latent factor models with isotropic variance enables a probabilistic interpretation of the image space. This offers an empirical and non-categorical approach to characterizing galaxy morphology. Using a sample of 10 243 galaxies from the Galaxy Zoo–CANDELS survey, chosen to lie in the deep GOODS–S field, we use three-band *HST* ACS and WFC3 imaging to derive just 12 eigengalaxies that are sufficient to provide robust similarity preservation in the resultant image space. We explored four applications of our framework:

(i) we have shown how the probabilistic interpretation can be used to assign likelihoods and identify outliers as objects with a low likelihood.

(ii) we have shown how PPCA can be used to predict missing values in imaging, for example due to bad data or partial coverage in a given band.

(iii) we have shown how the projection of eigengalaxies on to a 12D space naturally facilitates similarity searches, where galaxies can be sorted relative to their Euclidean distance from an exemplar, thus quickly returning samples of galaxies that are morphologically similar to the exemplar as defined by their eigengalaxy components.

(iv) we have shown how the Euclidean distances between galaxies in the 12D space can be compiled into a distance matrix that can provide the input for unsupervised clustering algorithms to discover groups of similar objects. We have demonstrated this using affinity propagation to show how morphologically similar groupings can be identified in large samples.

We have described how the methods may be used together in more advanced use cases and how working in eigengalaxy space may present a novel solution to outlier detection, as well as search and clustering problems in forthcoming massive imaging data sets such as LSST. We argue that these results and illustrations underscore the suitability of our PCA based probabilistic eigengalaxy framework for the study of morphology, especially in the era of big data astronomy, where representational efficiency and relevance will pay dividends.

DATA AVAILABILITY

Code for all results and figures can be found at https://emiruz.com/eigengalaxies.

[5]http://www.astropy.org

REFERENCES

Abraham R. G., Valdes F., Yee H. K. C., van den Bergh S., 1994, ApJ, 432, 75
Aggarwal C. C., Hinneburg A., Keim D. A., 2001, International conference on database theory, Springer, Berlin, Heidelberg, p. 420
Almeida J. S., Prieto C. A., 2013, ApJ, 763, 50
Almeida J. S., Aguerri J. A. L., Munoz-Tunón C., De Vicente A., 2010, ApJ, 714, 487
Anderson B., Moore A., Connolly A., Nichol R., 2004, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. p. 40
Astropy Collaboration, 2013, A&A, 558, A33
Bamford S. P. et al., 2009, MNRAS, 393, 1324
Baron D., Poznanski D., 2017, MNRAS, 465, 4530
Beck M. R. et al., 2018, MNRAS, 476, 5516
Beis J. S., Lowe D. G., 1997, Proceedings of IEEE computer society conference on computer vision and pattern recognition, IEEE, p. 1000
Berndt D. J., Clifford J., 1994, KDD workshop, Vol. 10, Seattle, WA, USA. p. 359
Bluck A. F. L., Mendel J. T., Ellison S. L., Moreno J., Simard L., Patton D. R., Starkenburg E., 2014, MNRAS, 441, 599
Bundy K., Ellis R. S., Conselice C. J., 2005, ApJ, 625, 621
Cappellari M. et al., 2011, MNRAS, 416, 1680
Cheng T. Y., et al.,2020, MNRAS, 493, 4209
Codis S., Pichon C., Devriendt J., Slyz A., Pogosyan D., Dubois Y., Sousbie T., 2012, MNRAS, 427, 3320
Conselice C. J., 2003, ApJS, 147, 1
Conselice C. J., 2006, ApJ, 638, 686
Darg D. W. et al., 2010, MNRAS, 401, 1043
de Vaucouleurs G., 1948, Ann. Astrophys., 11, 247
de Lapparent V., Bellanger C., Arnouts S., Mathez G., Mellier Y., Mazure A., 1993, Messenger, 72, 34
De La Calleja J., Fuentes O., 2004, MNRAS, 349, 87
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Djorgovski S. G., Mahabal A. A., Donalek C., Graham M. J., Drake A. J., Moghaddam B., Turmon M., 2012, 2012 IEEE 8th International Conference on E-Science, IEEE, p. 1
Dressler A. et al., 1997, ApJ, 490, 577
Duc P.-A. et al., 2011, MNRAS, 417, 863
Dueck D., 2009, Affinity propagation: clustering data by passing messages. Citeseer, Toronto, Canada
Dutta H., Giannella C., Borne K., Kargupta H., 2007, Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, p. 473
Eckart C., Young G., 1936, Psychometrika, 1, 211
Fukugita M., et al., 1996, Technical report, The sloan digital sky survey photometric system. SCAN-9601313
Galaz G., de Lapparent V., 1998, A&A, 332, 459
Gloub G. H., Van Loan C. F., 1996, Matrix Computations, Johns Hopkins University Press, Baltimore, USA
Goulding A. D. et al., 2018, PASJ, 70, S37
Grogin N. A. et al., 2011, ApJS, 197, 35
Hocking A., Geach J. E., Sun Y., Davey N., 2018, MNRAS, 473, 1108
Hubble E. P., 1926, ApJ, 64, 321
Huertas-Company M. et al., 2015, ApJS, 221, 8
Jackson R. A., Martin G., Kaviraj S., Laigle C., Devriendt J. E. G., Dubois Y., Pichon C., 2019, MNRAS, 489, 4679
Jackson R. A., Martin G., Kaviraj S., Laigle C., Devriendt J., Dubois Y., Pichon C., 2020, MNRAS, 494, 5568
Jackson R. A. et al., 2020b, preprint (arXiv:2007.06581)
Kaviraj S., 2010, MNRAS, 406, 382
Kaviraj S., 2014, MNRAS, 440, 2944
Kaviraj S. et al., 2007, ApJS, 173, 619
Kaviraj S. et al., 2012, MNRAS, 423, 49
Kaviraj S., Devriendt J., Dubois Y., Slyz A., Welker C., Pichon C., Peirani S., Le Borgne D., 2015, MNRAS, 452, 2845
Kaviraj S., Martin G., Silk J., 2019, MNRAS, 489, L12

Koekemoer A. M. et al., 2011, ApJS, 197, 36
Kohonen T., 1990, Proc. IEEE, 78, 1464
Komodakis N., Paragios N., Tziritas G., 2009, Advances in neural information processing systems, NeurIPS. p. 865
Lackner C. N., Gunn J. E., 2012, MNRAS, 421, 2277
Lahav O. et al., 1995, Science, 267, 859
Le Borgne J.-F. et al., 2003, A&A, 402, 433
Li C., Wang T.-G., Zhou H.-Y., Dong X.-B., Cheng F.-Z., 2005, AJ, 129, 669
Lintott C. et al., 2011, MNRAS, 410, 166
Lotz J. M., Primack J., Madau P., 2004, AJ, 128, 163
Ma Z. et al., 2019, ApJS, 240, 34
Mager V. A., Conselice C. J., Seibert M., Gusbar C., Katona A. P., Villari J. M., Madore B. F., Windhorst R. A., 2018, ApJ, 864, 123
Martin G. et al., 2018a, MNRAS, 476, 2801
Martin G., Kaviraj S., Devriendt J. E. G., Dubois Y., Pichon C., 2018b, MNRAS, 480, 2266
Martin G. et al., 2019, MNRAS, 485, 796
Martin G., Kaviraj S., Hocking A., Read S. C., Geach J. E., 2020, MNRAS, 491, 1408
Menanteau F., Ford H. C., Motta V., Benítez N., Martel A. R., Blakeslee J. P., Infante L., 2006, AJ, 131, 208
Menou K., 2018, MNRAS, 489, 4802
Naim A., Ratnatunga K. U., Griffiths R. E., 1997, ApJS, 111, 357
Odewahn S. C., Cohen S. H., Windhorst R. A., Philip N. S., 2002, ApJ, 568, 539
Ostrovski F. et al., 2017, MNRAS, 465, 4325
Peirani S., Crockett R. M., Geen S., Khochfar S., Kaviraj S., Silk J., 2010, MNRAS, 405, 2327
Peth M. A. et al., 2016, MNRAS, 458, 963
Porta J. M., Verbeek J. J., Kröse B. J., 2005, Auton. Robots, 18, 59
Postman M. et al., 2005, ApJ, 623, 721
Price-Whelan A. M. et al., 2018, AJ, 156, 123
Protopapas P., Giammarco J., Faccioli L., Struble M., Dave R., Alcock C., 2006, MNRAS, 369, 677

Refregier A., Amara A., Kitching T. D., Rassat A., Scaramella R., Weller J., 2010, preprint (arXiv:1001.0061)
Robertson B. E. et al., 2017, preprint (arXiv:1708.01617)
Ross D. A., Lim J., Lin R.-S., Yang M.-H., 2008, Int. J. Comput. vis., 77, 125
Rubin D. B., Thayer D. T., 1982, Psychometrika, 47, 69
Ryan R. E. J. et al., 2012, ApJ, 749, 53
Sart D., Mueen A., Najjar W., Keogh E., Niennattrakul V., 2010, 2010 IEEE International Conference on Data Mining, IEEE. p. 1001
Scarlata C. et al., 2007, ApJS, 172, 406
Schawinski K. et al., 2014, MNRAS, 440, 889
Schawinski K., Zhang C., Zhang H., Fowler L., Santhanam G. K., 2017, MNRAS, 467, L110
Sérsic J. L., 1963, Bol. Asoc. Argentina Astron. Plata Argentina, 6, 41
Simard L. et al., 2002, ApJS, 142, 1
Simmons B. D. et al., 2014, MNRAS, 445, 3466
Simmons B. D. et al., 2016, MNRAS, stw2587
Simmons B. D. et al., 2017, MNRAS, 464, 4420
Skibba R. A. et al., 2009, MNRAS, 399, 966
Skrutskie M. et al., 2006, AJ, 131, 1163
Smethurst R. J. et al., 2015, MNRAS, 450, 435
Soo J. et al., 2018, MNRAS, 475, 3613
Strateva I. et al., 2001, AJ, 122, 1861
Tipping M. E., Bishop C. M., 1999, J. R. Stat. Soc. B, 61, 611
Valenzuela L., Pichara K., 2018, MNRAS, 474, 3259
Walmsley M., Ferguson A. M. N., Mann R. G., Lintott C. J., 2019, MNRAS, 483, 2968
Weltman A. et al., 2020, Publ. Astron. Soc. Austr., 37, e002
Whittle P., 1952, Scand. Actuar. J., 1952, 223
Wild V. et al., 2014, MNRAS, 440, 1880
Willett K. W. et al., 2015, MNRAS, 449, 820
Willett K. W. et al., 2017, MNRAS, 464, 4176
Wollaeger R. T. et al., 2018, MNRAS, 478, 3298

This paper has been typeset from a TEX/LATEX file prepared by the author.