Eigengalaxies: galaxy morphology as a linear image space and its applications

Submitted to the University of Hertfordshire in partial fulfilment
of the requirement of the degree of MSc by Research

Emir Uzeirbegovic

June 19, 2020

Abstract

In this thesis I contextualise the history of morphology as underpinned by Hubble's scheme, discrete in nature, and deeply connected to theories of galaxy formation history. I set out in contrast, to describe a purely empirical morphology, continuous in nature, in which surveys become image spaces and galaxies become points, the meaning of which is sought by the quantifiable differences of their relative spatial positions. I show how an image space can be robustly constructed and then build upon it to illustrate important applications such as approximating surveys with small samples, detecting outliers, clustering, similarity search and missing data prediction.

The thesis proceeds as follows. Section 1 briefly surveys the importance, genesis and recent history of galaxy morphology. It also lays out the objectives of the thesis and information about the survey data which I have used. Section 2 describes how galaxy images can be processed and projected to a defensible low dimensional space in a morphology preserving way. Several analyses are then performed to test the fidelity of the projection. It is also shown how the image space can be given a probabilistic interpretation. Section 3 discusses methods for approximating surveys by reducing the number of objects under consideration. The section starts by describing simple random sampling and its limitations. It then shows how means and covariances can be used to summarise image spaces and how differences between image spaces can be quantified using the Kullback-Leibler divergence. This concept is then used to apply "leverage scores" sampling as a means to use information from the galaxy population to create a weighted sampling scheme which preserves mean and covariance better than random sampling and therefore enables much smaller representative samples. I also motivate and describe a cutting edge "coresets" methodology which I intend to more fully explore in future work. Section 4 demonstrates parsimonious applications of the image space framework to common use cases such as clustering, similarity search and outlier detection. It is a modified and abridged version of a paper to be published in MNRAS with some modification. Finally, section 5 draws summary conclusions and highlights important directions for the future.

Contents

1 Introduction

1.1 Morphology in astronomy

"Galaxy morphology" refers to the distribution of light in galaxies. Morphology strongly correlates with the physical properties of a galaxy, such as its stellar mass (e.g. Bundy et al. 2005), star formation rate (e.g. Bluck et al. 2014, Smethurst et al. 2015, Willett et al. 2015), surface brightness (e.g. Martin et al. 2019), rest frame colour (e.g. Bamford et al. 2009, Skibba et al. 2009, Strateva et al. 2001) 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. Jackson et al. 2020, Martin, Kaviraj, Devriendt, Dubois & Pichon 2018). 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 sign posts 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, Kaviraj, Volonteri, Simmons, Devriendt, Lintott, Smethurst, Dubois & Pichon 2018). In a similar vein, morphological details such as extended tidal features suggest recent mergers and/or interactions (e.g. Jackson et al. 2019, Kaviraj 2014, Kaviraj et al. 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. Kaviraj et al. 2015, Schawinski et al. 2014). The morphological analysis of galaxy populations, especially in the large surveys that underpin our statistical understanding of galaxy evolution, is therefore of fundamental importance.

1.2 Genesis of the study of galaxy morphology

The genesis of the study of galaxy morphology can be traced back to the visual classifications by Hubble (1926), which divided galaxies into ellipticals, lenticulars, spirals and irregular shapes. The classifications are typically presented as a bifurcating sequence known as the "tuning fork" diagram. The sequence starts with the varying degrees of ellipticals (E0, E3, E5, E7), reaches the lenticulars (S0), then bifurcates into non-barred spirals (Sa, Sb, Sc) and barred spirals (SBa, SBb, SBc). Many extensions and modifications have been proposed to Hubble's system (De Vaucouleurs 1959, Kormendy & Bender 1996, Van den Bergh 1976). As an example, Figure 1 illustrates modifications to Hubble's scheme proposed in Van den Bergh (1976) in which the lenticular range is extended into a parallel sequence, and a new branch of "anemic" spirals is added. The sequence order is determined by the bulge-to-disk ratio.

Hubble's scheme still underpins the broad morphological classes into which galaxies are split in modern studies, and relatively recent massively distributed citizen science systems like Galaxy Zoo have revolutionised the use of Hubble's classifications on survey datasets (e.g. Lintott et al. 2011, Simmons et al. 2017, Willett et al. 2017), discovering amongst other things, that many central implications of Hubble's model (e.g. the relationship between galaxy bulge and spiral windings) could not be confirmed.

1.3 Morphological classification

The "eyeball" (human expert) classifier is still arguably the gold standard, but even with distributed schemes like Galaxy Zoo, it is not always possible to scale the eyeball to the number of galaxies in a survey, especially if the morphology of interest is specialised or requires domain knowledge to spot. To this end, early efforts often sought to automate Hubble-like classifications or provide measures which correlated with them. Such methods can be roughly partitioned into parametric and non-parametric approaches. In the parametric approach, aspects of a galaxies morphology are fitted to parametric models and these models may in turn be combined together to create classification systems. Two such examples are Sérsic profile (Sérsic 1963) and Nuker profile which seek to index the relationship between intensity and distance from center, and summarise the surface brightness profile of galactic nuclei respectively. Peng et al. (2002) describe their

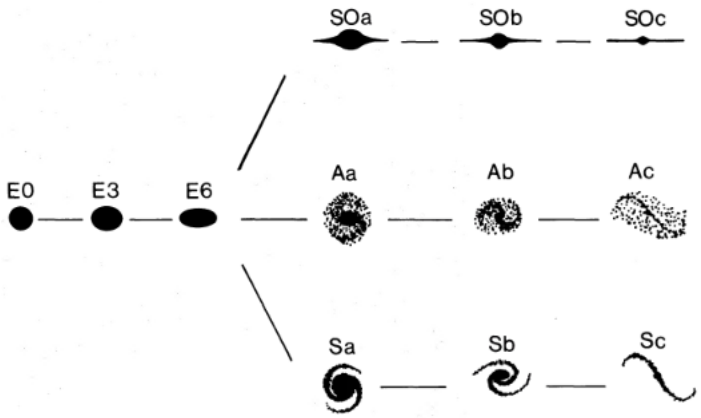


Figure 1: Galaxy classification as proposed in Van den Bergh (1976). Normal spirals (Sa, Sb, Sc), anemic spirals (Aa, Ab, Ac) and lenticulars (S0a, S0b, S0c) order by disk-to-bulge ratio.

canonical optimisation algorithm called GALFIT aimed at simultaneously fitting many parametric models to large galaxy images.

Amongst the problems with the parametric approaches are violations of distributional assumptions and dealing with irregularities. For example, Lotz et al. (2004) note that irregular, tidally disturbed, and merging galaxies violate the typical symmetry and smoothness operating constraints and lead parametric models to break down. Graham et al. (2003) show that none of the Nuker model parameters are robust and that they vary with the fitted radial extent often by more than 100%.

Non-parametric approaches aim to have fewer distributional assumptions and therefore be more robust to irregularity. Similarly to the parametric approach, systems are made up of simpler models which aim to describe specific phenomena. One such system called "CAS" is presented in Conselice (2003). The authors argue that three non-parametric models respectively measuring concentration (C), asymmetry (A) and clumpiness (S) taken together are enough to robustly classify resolved galaxies into a Hubble type scheme. Lotz et al. (2004) argue that CAS is implicitly dependent on symmetries and requires careful calibration of hyper-parameters (such as smoothing length) in order to work well. They introduced the Gini coefficient and M_{20} as more robust counterparts.

1.4 Machine learning in morphology

By the early 90's machine learning techniques were already being applied to morphology. Given labelled samples and a set of features, machine learning methods enabled relationships to be learned automatically, shifting effort away from analytical model building, to machine model training. For example, Weir et al. (1995) used 8 attributes calculated from the light distribution of galaxies, and a set of labelled galaxies to train a decision tree classifier. Storrie-Lombardi et al. (1992) and Odewahn et al. (1992) used a similar methodology with different image features to train neural network classifiers from labelled samples.

An informative (but false) partitioning of machine learning applications in astronomy may be sought in terms of those models utilising ground truth to regress/classify (supervised) and those which aim to partition/cluster/summarise without recourse to ground truth (unsupervised). A further useful (but false) dichotomy may be drawn between those models which are applied to pre-calculated explanatory features (explicit), and those to which calculation of explanatory features is implicit to the model (implicit). Astronomy has many examples in the supervised, explicit category (Goulding et al. 2018, Ostrovski et al. 2017, Weir et al. 1995) especially those using neural networks (Odewahn et al. 1992, Storrie-Lombardi et al. 1992). Supervised, implicit models have been rare until recently where convolutional neural networks (Krizhevsky et al. 2012) (CNNs), which are particularly well suited to image inference classification, became popular. Examples of CNNs in astronomy include Dieleman et al. (2015), Huertas-Company et al. (2015) and Cheng et al. (2019).

The unsupervised category is fairly sparse in astronomy. An example in the unsupervised explicit category

is Naim et al. (1997) in which 4 morphological parameters are calculated for a set of galaxies on the basis of which a self organising map (Kohonen 1990) (SOM) is fitted to produce a topology. A unsupervised implicit example is given in Hocking et al. (2018) in which the features are derived from the reduction and clustering of image "patches" using the Growing Neural Gas algorithm (Fritzke 1995). Patches are then used to create descriptive vectors for each galaxy which are in turn clustered.

1.5 Objectives and outline

In this thesis I aim to develop a view of morphology as a continuous space that may be thought about and characterised without recourse to a categorical classification of morphological types or an underpinning theory regarding the correspondence between formation history and present-day morphology. I aim to do this by projecting galaxy images into a defensibly constructed, low-rank linear space in which distances and proximities are morphologically significant, and to which a vast range of linear algebra and statistics-oriented methods may be applied. I will be solely interested in "spatial morphology" by which I mean the pattern and pixel distribution that gives a galaxy its visual appearance as it may be seen in grayscale. I also aim to demonstrate how this morphological "image space" serves as a basis and naturally extends to the characterisation, comparison and sampling of datasets which is so vital in Big Data use cases, and how the image space can be used to find outliers, cluster data, predict missing values and perform similarity searches, amongst other things.

The thesis proceeds as follows. Section 1 has so far explained morphology as largely underpinned by Hubble's scheme, discrete in nature, deeply connected to galaxy formation history, and initially driven by methods aimed at describing aspects of galaxies which are thought to be crucial in their formation history. The section also explains how machine learning had enabled a new approach to modelling by shifting focus away from analytical formulation to model training and brought with it a diversity of empirically centred machine learning models. This section also describes the data that will be used in this thesis, how its made useful and details about its accompanying catalogues. Section 2 describes how galaxy images can be processed and projected to a defensible low dimensional space in a morphology preserving way. Several analyses are then performed to test the fidelity of the projection. It is also shown how the image space can be given a probabilistic interpretation. Section 3 discusses methods for approximating surveys by reducing the number of objects under consideration. The section starts by describing simple random sampling and its limitations. It then shows how means and covariances can be used to summarise image spaces and how differences between image spaces can be quantified using the Kullback-Leibler (Kullback & Leibler 1951) divergence. This concept is then used to motivate "leverage scores" sampling as a means to use information from the galaxy population to create a weighted sampling scheme which preserves mean and covariance better than random sampling and therefore enabled smaller representative samples. The results of applying this method on our image space is discussed. I also motivate and describe a cutting edge "coresets" methodology which I intend to more fully explore in future work. Section 4 demonstrates parsimonious applications of the image space framework to common use cases such as clustering, similarity search and outlier detection. It is a modified and abridged version of a paper to be published in MNRAS with some modification (Uzeirbegovic et al. 2020). Finally, section 5 draws summary conclusions from the analysis performed and highlights important directions for the future.

1.6 Optical data and catalogues used

This thesis aims to develop methodology which may be applied to optical surveys broadly. Throughout the thesis I use the *HST* CANDELS (Grogin et al. 2011, Koekemoer et al. 2011) survey because it 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, I focus on GOODS-S, one of the deep tier (at least four-orbit effective depth) fields. I select the sample of 11,469 galaxies present in the 'Galaxy Zoo: CANDELS' (GZ-CANDELS) GOODS-S catalogue (Simmons et al. 2016), which also 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). The motivation for using the GZ-CANDELS catalogue is so that in future work, the results can be compared to the morphological and structural classifications provided by the Galaxy Zoo (GZ) citizen

science project. GZ asks participants to classify optical imagery into a pre-determined taxonomy. The GZ-CANDELS catalogue aggregates these classifications into fractions of votes and provides indicators for several morphological features.

For each object in the catalog I extracted a 40×40 ($2.4''$) pixel cut-out from the original CANDELS FITS files, using the catalogued sky coordinate as a centroid. ACS images are down-sampled by a factor of two to match the pixel scale of the WFC3 images so that for each band I obtained a 40×40-pixel image at the position of each galaxy.

2 Morphology as an image space

2.1 Introduction

We may gain considerable flexibility by thinking about morphology as a continuous space which does not presume the existence of sharp dividing lines and categorical differences. In this thesis, I begin by projecting images to points in a j-dimensional linear space $\mathbb{R}^j$. This new "image space" facilitates the definition of morphology as the relationships between points in space, to which an arsenal of tools from linear algebra and statistics may be brought to bear.

Let an image be a 3 dimensional array $N \times M \times B$ where N and M are the number of rows and columns in the image and B is the number of bands. Thus, each cell (n, m, b) represents the flux density at row n, column m and band b. The total number of cells is therefore given by $N \cdot M \cdot B$ and a trivial projection to an image space could be achieved by flattening the whole array into a single long vector such that each image can be represented as $v_i \in \mathbb{R}^{N \cdot M \cdot B}$. If $v_i = v_j$ then it must be that $i = j$, but what about when they are not the same? What do the flattened dimensions represent, are they all meaningful, how do we compare them and what do the differences mean?

2.2 Variance of interest

Our linear space can usefully represent morphology if it can make morphologically meaningful the differences between image vectors. However, the vectors presently carry a lot of information including band magnitudes, relative brightness, spatial correlations and background. I am only interested in the spatial morphology which I have defined as the pattern and pixel distribution that gives a galaxy its visual appearance as it may be seen in grayscale. In this thesis, I have employed the following preprocessing steps to select primarily for spatial variance so that differences in spatial morphology may be modelled to the exclusion of other effects.

Confusion noise and background clipping

Sometimes attenuating dust and debris is crucial to morphology (such as in spiral galaxies for example), however a typical thumbnail of a galaxy often also contains superfluous confusion noise and "background" brightness beyond that which is useful for determining morphology. The shape and relationship of these artefacts to the galaxy may have an arbitrary effect on considerations of its morphology so it makes sense to try and reduce them. Confusion noise is light from other galaxies visible on the periphery, too dim to be resolved but not zero. Meanwhile, I use the fact that the images are centered on the brightest point and that in a 40×40 thumbnail, most pixels are not of the galaxy, to define as "background" those pixels which are close to the median brightness. More formally, for each band, the survey images were flattened into 1D arrays and concatenated, and the quantiles .25,.5,.75 were calculated. On the assumption that confusion noise is Gaussian with parameters μ, σ^2, its distribution can be fitted by recalling that in the normal case $\mu = median(x), \sigma = \frac{iqr(x)}{2\Phi^{-1}(.75)}$, where x is the pixel array, Φ^{-1} is the quantile distribution and iqr is the interquartile range function. Since the median and IQR are robust statistics with very high breakdown points (i.e. 25% and 50% of the data must be contaminated for the IQR and median to fail respectively), I can treat resolved galaxies as contamination and ignore them using robust statistics in order to fit confusion noise. We can specify the threshold as the percentile that covers at least 99% of the data: $\Phi^{-1}(.99 \mid \mu, \sigma)$, where 99 is a large arbitrary constant. For my data, $\mu, \sigma = 0$ for both bands F160W, F125W hence their thresholds are 0. For F814W, $\mu = 0, \sigma \approx 0.001$, and $\Phi^{-1}(.99 \mid \mu, \sigma) \approx 0.003$. I concluded that the deviations from zero are so low that an explicit correction is not needed.

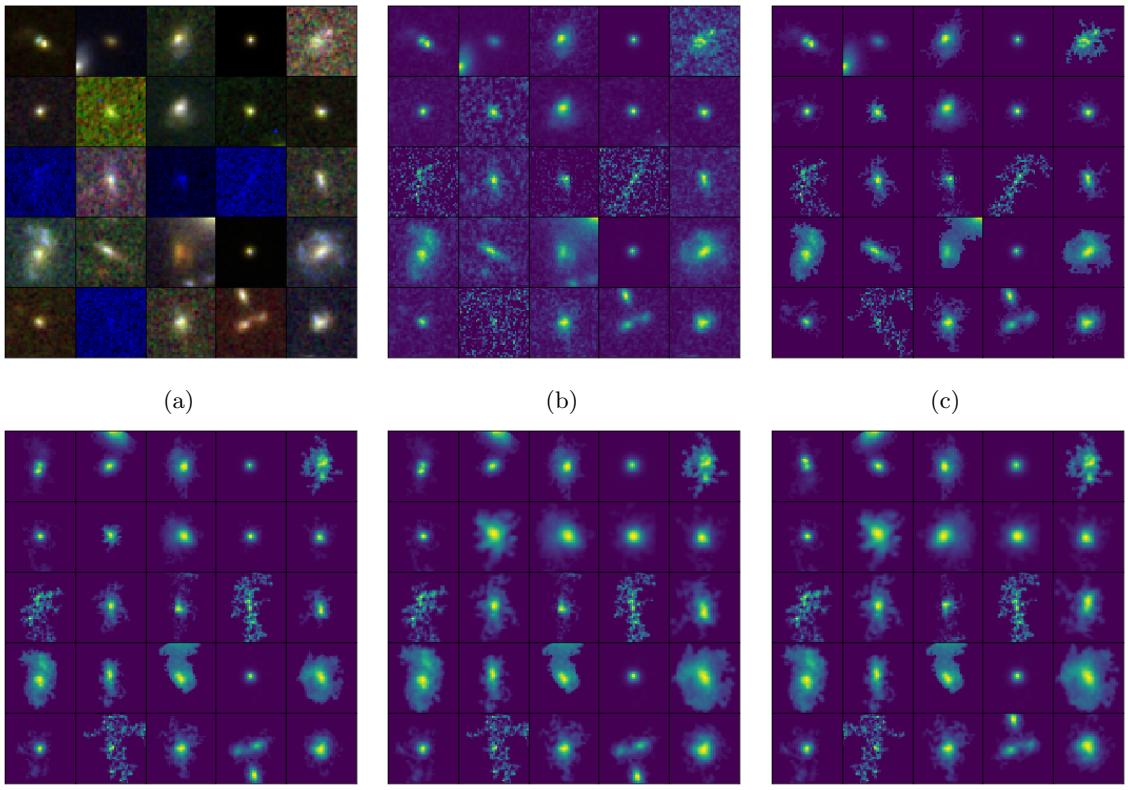


Figure 2: Stages of the image processing pipeline: (a) RGB composite of 25 random galaxies (b) max images created by taking max pixel values across bands to create a compound image (c) background clipping and biggest connected component extraction (d) rotation to vertical (e) image scaling (f) flipping to align brightness to top left.

Background was treated in much the same way however rather than being calculated over survey images, it was calculated for each galaxy thumbnail. Even at 40×40, the majority of each thumbnail is background. As before, it was assumed that the background scene was Gaussian and contaminated by the galaxy and hence the median and IQR were used to estimate distributional parameters. All pixels below the background threshold were clipped to zero. Higher thresholds progressively clip more of the thumbnail towards the center. The threshold was chosen to be $\Phi^{-1}(.65 \mid \mu, \sigma)$ by visually observing a large set of galaxies to see what gets clipped at various values. The threshold parameter is robust to a wide range and anything between $0.6 - 0.75$ has limited down-stream impact.

Image compounding

The band spectrum contains more information than is relevant to just the spatial distribution of a galaxy, and the spatial distribution may not be optimally expressed in any individual band. I produce composite images out of the 3 available bands in my dataset by taking, for each pixel, the maximum flux density available in any band, thus creating a "max image". Band magnitudes tend to be heavily correlated but max images nonetheless have the effect of emphasising relative brightness regardless of which band it happened to occur in. Figure 2 panel (a) and (b) illustrates max images created for some random galaxies. The wide and varying morphologies are clearly preserved in the max images although all band source information is removed.

Magnitude scaling

The difference in brightness between galaxies can be large, and even on a logarithmic scale, brightness can dominate all other differences between vectors. However, only the relative brightness of the pixels in the same image determine its spatial morphology. To this end, all max image pixels were divided by the standard deviation of the pixels in the image. This retains the relative magnitudes within an image but removes the effect of brightness between images. Since all pixels in an image are normalised by the same constant there is no visible difference when presented on a univariate scale (e.g. grayscale).

2.3 Representational invariance

Dimensions so far represent normalised flux densities of galaxy pixels specialised to spatial morphology to the exclusion of absolute brightness and derivatives of band distributions (e.g. colours). For our vector representation to be meaningful, the ith components of all vectors must be comparable to each other. That is, we must conceptualise the projection so that dimensions have fixed meanings. Another way to construe this problem is in terms of invariance in the presence of confounders. That is, every dimension would have a fixed meaning if it was not for some set of effects which prevented it from doing so. In this thesis, I have identified the following effects as being most important.

Centering

This refers to the position of the galaxy in relation to the frame (thumbnail). Photometric imagery tends come with catalogues of RA, DEC coordinates designating object centres. In this thesis, I use the centres from the GZ-CANDELS catalogue to create 40×40 max images as described above.

Rotations and flips

It should be possible for a galaxy image to be rotated or flipped in any way without the galaxy being considered different as a result. That is, the mapping of a galaxy to a point in image space should be invariant to rotation and flipping. If this is not the case then two very similar galaxies may be considered very different just because they are rotated/flipped differently relative to the observer. To account for this effect, for each max image, I use the locations of all pixels with values exceeding the 75^{th} brightness percentile[1] 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 the image and the new variance-preserving one is then given by $\theta = atan2(y_0, x_0)$. Finally, I transform the original max image by rotating it by $\theta - \pi/2$ radians which results in vertically-aligned brightness. Figure 2 panel (d) illustrates the effect of rotation to the vertical axis.

Combined with vertical alignment, the image can be made approximately invariant to flips by flipping the image along the horizontal axis if the lower half is brighter than the upper half, and by flipping the image along the vertical axis if the left side is brighter than the right side. This has the effect of approximately aligning any peripheral brightness to the top left quadrant and so making it more likely that such brightness overlaps with other images[2]. Figure 2 panel (f) illustrates the effect of flipping.

Scaling

Symmetrically scaling an image to fill the thumbnail is a crucial step because it increases the chances that the same part of each galaxy is being represented by any given pixel. The background clipping of the compounded images has the effect of creating a border around the galaxy and/or of sparsifying the background. This makes it much easier to identify the galaxy as the "connected component" (contingent of non-zero adjacent pixels) containing the centre pixel. Once the galaxy component is identified all pixels not in the component

[1] Anything between 60-90% can be used as the percentile; the result is robust

[2] Since thumbnails are square, they technically have 4 symmetries around which to flip, and whilst not pursued in this thesis, it may be possible to go further and find an optimal set of flips to concentrate the most brightness in the top left corner

can be clipped to zero. The image may then be symmetrically shrunk so as to remove as much empty border as possible whilst preserving the image centre, and then stretched back to the standard 40×40 size by utilising cubic spline interpolation (Parker et al. 1983). This has the effect of "zooming in" on galaxies. Figure 2 panels (c) and (e) illustrate the effect of background clipping and image scaling respectively.

2.4 Low-rank approximation

Having fixed how images are to be represented in our projection, we can further improve the image space by noticing that image pixels are spatially correlated. This implies that it may be possible to closely approximate many of the dimensions as a linear combination of the other dimensions, and so reduce the total number of dimensions under consideration. This intuition can be formalised by expressing a "survey matrix" as a stack of projections each considered as row vectors and then looking for the best approximate basis in a lower dimensional space. Most importantly this requires us to fix the meaning of "approximate" by defining how variance will quantified.

Let $S = (v_1, ..., v_L)$ be a survey matrix in which the row vectors v_i with $j = N \cdot M \cdot B$ components are the projected images. I will drop B since I am using max images as explained above. We are interested in $\hat{S}$ with dimensions $L \times k$ where $k < j$. The associated minimisation problem can be expressed as:

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

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

$$||X|| = \sqrt{\sum_{i=0}^{m} \sum_{j=0}^{n} |x_{i,j}|^2} \tag{2}$$

It is commonly used and can be thought of as a matrix generalisation of the Euclidean distance. Since I have gone to some length to make our image space invariant to likely confounders, I am happy to treat all variance equally.

Given a Frobenius norm, the minimisation 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 $S = U\Sigma V^T$ where U, V are orthogonal matrices and Σ 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 $\hat{S}$ of rank k is given by $\hat{S} = \sum_{i=1}^{k} \sigma_i u_i v_i^T$, where $k \leq l$, and u_i, v_i are vectors from U, V respectively.

Principal component analysis (PCA) is a recursive procedure in which a plane (principal component, PC) which minimises the 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 onto 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 is centered, it turns out that PCs discovered in the iterative process above are equivalent to the eigenvectors of the covariance matrix $S^T S$ 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 minimisation problem posed in equations 1.

In this thesis, I use the Frobenius norm as a measure of difference, and I use PCA to optimally reduce the dimensionality of the image space. Following De La Calleja & Fuentes (2004), I term the eigenvectors as "eigengalaxies" because they can be reshaped into 40×40 thumbnails which map their emphasis as in Figure 5. I preprocessed the images as outlined above, and fitted a PCA model to the resultant 1600D space. Figure 3 illustrates how the EVR corresponds to the number of eigenvectors for PCA performed on the data

in three scenarios: (a) when the data is randomly shuffled. (b) when the data is shuffled within columns thus preserving marginal distributions but removing covariance. (c) when the data is not randomised. It is noteworthy that covariance makes a drastic difference to the amount of EVR captured. For example, the EVR at $k = 50$ (close to the elbow in panel (c)) is ~ 0.06, ~ 0.29 and ~ 0.86 respectively across the panels. It is a reassuring observation that indicates our space is indeed heavily affected by spatial correlation as it was designed to be.

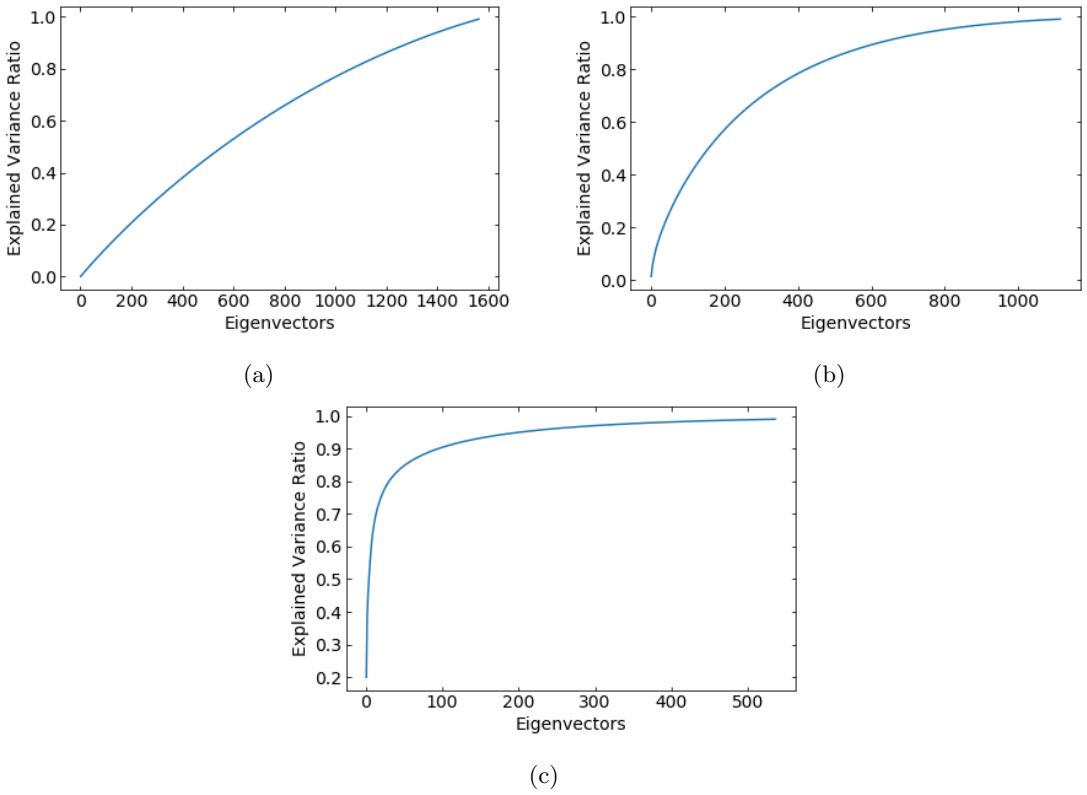
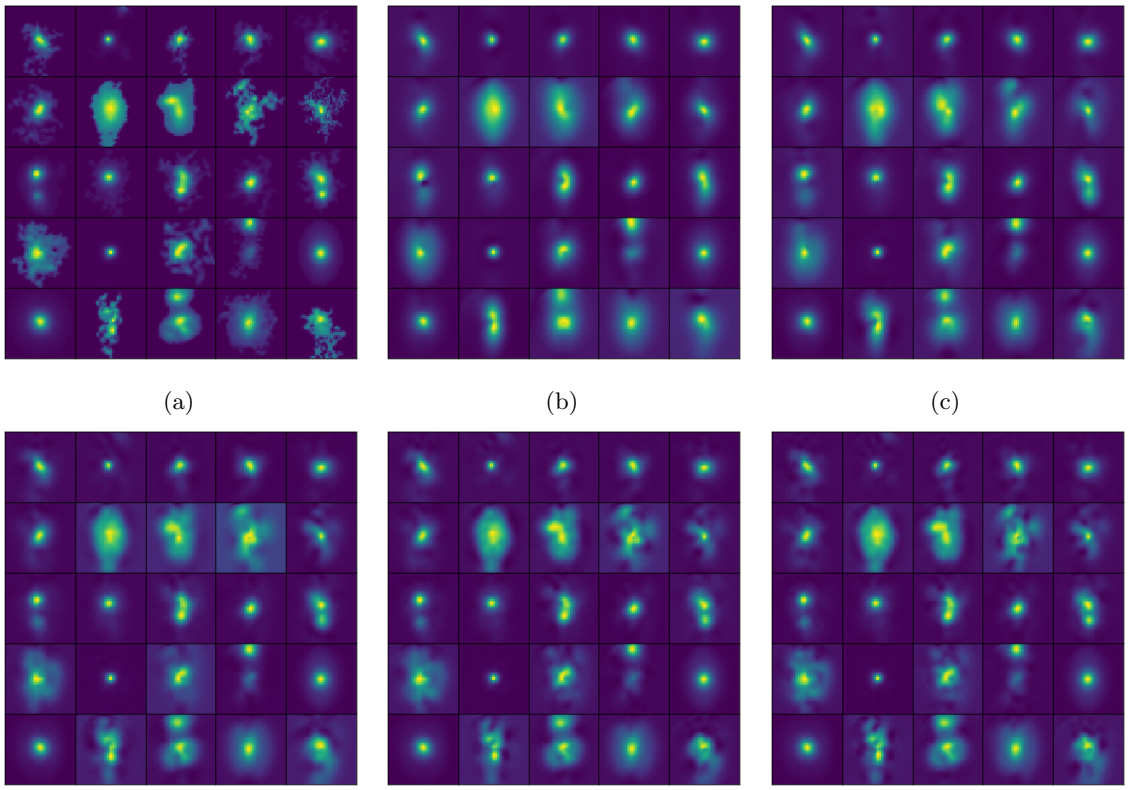


Figure 3: Explained variance ratio as a function of eigenvectors for PCA fitted to the preprocessed GZ CANDELS dataset. In panel (a) the data randomly shuffled. In panel (b) the data is shuffled only within columns so that dimensions retain their marginal distributions but all covariance is lost. In panel (c) the data is not randomised. For example, the EVR at $k = 50$ (close to the elbow in panel (c)) is ~ 0.06, ~ 0.29 and ~ 0.86 respectively across the panels, thus illustrating the drastic impact of covariance EVR captured.

Figure 4 helps clarify how EV corresponds to what morphology is captured. The figure illustrates a random selection of galaxy max images and their reconstruction at 10, 20, 40, 60 and 80 eigenvectors. The series shows that the fidelity of the reconstruction is directly related to the retention of eigenvectors. After some more granular experiments with bigger samples, I concluded that $k = 49$ produces sufficient reconstructions and accounts for $\sim 86\%$ of the total variance.

Eigenvectors are 1600D, and their component values indicate the relative weighting on each dimension. We can reshape the eigenvectors back into a 40×40 image to get a sense of what each eigenvector is emphasising. Figure 5 shows the first 49 eigenvectors depicted as "eigengalaxies". It is evident that the sequence of eigenvectors gets successively more complicated and is therefore more likely that later eigenvectors are involved in expressing details, whilst the early eigenvectors seem more oriented to gross shapes like central or peripheral bulges and rings.

Another way to examine how well the preprocessing and PCA does at creating a morphological space is to pick random galaxies and then look to see what their nearest neighbours in the space look like. Figure 6 illustrates 6 panels, in each the top left galaxy is the random pick, and the remaining galaxies are its nearest neighbours arranged in order of proximity row-wise from top left to bottom right. Given the other analysis



(a) (b) (c)

(d) (e) (f)

Figure 4: Panel (a) illustrates some random galaxy max images. The sequence from b-f shows the reconstruction of the galaxies in panel (a) using 10,20,40,50 and 60 eigenvectors respectively. Texture starts to meaningfully emerge at 40 eigenvectors i.e. panel (d).

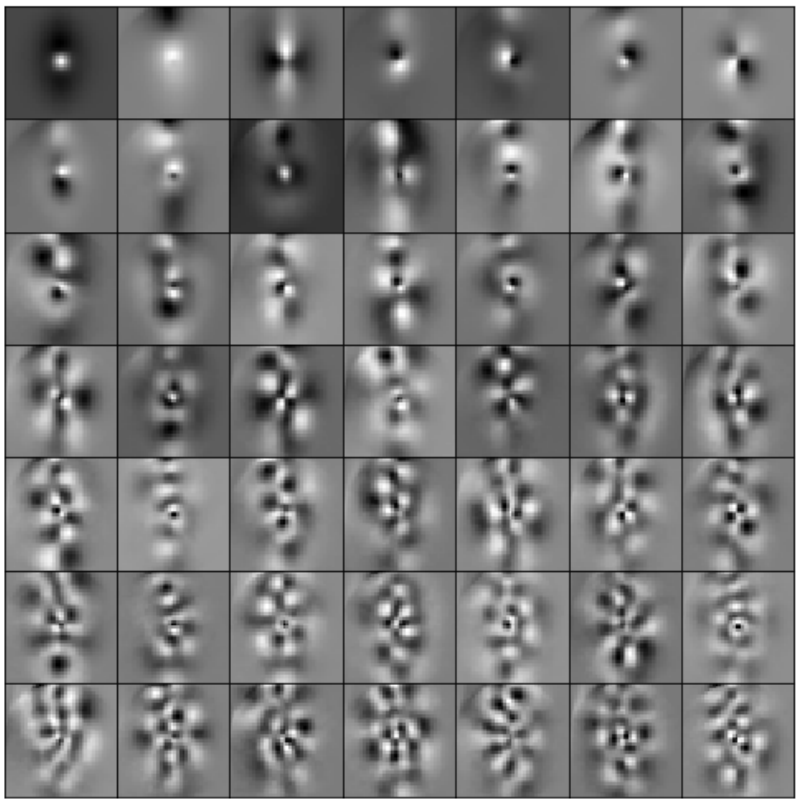


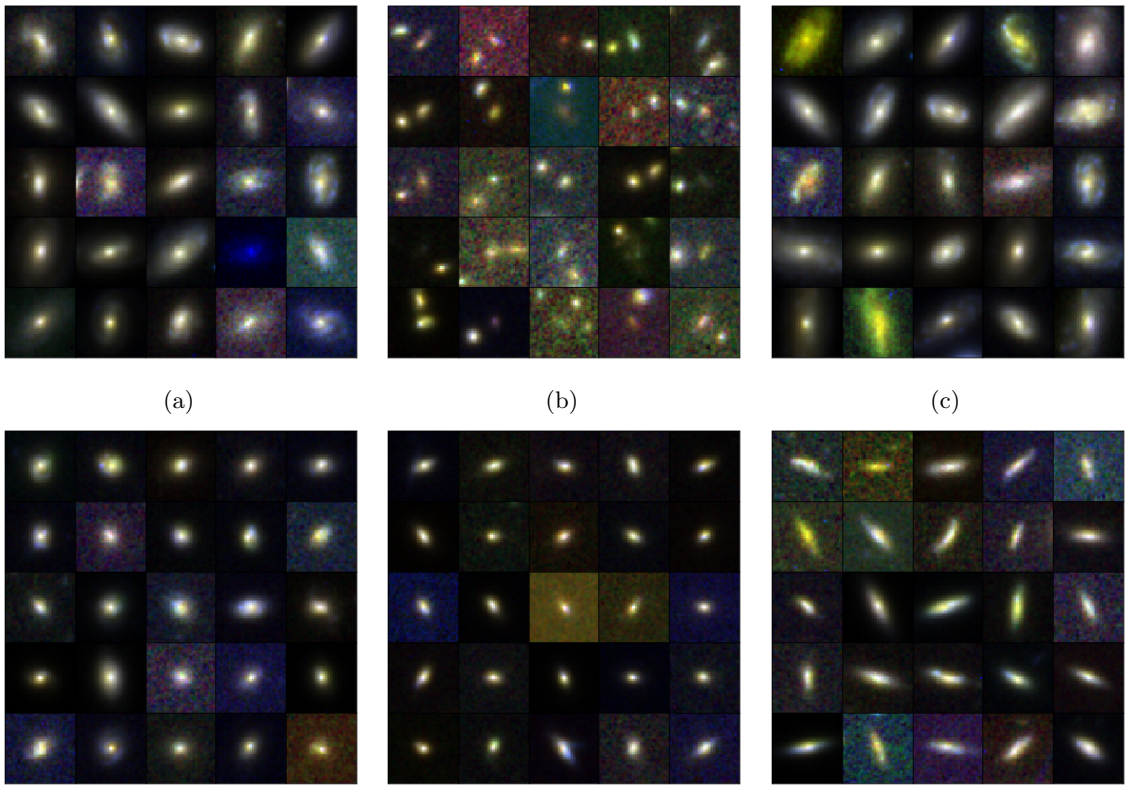
Figure 5: The first 49 eigenvectors laid out by row from top left to bottom right. Brightness and darkness represent higher factor loadings in the positive and negative direction respectively. Components are constrained to be orthogonal to each other which makes them difficult to interpret individually but it is noteworthy that eigenvectors grow sequentially more complicated as the variance they represent becomes less and less general.

above and having considered hundreds of panels such as in Figure 6, I concluded that the system does a remarkably good job at preserving spatial morphology in the new image space.

2.5 Limitations, mitigations and goodness of fit

The preprocessing steps have selected for spatial morphology and normalised dimensions so as to be meaningful and invariant to common confounders. Together with mean centering, this fulfils the basic criteria necessary for PCA. Generally speaking, PCA only fails when there are multiple identical eigenvalues which results in the associated eigenvectors not being unique thus rendering the result meaningless. If PCA fits, how well it fits is ill-defined. In our case, how images are projected greatly determines the significance of the explained variance ratio for example, and we have already gone a long way to show that the EVR captured is indeed meaningful. Otherwise, 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 brightness, but this is controlled for by magnitude scaling in preprocessing. Further, there is no obvious reason to expect large non-linear sub-spaces; and the thesis expects that the transition between clusters is likely to be smooth rather than very well separated (especially given a 1600D space).

At the time of writing, there are no standard (well known) tests which are able 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 $k = 49$ which does a good job as described above an $k = 1$ as the first eigenvector explains the



(a) (b) (c)

(d) (e) (f)

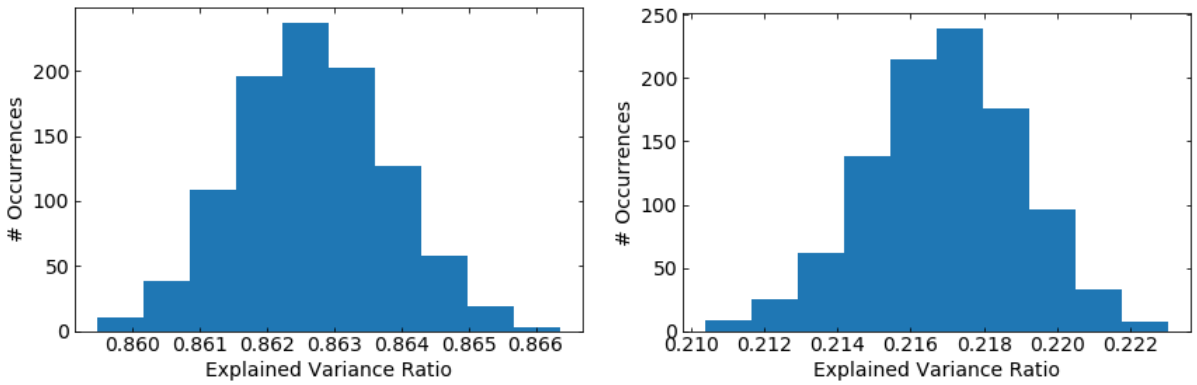
Figure 6: In each panel the top left image is a randomly chosen galaxy and all the other images in the same panel are its nearest neighbours in the image space, ordered row-wise by proximity from top left to bottom right.

most variance and is hence more likely to be affected by degeneracies. If our data is summarised robustly by PCA, the variance of the EVRs given k components should be small and symmetrically distributed. If the EVRs are oddly shaped and/or large then it gives us reason to believe that presence/absence of outliers, clusters, non-linear sub-spaces, or some other degeneracy is causing the fitting to vary. It is important that the sample size taken on every iteration is large enough to make catching these effects likely. Figure 7 illustrates a histogram of the EVR values achieved by performing PCA on 1000 random 70% samples of the data for $k = 49$ (a) and $k = 1$ (b) respectively. The figure shows tight and mostly symmetrical distributions and thus provides some confidence that the data is appropriate for PCA as described herein.

Preprocessing also includes several design decisions which imply limitations and consequences, the most important of which I list below:

- Images are individually normalised by scaling their flux densities to a unit variance (by dividing all pixels by the standard deviation of the flux in the image). This reduces the effect that images with very bright components have on the survey variance, without which few eigenvectors would appear to account for the vast majority of the EVR whilst the variance of interest may remain unaccounted for. However, this adjustment also implies that very bright objects can no longer be separated on the basis of their brightness. Alternative normalisations and standardisations exist and may lead to better or worse results. For example, I could have normalised the flux densities in all images to a fixed range (e.g. 0 to 1) which would temper the effects of variable brightness between images whilst also retaining heterogeneous variances which may have allowed for images with bright components to be discerned.

- Background clipping removes background brightness which is usually unimportant to the morphology of the galaxy, and facilitates the extraction of the galaxy as the biggest (most connected) mass of



(a) (b)

Figure 7: Both graphs show histograms of 1000 PCA fittings on 70% of the data randomly sampled on each iteration. Panel (a) shows the explained variance ratio (EVR) for $k = 49$. Note that the EVR at $k = 49$ when the whole data is retained is ~ 0.862. Panel (b) shows the EVR for the first eigenvector. Note that the EVR at $k = 1$ when the whole data is retained is ~ 0.217.

remaining pixels. Resultantly, this adjustment reduces the variance which must be accounted for in each image and so reduces the number of eigenvectors required to approximate the survey. However, it is possible that at times background clipping may result in the severing of the connectedness between the central galaxy component and peripheral components resulting in important periphery also being clipped and the galaxy neighbourhood being misrepresented.

- Image scaling stretches galaxies out around the center point to increase the likelihood that similar regions of galaxies are being compared to each other. This has the effect of rendering similar those galaxies which may have similar spatial distributions but are of different sizes. However, image scaling is dependent on background clipping so sometimes how much an image is scaled may be arbitrarily decided by the amount of background retained or the presence of peripheral objects. This may lead to the misrepresentation of affected galaxies.

I consider my preprocessing decisions as on the whole justified by the satisfying results however it should be noted that other preprocessing configurations may have worked as well or better.

2.6 Probabilistic interpretation

There are many applications (e.g. outlier detection, dataset comparison, missing value prediction; see sections 3, 4) that require some way of taking the distribution of galaxy points into account. A probabilistic interpretation of our image space able to define its structure and assign likelihood to its points fits the purpose in a very general way. This section describes a way of interpreting PCA as a probabilistic model which will enable advanced applications in the following sections.

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 down as follows:

$$t = Wx + \mu + \epsilon \tag{3}$$

Here $t \in \mathbb{R}^d$ is the data vector, $x \in \mathbb{R}^q$ 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 $x \sim N(0, I)$, 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 $t \sim N(\mu, WW^T + \Psi)$. Whittle (1952) showed that in the case that $\Psi = \sigma^2 I$ (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, W spans the same subspace as PCA and hence

is also an optimal solution to the low rank approximation problem formalised in equation 1. However, the formulation in Whittle (1952) is highly limiting since it is unlikely that in real data the covariance structure is entirely known or that the model and sample covariance are exactly the same. Enter Tipping & Bishop (1999) who show that maximum likelihood (ML) estimates for W and σ do exist without requiring the covariance to be known and, that the scaled principal eigenvectors make up the columns of W when the estimators are at their global maximum. The result (exposed more formally below) is that given the PCA low rank approximation performed in the previous sections, I can directly write down an equivalent factor model which induces a multivariate Gaussian distribution over my image space and hence I am able to assign likelihoods in that space to every galaxy point. Other than the direct applications of this formulation covered in section 4, the factor model allows me to compare image spaces by quantifying the implications of their structural differences on likelihood assignment to galaxy points, which I later leverage to create representative samples.

I will now sign-post 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 $\epsilon \sim N(0, \sigma^2 I)$, it follows that t conditional on x is given by $t \mid x \sim N(Wx + \mu, \sigma^2 I)$. Since $x \sim N(0, I)$ it is easy to marginalise over x to obtain $t \sim N(\mu, WW^T + \sigma^2 I)$. The corresponding log likelihood function is given by:

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

where S is the sample covariance matrix. The ML estimator for μ is the sample mean. Meanwhile, globally optimal estimates for σ, W can be obtained using iterative maximisation algorithm such as those given in Rubin & Thayer (1982) but 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:

$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j \tag{5}$$

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 W ML estimate is given by:

$$W_{ML} = U_q(\Delta_q - \sigma^2 I)^{\frac{1}{2}}R \tag{6}$$

were Δ_q is a $q \times q$ diagonal matrix with the retained eigenvalues $\lambda_1, ..., \lambda_q$ on its diagonal. R is an arbitrary rotation matrix and can be dropped for our purposes by setting $R = I$.

Thus, it is the case then that having calculated PCA, I can use $\hat{\mu}, \sigma_{ML}, W_{ML}$ to immediately write down a multivariate Gaussian which induces a probability distribution over the image space. Sections 3, 4 will make extensive use of these results. Section 3.3 discusses some considerations regarding Gaussian assumptions and the interpretation of probability in this context.

2.7 Summary

In this section I have shown how an image space can be constructed from preprocessed galaxy images. I have shown that spatial morphology can be selectively preserved by actively removing absolute brightness and band information whilst retaining pixel correlations and relative brightness through a set of preprocessing steps. I have also shown how the dimensions in the new space can be given fixed meanings by further preprocessing aimed at making dimensions invariant to a set of likely confounders such as rotation, scaling and background. I showed that the resultant space can be drastically reduced using low rank approximation and that PCA offers an optimal and analytical solution. I discussed various results of applying PCA to our dataset and showed how the fit can be evaluated through the study of image reconstruction, the nearest neighbours of random galaxies, EVR curves and the eigenvectors themselves. I discussed problems and limitations with PCA, and demonstrated how a non-parametric bootstrap procedure can be used as a heuristic to examine the robustness of a PCA solution and to detect the potential presence of more serious issues like outliers, highly clustered data and non-linear subspaces. I have concluded that an image space thus constructed is

robust for the GZ CANDELS dataset and suitable to be further built upon. Finally, I have shown how PCA can be given a probabilistic interpretation as a factor model which furnishes the image space with a general approach to dealing with galaxy point distribution and enables more advanced applications in sections 3, 4.

3 Approximating big surveys with small samples

3.1 Introduction

There are several existing or expectant surveys which will have coverage over billions of objects. Examples include LSST (Robertson et al. 2017), *Euclid* (Refregier et al. 2010) and the Square Kilometre Array (Weltman et al. 2020). The sheer volume of data has the effect of making expensive or intractable traditional and machine learning methods alike. This section investigates survey approximation methods which are able to reduce the size of the data needed for analysis whilst preserving its morphological diversity. The section starts by discussing the ubiquitous simple random sampling (SRS) procedure and its limitations. It then discusses how an image space can be summarised as a mean vector and covariance matrix to make explicit what needs to be preserved. These summaries are then used to compare the representativeness of subsets of the image space to the whole using Kullback-Leibler (Kullback & Leibler 1951) (KL) divergence. Some benchmarks for the new measure of representativeness are established using SRS and finally methods to reduce the sample size below what is possible with SRS are described; some of which will be part of future work.

3.2 Row sampling

Sampling is common in science but usually in the context of cost effectively learning about populations when a census of it is not available. A slightly stranger sampling problem is how to pick a minimal subset from a census such that it can be used instead of the census for some analysis. Yet, it seems a pertinent problem to solve in the era of big data astronomy. Simple random sampling (SRS) is the standard data reduction strategy and involves picking items from a population such that each item has a uniform probability of being selected. SRS is unbiased, so if the population has sub-populations or categories of interest within it, big enough samples may preserve their co-occurrence ratios. To calculate precisely what "big enough" means in this context, we can consider a situation where we would like to apply SRS to a big survey in such a way that it preserves the proportions of some M different morphological types. We can thus imagine SRS to be as if sampling from a multinomial distribution. Tortora (1978) elaborate work originally featured in Goodman (1965) regarding how to calculate sample size in multinomial cases. Let there be M mutually exclusive and exhaustive categories $i = 1, ..., M$, and let π_i and n_i be the true proportion and observed frequency in the ith category respectively. The aim then, for a specific confidence level α, is to specify a set of intervals S_i such that $P\Big[\cap_{i=1}^{M}(\pi_i \in S_i)\Big] \geq 1 - \alpha$. Given that each category is required within an absolute precision b_i, the relationship between sample size and precision is given by:

$$b_i = \sqrt{\frac{B\pi_i(1-\pi_i)}{n}} \tag{7}$$

where B is the upper $\frac{\alpha}{M} \times 100$th percentile of the χ^2 distribution with 1 degree of freedom. Hence, given some fixed α, absolute precision is improved proportionally to $\frac{1}{\sqrt{n}}$, and the total sample size n can be picked such that the worst precision is good enough. This makes evident at least three complications. Firstly, there are diminishing returns to larger sample sizes as precision increases with $\sqrt{n}$. Secondly, it is an assumption that mutually exclusive morphological categories exist in the data. If objects are able to occupy multiple categories (at least due to vagueness) then the intersections of categories make true proportions smaller, and the needed sample size arbitrarily bigger. Thirdly, the proportions π_i in the population are typically not known in advance. This makes it difficult to calculate sample size unconditionally, and in practice often reduces scientists to using the biggest affordable sample.

A key argument in this thesis is that morphology can be captured as a continuous space without recourse to categorisation. If I can establish comparable summaries for image spaces then I would be able to measure

the differences between them without needing to consider how the image spaces might be partitioned or categorised. The following section aims to do just that.

3.3 Summarising and comparing image spaces

For the purpose of this thesis, the primary use of summary statistics is to characterise image spaces such that they may be compared with each other. The key result of section 2.6 was that the PCA low-rank approximations can be represented as a multivariate Gaussian distribution parameterised by its mean vector μ and the covariance of the latent space given by $C = WW^T + \sigma^2 I$. These parameters double up as detailed and convenient summaries with regards to the covariance structure and location of the data. Neither parameter is sufficient by itself because (i) the mean vector says nothing about how the dimensions of the data are related and (ii) the covariance is invariant to addition (it is unaffected by the mean) and hence says nothing about the mean of the data. However, it is unlikely that a real image space is Gaussian; not least because image spaces have no obvious reason to be symmetrical or unimodal. It would also be wrong to state that the image space has a probability distribution as if making a claim about the statistics of how the universe generated galaxies. For my purposes, I will avoid making either claim, and I will instead use the multivariate Gaussian as a device to quantify the implications of the differences between the mean and covariance parameters of image spaces. The Gaussian implication that the likelihood assignment is symmetrical around the mean and shaped by the covariance should not be problematic if I am primarily interested in comparing the locations and covariances themselves.

Any matrix norm could in theory be used for the purpose comparing covariance matrices of the same size (e.g. L2, Frobenius, etc) but typically these measures do not provide a ready way to take the mean vector into account. I thus propose that the KL divergence be used as it is able to assess the effect of both statistics simultaneously. KL divergence is a measure of how one probability distribution differs relative to some reference distribution. By treating two image spaces as Gaussians we may use the KL divergence to meaningfully compare them in terms of how they assign probability to the full range of possible galaxy points. The KL divergence measure spans $[0, \infty]$ where lower values indicate more similarity and 0 means that the two distributions are the same. In the continuous case, the KL divergence of a distribution P relative to a distribution Q is calculated as follows:

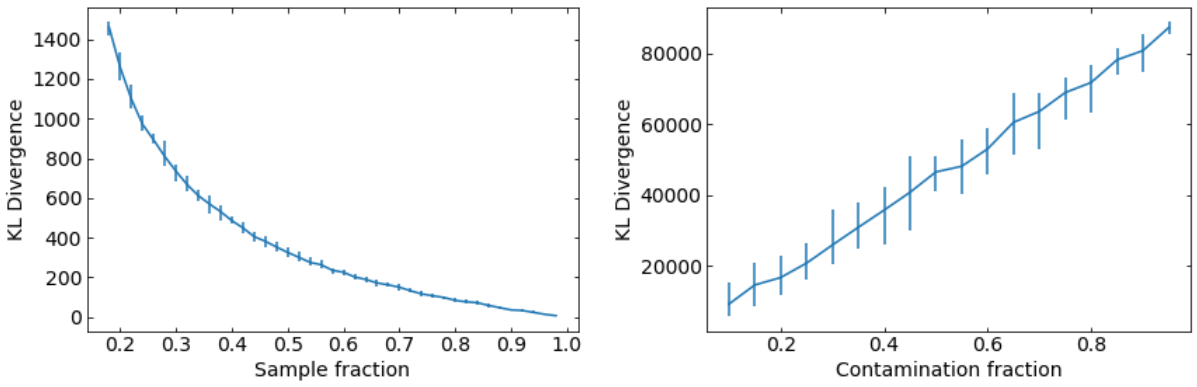
$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log\left(\frac{p(x)}{q(x)}\right) dx \tag{8}$$

KL divergence does not always have a analytical form but it turns out that it does in the case that P, Q are multivariate Gaussians. It is given as follows:

$$D_{KL}(N_0||N_1) = \frac{1}{2}\left(tr(C_1^{-1}C_0) + (\mu_1 - \mu_0)^T C_1^{-1}(\mu_1 - \mu_0) - k + \ln\left[\frac{\det C_1}{\det C_0}\right]\right) \tag{9}$$

where N_0 and N_1 are two Gaussians with means μ_0, μ_1, covariances C_0, C_1, k is the number of dimensions in the covariance matrices. I can examine the implications of summarising image spaces by their means and covariances and quantifying their differences using the KL divergence under Gaussian assumptions by empirically considering how sampling and adding noise affects comparisons. In figure 8, factor models are fitted to various samples and compared to a factor model fitted over the whole image space. Experiments were repeated 10 times for each sample size, and the vertical lines represent the interquartile range. Panel (a) illustrates KL divergences calculated over random samples of the image space of increasing sizes. As sample sizes increase they become more representative of the whole, and KL divergence decreases proportionally. It is noteworthy that the graph shows diminishing returns not unlike a $\frac{1}{\sqrt{n}}$ relationship. Since a KL divergence other than zero does not have an intuitive interpretation, I will use this curve to benchmark the effectiveness of other sampling methods. Panel (b) illustrates the KL divergence of the image space wherein some proportion of the data has been replaced by a column-wise randomised version of the image space (such that marginal distributions of dimensions are retained by the covariance is lost). Since the mean vector of the components of a Gaussian is just the concatenation of its marginal means, the randomisation should not affect the mean estimate hence the contamination should mainly affect the covariance. The effect is severe compared to error

associated with sampling and increases linearly with the fraction of contamination, thus indicating that KL divergence, as calculated, is very responsive to change in covariance structure.



(a) (b)

Figure 8: Panel (a) shows the KL divergence of the factor model at increasing random sample sizes, relative to a factor model calculated on the whole data. At each size, 10 random samples were taken, the vertical lines represent the interquartile range. There is a diminishing return to larger samples. Panel (b) shows the KL divergence of the factor model at increasing levels of contamination, relative to a factor model calculated on uncontaminated data. Contamination was generated by creating a copy of the original data and shuffling the data within columns so as to remove covariance but preserve marginal distributions. There is a linear relationship between contamination and divergence, and it indicates that KL divergence is very responsive to structural change.

3.4 Weighted sampling

If we should pick a suitably small sample size and then calculate the KL divergence of every subset of the image space of that size, it is likely that (i) there would be an ordering of subsets by KL divergence (ii) the set variance would be large enough that it is unlikely we would randomly pick the subset with the lowest (or close to lowest) divergence. Figure 8 panel (a) supports this view and shows the highest interquartile ranges in the smallest sample sizes. One would expect samples to become increasingly representative as the sample size increases. We can attempt to do better than random by adopting a sampling distribution that is not uniform but instead makes it more likely that some points will be selected than others. One such approach is known as leverage score sampling (Papailiopoulos et al. 2014), or as spectral ϵ-approximation (Kumar et al. 2009).

If we describe a linear regression as the picking x to minimise $||Ax - b||$ then the objective is to find $\hat{A}$ with fewer rows that A such that for all x:

$$(1-\epsilon)||Ax|| \leq ||\hat{A}x|| \leq (1+\epsilon)||Ax|| \tag{10}$$

where ϵ is the factor governing how approximate the solution is. For any given row in A, the *leverage score* is given by $\tau_i(A) = a_i^T(AA^T)^+a_i$, where the $+$ indicates the Moore-Penrose inverse (Moore 1920). Tropp (2012) establishes a concentration result showing that if each row i is sampled with a probability $p_i \geq min(2\epsilon^{-2}\tau_i(A)log(\frac{n}{\delta}), 1)$ at most $2\epsilon^{-2}nlog(\frac{n}{\delta})$ points will be chosen. If we construct a diagonal matrix D with weights set according to the leverage scores, then DA will be an ϵ-approximation of A with probability of at least $1-\delta$ (Kyng 2018). The hope is then that if we treat our image space as A, and sample by calculating $\hat{A}$, the information conserved which is relevant to optimal regression is also relevant to preserving the mean and covariance. If we are happy to accept an approximation of factor $\epsilon = 0.1$ then the largest number of points required with $\delta = .1$ probability of failure is $.2^{-2}11469\log(\frac{11469}{.1}) = 26,722,743$. This is orders of magnitude more rows than the total dataset so a somewhat useless guarantee, making the method very dependent on $\tau_i(A)$. To test its efficiency empirically, I calculated the leverage scores for our image space

and and drew samples at different sizes by varying the ϵ parameter. Figure 9 shows increasing sample sizes versus the KL divergence of SRS (blue) and leverage scores (orange). At each size 10 experiments were performed, the vertical lines indicate the interquartile range. It is noteworthy that a $\sim 20\%$ leverage scores sample has the equivalent KL divergence of $\sim 50\%$ SRS sample: 2.5 times fewer rows.

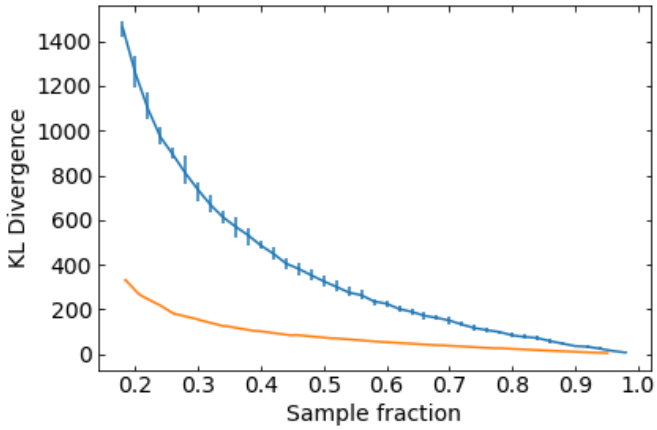


Figure 9: Graph shows increasing sample sizes versus the KL divergence of SRS (blue) and leverage scores (orange). Its noteworthy that a $\sim 20\%$ leverage scores sample has the equivalent KL divergence of $\sim 50\%$ SRS sample: 2.5 times fewer rows.

Our image space is too small to to calculate KL divergence for sample sizes of $\leq 18\%$ because the number of dimensions is so large it leads to zero determinants for which KL divergence is not defined. However, future work will utilise larger surveys which I expect to show even bigger differences in the small sample fraction range.

3.5 Future work

Recently an alternative approach described in Feldman et al. (2016) holds more promise still. PCA is a low-rank approximation technique made special by the fact that it optimally minimises the expected square distance between data $X \subseteq \mathbb{R}^d$ and a lower rank subspace $A \subseteq \mathbb{R}^k$, where A has the orthogonal basis of eigenvectors given by PCA. Feldman et al. (2016) demonstrate an algorithm able to pick a weighted subset of any $n \times d$ matrix X such that:

$$| \, dist^2(X,S) - dist^2(C,S) \, | \leq \epsilon dist^2(X,S) \tag{11}$$

where $dist^2(X,S) = \sum_{i=1}^{n} dist^2(x_i,S)$ and $dist^2(x,S) = min_{x \in S}||x-s||^2$. This implies that PCA carried out on the coreset $C \subset X$ will approximate PCA carried out on X within a multiplicative factor of $1 \pm \epsilon$. More remarkably the authors show that the coreset is of size $\frac{k^2}{\epsilon^2}$. In our case given $\epsilon = .1$ we would expect to require a sample size of $\frac{49^2}{.1^2} = 240,100$ which is also many times bigger than the whole data but most importantly it is an estimate independent of both n (rows in the data) and j (the columns in the data). In future work I intend to implement the coreset construction algorithm for PCA featured in Feldman et al. (2016) and use it to create minimal samples. I intend to test it with an image space built using a much larger dataset wherein I hope to show that a coreset many times smaller than a random sample may be just as efficient in terms of KL divergence.

3.6 Summary

In this section I have explained the main motive behind survey approximation, I have shown how its usually approached with SRS and discussed the inherent limitations. I have shown how the mean vector and covariance matrix can serve as a detailed summary of the data, and how in turn image spaces can be compared by treating their means and covariances as parameters of Gaussian distributions and using KL

divergence to quantify their differences. I have used KL divergence with SRS to establish a benchmark by which other sampling methods may be measured. I have described leverage scores and shown that leverage score samples can reduce size by as much as 2.5 times whilst retaining the same KL divergence. Finally, I have explained the basic principle of coreset construction for PCA and motivated future work to explore designs which may still further improve sample size efficiency.

4 Further applications

4.1 Introduction

The image space established can be considered as a framework within which many applications can parsimoniously be expressed and made principled. For example, the similarity of nearest neighbours established in section 2.4 provides a simple basis for similarity search and clustering, whilst the probabilistic interpretation of an image space enables us to define "outliers" as the galaxies assigned the lowest likelihood and the prediction of missing data as that which is most likely conditional on the rest. In this section I apply the image space to deal with a selection of common use cases.

4.2 Clustering

Unsupervised clustering is particularly useful for astronomy datasets because it provides a method of 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 dataset by characterising its *morphological centers* 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) utilise k-means to classify spectra from SDSS into fewer base types. Valenzuela & Pichara (2018) use k-medoids[3] to cluster and map sequences of light curve segments to variational trees.

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 a $N \times N$ 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 minimised. Formally, for some set of galaxy points $s \in S$:

$$\min_{\{q_i\}_{i=1}^{m}} \left(\sum_{s \in S} \min_{i} d_{s,q_i} + \sum d_{q_i,q_i} \right) \tag{12}$$

where $q_i, ..., q_m \in S$ are m examplars. The second summation acts as a regulariser barring trivial solutions such as picking every point as its own exemplar, and ultimately controlling 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 centers and their membership. The problem is NP-hard (Komodakis et al. 2009) but there are myriad relaxations and approximations that deliver near optimal solution in empirical tests. One such approximation is known as affinity propagation (Dueck 2009) (AP) which minimises a similar equation using message passing over factor graphs. I used the image space to fit AP with its default parameters. It resulted in 462 clusters of median size 14, the sizes of the 6 biggest clusters were 91-115. Figure 10 illustrates 25 random examples from a selection of diverse clusters. Due to preprocessing galaxies within clusters are similar despite differing band distributions or heavy speckling in some bands.

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

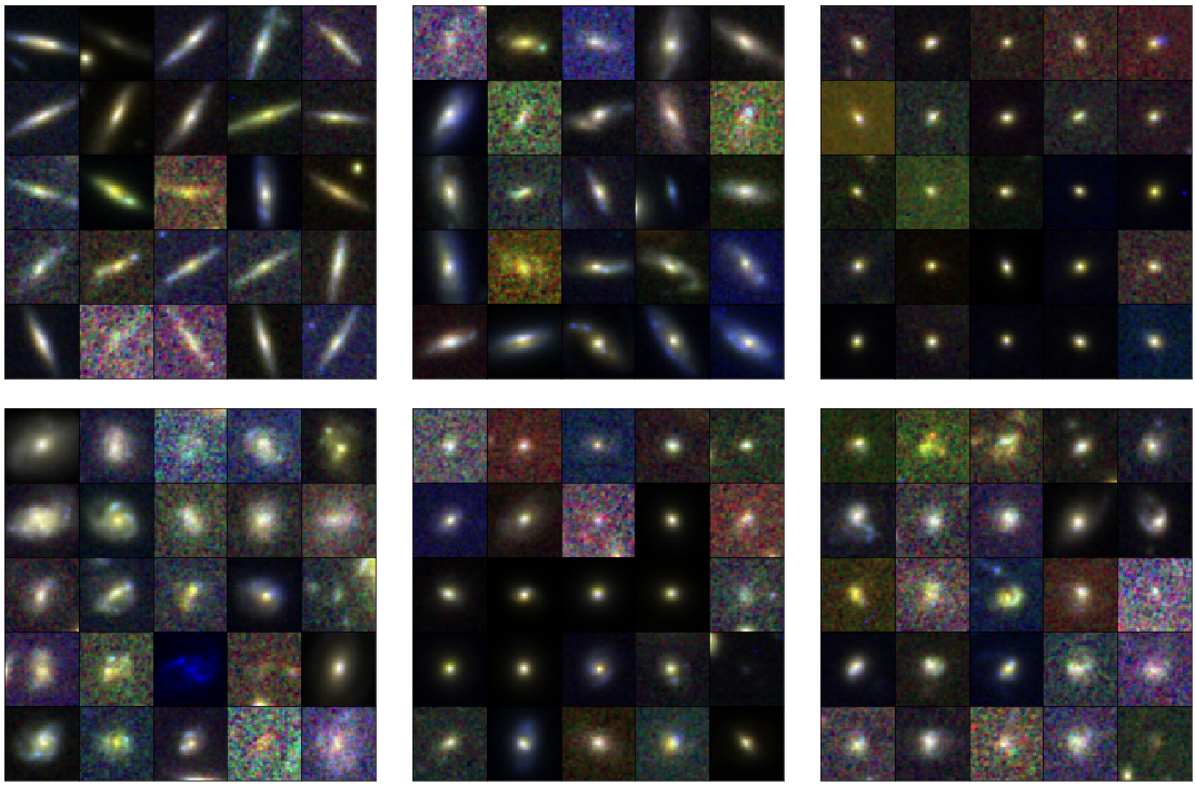


Figure 10: Composite image samples of six morphological clusters from a total of 462 created using affinity propagation clustering of a distance matrix defined by pairwise Euclidean distances in 49D image space. Each 40×40 pixel ($2.4''$) thumbnail image is an RGB composite of the F160W, F125W, and F814W bands.

Although not explored in this thesis, since exemplars are locally representative in the image space, discovering exemplars using AP may also be a sampling method more efficient at preserving morphological variation than SRS. I intend to consider this in future work.

4.3 Similarity search

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) utilise 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 settle on using cosine distance to measure similarity from a description generated by growing neural gas prior to hierarchical clustering.

In section 2.4, I explained that the suitability of the image space can be inspected by considering the similarity of the nearest neighbours are for any given galaxy, and I provided some examples in Figure 6. "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 k-dimensional image 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 spatial morphology and that the number of them retained pertains directly to how well they can reconstruct images.

Therefore, the closer two galaxy points are in image space the more likely they are to be reconstructed in the same way. The method is also straightforward to extend to the case of multiple reference objects are desirable by calculating for each other object the median distance from the reference points and then ordering as before. Figure 11 illustrates six examples of similarity searches, showing the 24 most similar galaxies to the exemplar which is always in the top left. With the exception of the point sources, I have deliberately picked fairly nuanced reference objects with complicated features such as noise, projections, secondary objects, or with particular profiles such as the edge-on galaxy example.

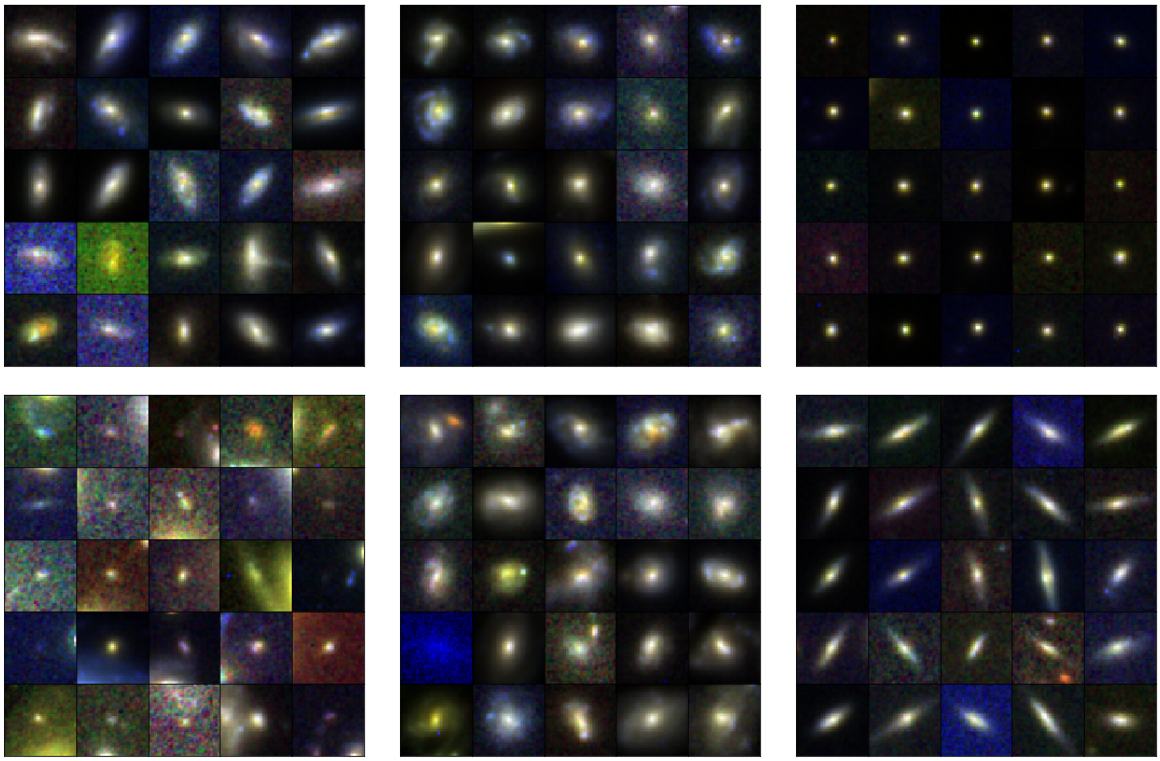


Figure 11: Examples of similarity searches. In each thumbnail image (an RGB composite of the F160W, F125W, and F814W bands), the exemplar galaxy is given in the top left, followed by 24 of its nearest neighbours by Euclidean distance in the 49D image space. I have deliberately picked fairly nuanced reference objects with complicated features such as noise, projections, secondary objects, or with particular profiles such as the edge-on galaxy example.

4.4 Missing data prediction

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 by leveraging the image space. Tipping & Bishop (1999) defines an expectation maximisation (EM) algorithm for fitting the factor model interpretation of PCA in the presence of missing data. It works by treating missing values as jointly distributed with the latent variables and maximising the expectation of the joint likelihood function. Our image space is not suited to this purpose because it intentionally ignores band distribution and magnitude, so for this example I adopt the image space from Uzeirbegovic et al. (2020) which is sensitive to absolute magnitudes and band distribution. I randomly omitted a band with equal probability from our galaxy dataset for 5% of rows chosen at random. I used the data as processed in Section 2.2 of Uzeirbegovic et al. (2020) and an efficient variational EM version of the Tipping & Bishop (1999) algorithm [4] set out in Porta et al. (2005) to fit our factor model. I used

[4] Available in Python package `pyppca`.

the ratio of the sum of differences between the original and predicted pixel flux, and the sum of the flux of the original image as an intuitive measure of reconstruction error. Figure 12 illustrates the distribution of prediction error. The distribution is roughly symmetrical and is centred on zero. The body of the distribution extends to ±20% error but ∼ 72% of objects are accounted for within ±10% error showing that high fidelity prediction is possible for most objects.

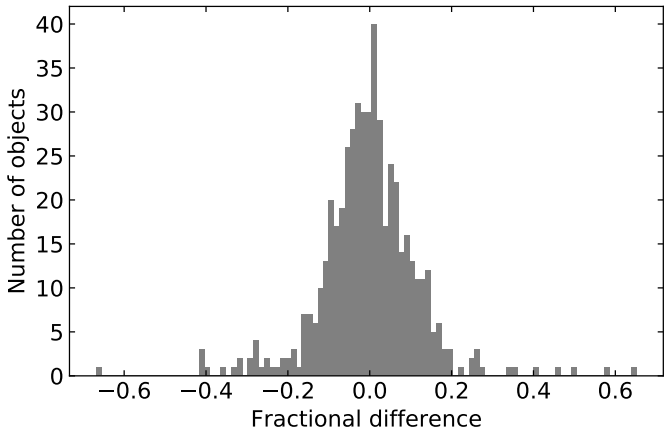


Figure 12: The frequency of the ratio of the sum of deltas (difference between the original and predicted pixel flux) and the sum of the flux of the original images for galaxies with missing data. The body of the distribution extends to ±20% error but ∼ 72% of objects are predicted within ±10% error, indicating that factor model can predict missing data in a 2D sense with a high level of fidelity.

Additionally, Figure 13 provides some examples of predicted data for different bands. It is noteworthy that by using information contained in the latent factors, the factor model can be 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 (SED).

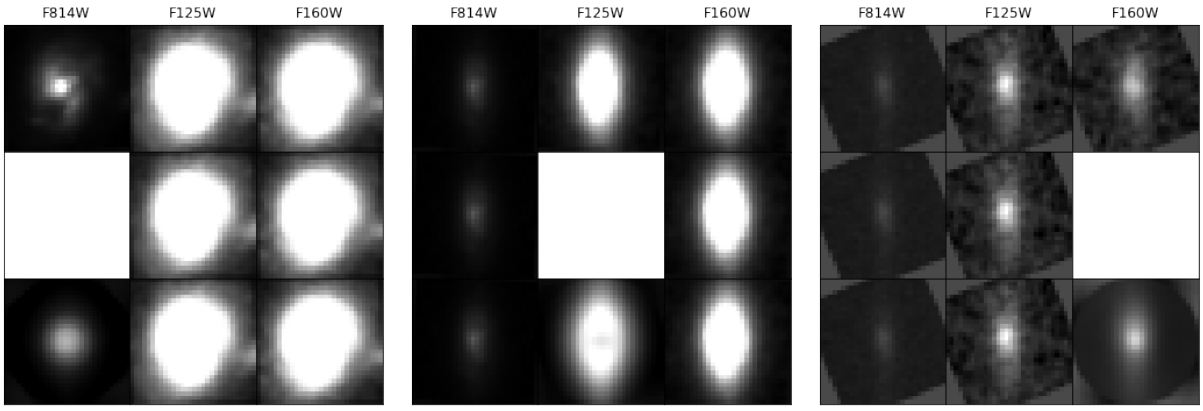


Figure 13: An example of using a factor model for image imputation. In each figure, the first row shows the original galaxy across all three bands, the second row shows the same galaxy with a random band censored, the third row shows the censored galaxy predicted by factor model. It is noteworthy that by using information contained in the latent factors, the factor model can be successful in estimating brightness correctly even when it bears no resemblance to that of the other bands, as illustrated in the left most image.

4.5 Outlier detection

A key utility of outlier detection is to make discoverable rare phenomena buried in enormous datasets. 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) datasets. In defining an outlier they search for galaxies which are over-represented 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) utilise a random forest and fit to discriminate between real and synthetic data[5]. For every pair of objects, they count the number of trees in which each pair is labelled 'real' in the same leaf. The output is a $N \times N$ similarity matrix which is then searched for 'outliers', defined as objects with a large average distance from all other objects.

I focus on a simple and principled definition for an 'outlier', proceeding directly from probabilistic interpretation of the image space: outliers are the objects assigned the least likelihood. Outliers can thus be interpreted as the objects least representative of the object population described by the mean and covariance structure of the factor model. Note that the approach of calculating PCA for a set of features, converting it to a factor model and then using our outlier definition is general and could be applied not only to imaging, but also to spectroscopy, light curves, catalogues and other kinds of data. Formally, given a generative factor model $N(Wx + \mu, \sigma^2 I)$, an outlier x is an object such that $p(x \mid Wx + \mu, \sigma^2 I) < T$ where T is a threshold probability density, and can be set according to the purpose at hand.

To illustrate the concept with the GZ-CANDELS dataset, I use the factor model representation to assign a log likelihood to every galaxy. I sort the data by likelihood and present in figure 14 panel (a) the 25 most rare and in panel (b) the 25 least rare objects according to the likelihood. Nothing is more rare than sparseness and random noise, so the rarest images tend to be misdetections or confounding effects, whilst the least rare objects are point sources. Panel (c) shows the 25 least likely galaxies under a different image space which is sensitive to band distribution and absolute magnitude (Uzeirbegovic et al. 2020). It 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.

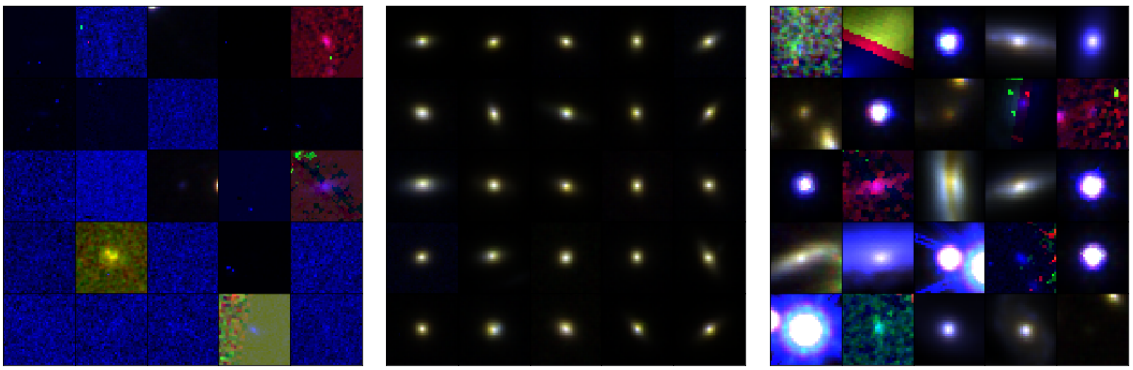
4.6 Summary

In this section I have shown how the image space can be used as a unifying framework to facilitate multiple applications including the following:

- Clustering - which utilised the locality of similar objects together with a exemplar based clustering methodology to discover cluster centres.
- Similarity search - which also used the locality of similar objects to induce an ordering on all galaxies relative to a reference galaxy.
- Missing data prediction - which utilised the probabilistic interpretation of PCA along with an EM algorithm to simultaneously fit a factor model and estimate its missing values.
- Outlier detection - which used the probabilistic interpretation to define as outliers the objects with least likelihood. It also showed how outliers differ under multiple image spaces.

In each case I have shown how the method applied to our image space and demonstrated some motivating results.

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



(a) (b) (c)

Figure 14: Panel (a) shows the 25 least likely galaxies. Noise or extremely sparse signal is predominant. This may be expected since the image space is mostly sensitive to visual patterns. Panel (b) shows the 24 most likely galaxies. Panel (c) shows the 25 least likely galaxies under a different image space which is sensitive to band distribution and absolute magnitude (Uzeirbegovic et al. 2020). It 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.

5 Conclusions

In section 1 I contextualised morphology as largely underpinned by Hubble's scheme, discrete in nature, and deeply connected to galaxy formation history. I described the early automated methods as metrics and simple models aiming at characterising specific aspects of a galaxy, which were succeeded by systems of such models and then by data-driven machine learning techniques; especially prominent amongst which were neural networks. I set out this thesis in contrast, to describe a purely empirical morphology without prerequisite categories in which surveys become image spaces and galaxies become points, the meaning of which is described by the quantifiable differences of their relative positions. It cannot be said that I have entirely succeeded in the matter in any definitive way, but I believe that steps in the right direction have been successfully taken for reasons to follow.

In section 2 I demonstrated how a space can be built which selects for mostly spatial morphology and which is somewhat invariant to absolute magnitudes, background noise, scaling, rotation and flipping. I showed that this space could be compressed using low-rank approximation and that it was robust by illustrating that its derivations were stable even if the sample was varied and that key properties such as the similarity of galaxies to their nearest neighbours were retained. Further, I showed that it could be interpreted probabilistically which provides the segue into image space summarisation, comparison and more advanced applications such as missing data prediction. In hindsight I see other key properties that need to be tested and invariances that need to be addressed. Here are some examples to be considered in future work:

- Many galaxies in the GZ-CANDELS survey are very noisy. The noise makes the galaxy appear unique and rare but in actuality it is just noise. Correlation between neighbouring pixels is very reliable in the galaxy thumbnails, so extremely noisy galaxies can likely be detected by a lack of spatial correlations. A more advanced method may employ techniques from blind source separation such as independent component analysis (Hyvärinen & Oja 2000) which may help separate the Gaussian components from the non-Gaussian ones, and so achieve a more robust image space.

- I should visualise random projections of galaxies which can be ordered by similarity to see how the image space progresses morphologies. Morphology should not be symmetric for example. E.g. I would not expect to see spirals on either side of point sources. I should also look to see if there is a way use the natural ordering constraints inherent in galaxy similarity to create a graph or tree which may be used to visualise the sequences of neighbourhoods that make up an image space. I think techniques

like these are key to establishing a trust that the image space is not arbitrary or specific to a particular purpose. Further, I think that image spaces could be very well complimented by sets of "unit tests" which aim to assert specific ground truths that the space should support as a further proof of coherence.

Section 3 discusses methods for approximating surveys by reducing the number of objects under consideration. I think I have been successful in showing the limitations of SRS, in establishing more general summarisation and comparison procedures and in motivating that better sampling schemes can be created by using information available in the image space. Leverage scores sampling was covered mostly because it is a staple of the data reduction literature, but I think that the coreset method outlined has considerably more potential because it is specific to PCA. I think this is a key area for future work because being able to greatly reduce the size of a large survey may make it tractable to visualisation techniques such as those discussed above which are so crucial to establishing the fidelity of an image space and being able to test it.

Section 4 is a showcase of applications which are made straightforward by features of the image space such as local similarty or the probabilistic interpretation. I think the section is largely successful in illustrating that there are central, reusable aspects of the image space which facilitate many different applications. I think that some applications are missing from the show case, and should be addressed in future work. The chief amongst which are classification and regression. By the former I mean using the image space to distinguish between different types of objects. Separating hyperplane methods such as support vector machines (Gunn et al. 1998) would be especially appropriate given how the image space was constructed. By regression I mean using the dimensions of the image space as covariates to explain or predict some ground truth such for example the fraction of spiral votes for each galaxy in the GZ catalogue. Further, all the methods described in section 4 can be used together to create additional capabilities. Future work should look to outline useful "pipelines" which may facilitate astronomers. Here is an example:

1. Outlier detection surfaces a rare galaxy.
2. Similarity search is used to find more instances of it.
3. A classifier is trained on a handful of instances to approximately count how often it occurs in the survey.
4. The nearest neighbours of the rare galaxies are considered to understand what nearby morphologies look like.

Finally, most methods mentioned herein (with the exception of AP clustering and missing data prediction) already have well documented paths to Big Data scalability which the thesis does not cover. Yet it is important to know that, for example, PCA can be scaled to apply to billions of objects and that most row sampling methods (such as leverage scores) were born in the Big Data world. I think that more time will naturally be devoted to this in future work as a consequence of using bigger surveys for research and analysis.

References

Almeida, J. S., Aguerri, J. A. L., Munoz-Tunón, C. & De Vicente, A. (2010), 'Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra', *The Astrophysical Journal* **714**(1), 487.

Almeida, J. S. & Prieto, C. A. (2013), 'Automated unsupervised classification of the sloan digital sky survey stellar spectra using k-means clustering', *The Astrophysical Journal* **763**(1), 50.

Bamford, S. P., Nichol, R. C., Baldry, I. K., Land, K., Lintott, C. J., Schawinski, K., Slosar, A., Szalay, A. S., Thomas, D., Torki, M., Andreescu, D., Edmondson, E. M., Miller, C. J., Murray, P., Raddick, M. J. & Vandenberg, J. (2009), 'Galaxy Zoo: the dependence of morphology and colour on environment*', **393**(4), 1324–1352.

Baron, D. & Poznanski, D. (2017), 'The weirdest sdss galaxies: results from an outlier detection algorithm', *Monthly Notices of the Royal Astronomical Society* **465**(4), 4530–4555.

Berndt, D. J. & Clifford, J. (1994), Using dynamic time warping to find patterns in time series., *in* 'KDD workshop', Vol. 10, Seattle, WA, pp. 359–370.

Bluck, A. F. L., Mendel, J. T., Ellison, S. L., Moreno, J., Simard, L., Patton, D. R. & Starkenburg, E. (2014), 'Bulge mass is king: the dominant role of the bulge in determining the fraction of passive galaxies in the Sloan Digital Sky Survey', **441**, 599–629.

Bundy, K., Ellis, R. S. & Conselice, C. J. (2005), 'The Mass Assembly Histories of Galaxies of Various Morphologies in the GOODS Fields', **625**, 621–632.

Cappellari, M., Emsellem, E., Krajnović, D., McDermid, R. M., Serra, P., Alatalo, K., Blitz, L., Bois, M., Bournaud, F., Bureau, M., Davies, R. L., Davis, T. A., de Zeeuw, P. T., Khochfar, S., Kuntschner, H., Lablanche, P.-Y., Morganti, R., Naab, T., Oosterloo, T., Sarzi, M., Scott, N., Weijmans, A.-M. & Young, L. M. (2011), 'The ATLAS^{3D} project - VII. A new look at the morphology of nearby galaxies: the kinematic morphology-density relation', **416**, 1680–1696.

Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., Li, N., Bluck, A. F. L., Hartley, W. G., Annis, J., Brooks, D., Doel, P., García-Bellido, J., James, D. J., Kuehn, K., Kuropatkin, N., Smith, M., Sobreira, F. & Tarle, G. (2019), 'Optimising Automatic Morphological Classification of Galaxies with Machine Learning and Deep Learning using Dark Energy Survey Imaging', *arXiv e-prints* p. arXiv:1908.03610.

Codis, S., Pichon, C., Devriendt, J., Slyz, A., Pogosyan, D., Dubois, Y. & Sousbie, T. (2012), 'Connecting the cosmic web to the spin of dark haloes: implications for galaxy formation', **427**(4), 3320–3336.

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

Conselice, C. J. (2006), 'Early and Rapid Merging as a Formation Mechanism of Massive Galaxies: Empirical Constraints', **638**, 686–702.

Darg, D. W., Kaviraj, S., Lintott, C. J., Schawinski, K., Sarzi, M., Bamford, S., Silk, J., Proctor, R., Andreescu, D., Murray, P., Nichol, R. C., Raddick, M. J., Slosar, A., Szalay, A. S., Thomas, D. & Vandenberg, J. (2010), 'Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies', **401**(2), 1043–1056.

De La Calleja, J. & Fuentes, O. (2004), 'Machine learning and image analysis for morphological galaxy classification', *Monthly Notices of the Royal Astronomical Society* **349**(1), 87–93.

De Vaucouleurs, G. (1959), Classification and morphology of external galaxies, *in* 'Astrophysik iv: Sternsysteme/astrophysics iv: Stellar systems', Springer, pp. 275–310.

Dieleman, S., Willett, K. W. & Dambre, J. (2015), 'Rotation-invariant convolutional neural networks for galaxy morphology prediction', *Monthly notices of the royal astronomical society* **450**(2), 1441–1459.

Djorgovski, S. G., Mahabal, A. A., Donalek, C., Graham, M. J., Drake, A. J., Moghaddam, B. & Turmon, M. (2012), 'Flashes in a Star Stream: Automated Classification of Astronomical Transient Events', *arXiv e-prints* .

Dressler, A., Oemler, Jr., A., Couch, W. J., Smail, I., Ellis, R. S., Barger, A., Butcher, H., Poggianti, B. M. & Sharples, R. M. (1997), 'Evolution since z = 0.5 of the Morphology-Density Relation for Clusters of Galaxies', **490**, 577–591.

Dueck, D. (2009), *Affinity propagation: clustering data by passing messages*, Citeseer.

Dutta, H., Giannella, C., Borne, K. & Kargupta, H. (2007), Distributed top-k outlier detection from astronomy catalogs using the demac system, *in* 'Proceedings of the 2007 SIAM International Conference on Data Mining', SIAM, pp. 473–478.

Eckart, C. & Young, G. (1936), 'The approximation of one matrix by another of lower rank', *Psychometrika* **1**(3), 211–218.

Feldman, D., Volkov, M. & Rus, D. (2016), Dimensionality reduction of massive sparse datasets using coresets, *in* 'Advances in Neural Information Processing Systems', pp. 2766–2774.

Fritzke, B. (1995), A growing neural gas network learns topologies, *in* 'Advances in neural information processing systems', pp. 625–632.

Fukugita, M., Shimasaku, K., Ichikawa, T., Gunn, J. et al. (1996), The sloan digital sky survey photometric system, Technical report, SCAN-9601313.

Gloub, G. H. & Van Loan, C. F. (1996), 'Matrix computations', *Johns Hopkins Universtiy Press, 3rd edtion* .

Goodman, L. A. (1965), 'On simultaneous confidence intervals for multinomial proportions', *Technometrics* **7**(2), 247–254.

Goulding, A. D., Greene, J. E., Bezanson, R., Greco, J., Johnson, S., Leauthaud, A., Matsuoka, Y., Medezinski, E. & Price-Whelan, A. M. (2018), 'Galaxy interactions trigger rapid black hole growth: An unprecedented view from the Hyper Suprime-Cam survey', *Publications of the Astronomical Society of Japan* **70**, S37.

Graham, A. W., Erwin, P., Trujillo, I. & Ramos, A. A. (2003), 'A new empirical model for the structural analysis of early-type galaxies, and a critical review of the nuker model', *The Astronomical Journal* **125**(6), 2951.

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

Gunn, S. R. et al. (1998), 'Support vector machines for classification and regression', *ISIS technical report* **14**(1), 5–16.

Hocking, A., Geach, J. E., Sun, Y. & Davey, N. (2018), 'An automatic taxonomy of galaxy morphology using unsupervised machine learning', **473**, 1108–1129.

Hubble, E. P. (1926), 'Extragalactic nebulae.', **64**, 321–369.

Huertas-Company, M., Gravet, R., Cabrera-Vives, G., Pérez-González, P. G., Kartaltepe, J. S., Barro, G., Bernardi, M., Mei, S., Shankar, F., Dimauro, P., Bell, E. F., Kocevski, D., Koo, D. C., Faber, S. M. & Mcintosh, D. H. (2015), 'A Catalog of Visual-like Morphologies in the 5 CANDELS Fields Using Deep Learning', **221**, 8.

Hyvärinen, A. & Oja, E. (2000), 'Independent component analysis: algorithms and applications', *Neural networks* **13**(4-5), 411–430.

Jackson, R. A., Martin, G., Kaviraj, S., Laigle, C., Devriendt, J., Dubois, Y. & Pichon, C. (2020), 'Why do extremely massive disc galaxies exist today?', *arXiv e-prints* p. arXiv:2004.00023.

Jackson, R. A., Martin, G., Kaviraj, S., Laigle, C., Devriendt, J. E. G., Dubois, Y. & Pichon, C. (2019), 'Massive spheroids can form in single minor mergers', **489**(4), 4679–4689.

Kaviraj, S. (2010), 'Peculiar early-type galaxies in the Sloan Digital Sky Survey Stripe82', **406**, 382–394.

Kaviraj, S. (2014), 'The importance of minor-merger-driven star formation and black hole growth in disc galaxies', **440**, 2944–2952.

Kaviraj, S., Devriendt, J., Dubois, Y., Slyz, A., Welker, C., Pichon, C., Peirani, S. & Le Borgne, D. (2015), 'Galaxy merger histories and the role of merging in driving star formation at z 1', **452**, 2845–2850.

Kaviraj, S., Martin, G. & Silk, J. (2019), 'AGN in dwarf galaxies: frequency, triggering processes and the plausibility of AGN feedback', **489**(1), L12–L16.

Kaviraj, S., Ting, Y.-S., Bureau, M., Shabala, S. S., Crockett, R. M., Silk, J., Lintott, C., Smith, A., Keel, W. C., Masters, K. L., Schawinski, K. & Bamford, S. P. (2012), 'Galaxy Zoo: dust and molecular gas in early-type galaxies with prominent dust lanes', **423**, 49–58.

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

Kohonen, T. (1990), 'The self-organizing map', *Proceedings of the IEEE* **78**(9), 1464–1480.

Komodakis, N., Paragios, N. & Tziritas, G. (2009), Clustering via lp-based stabilities, *in* 'Advances in neural information processing systems', pp. 865–872.

Kormendy, J. & Bender, R. (1996), 'A proposed revision of the hubble sequence for elliptical galaxies', *The Astrophysical Journal Letters* **464**(2), L119.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in neural information processing systems', pp. 1097–1105.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86.

Kumar, S., Mohri, M. & Talwalkar, A. (2009), On sampling-based approximate spectral decomposition, *in* 'Proceedings of the 26th annual international conference on machine learning', pp. 553–560.

Kyng, R. (2018), 'A tutorial on matrix approximation by row sampling'.

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., Szalay, A., Andreescu, D., Murray, P. & Vandenberg, J. (2011), 'Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies', **410**, 166–178.

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

Martin, G., Kaviraj, S., Devriendt, J. E. G., Dubois, Y. & Pichon, C. (2018), 'The role of mergers in driving morphological transformation over cosmic time', **480**, 2266–2283.

Martin, G., Kaviraj, S., Hocking, A., Read, S. C. & Geach, J. E. (2020), 'Galaxy morphological classification in deep-wide surveys via unsupervised machine learning', **491**(1), 1408–1426.

Martin, G., Kaviraj, S., Laigle, C., Devriendt, J. E. G., Jackson, R. A., Peirani, S., Dubois, Y., Pichon, C. & Slyz, A. (2019), 'The formation and evolution of low-surface-brightness galaxies', **485**, 796–818.

Martin, G., Kaviraj, S., Volonteri, M., Simmons, B. D., Devriendt, J. E. G., Lintott, C. J., Smethurst, R. J., Dubois, Y. & Pichon, C. (2018), 'Normal black holes in bulge-less galaxies: the largely quiescent, merger-free growth of black holes over cosmic time', **476**, 2801–2812.

Menou, K. (2018), 'Morpho-Photometric Redshifts', *arXiv e-prints* p. arXiv:1811.06374.

Moore, E. H. (1920), 'On the reciprocal of the general algebraic matrix', *Bull. Am. Math. Soc.* **26**, 394–395.

Naim, A., Ratnatunga, K. U. & Griffiths, R. E. (1997), 'Galaxy morphology without classification: Self-organizing maps', *The Astrophysical Journal Supplement Series* **111**(2), 357.

Odewahn, S. C., Stockwell, E., Pennington, R., Humphreys, R. M. & Zumach, W. (1992), Automated star/galaxy discrimination with neural networks, *in* 'Digitised Optical Sky Surveys', Springer, pp. 215–224.

Ostrovski, F., McMahon, R. G., Connolly, A. J., Lemon, C. A., Auger, M. W., Banerji, M., Hung, J. M., Koposov, S. E., Lidman, C. E., Reed, S. L., Allam, S., Benoit-Lévy, A., Bertin, E., Brooks, D., Buckley-Geer, E., Carnero Rosell, A., Carrasco Kind, M., Carretero, J., Cunha, C. E., da Costa, L. N., Desai, S., Diehl, H. T., Dietrich, J. P., Evrard, A. E., Finley, D. A., Flaugher, B., Fosalba, P., Frieman, J., Gerdes, D. W., Goldstein, D. A., Gruen, D., Gruendl, R. A., Gutierrez, G., Honscheid, K., James, D. J., Kuehn, K., Kuropatkin, N., Lima, M., Lin, H., Maia, M. A. G., Marshall, J. L., Martini, P., Melchior, P., Miquel, R., Ogand o, R., Plazas Malagón, A., Reil, K., Romer, K., Sanchez, E., Santiago, B., Scarpine, V., Sevilla-Noarbe, I., Soares-Santos, M., Sobreira, F., Suchyta, E., Tarle, G., Thomas, D., Tucker, D. L. & Walker, A. R. (2017), 'VDES J2325-5229 a z = 2.7 gravitationally lensed quasar discovered using morphology-independent supervised machine learning', **465**(4), 4325–4334.

Papailiopoulos, D., Kyrillidis, A. & Boutsidis, C. (2014), Provable deterministic leverage score sampling, *in* 'Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 997–1006.

Parker, J. A., Kenyon, R. V. & Troxel, D. E. (1983), 'Comparison of interpolating methods for image resampling', *IEEE Transactions on medical imaging* **2**(1), 31–39.

Peirani, S., Crockett, R. M., Geen, S., Khochfar, S., Kaviraj, S. & Silk, J. (2010), 'Composite star formation histories of early-type galaxies from minor mergers: prospects for WFC3', **405**, 2327–2338.

Peng, C. Y., Ho, L. C., Impey, C. D. & Rix, H.-W. (2002), 'Detailed structural decomposition of galaxy images', *The Astronomical Journal* **124**(1), 266.

Porta, J. M., Verbeek, J. J. & Kröse, B. J. (2005), 'Active appearance-based robot localization using stereo vision', *Autonomous Robots* **18**(1), 59–80.

Postman, M., Franx, M., Cross, N. J. G., Holden, B., Ford, H. C., Illingworth, G. D., Goto, T., Demarco, R., Rosati, P., Blakeslee, J. P., Tran, K.-V., Benítez, N., Clampin, M., Hartig, G. F., Homeier, N., Ardila, D. R., Bartko, F., Bouwens, R. J., Bradley, L. D., Broadhurst, T. J., Brown, R. A., Burrows, C. J., Cheng, E. S., Feldman, P. D., Golimowski, D. A., Gronwall, C., Infante, L., Kimble, R. A., Krist, J. E., Lesser, M. P., Martel, A. R., Mei, S., Menanteau, F., Meurer, G. R., Miley, G. K., Motta, V., Sirianni, M., Sparks, W. B., Tran, H. D., Tsvetanov, Z. I., White, R. L. & Zheng, W. (2005), 'The Morphology-Density Relation in z ˜ 1 Clusters', **623**, 721–741.

Protopapas, P., Giammarco, J., Faccioli, L., Struble, M., Dave, R. & Alcock, C. (2006), 'Finding outlier light curves in catalogues of periodic variable stars', *Monthly Notices of the Royal Astronomical Society* **369**(2), 677–696.

Refregier, A., Amara, A., Kitching, T. D., Rassat, A., Scaramella, R. & Weller, J. (2010), 'Euclid Imaging Consortium Science Book', *arXiv e-prints* p. arXiv:1001.0061.

Robertson, B. E., Banerji, M., Cooper, M. C., Davies, R., Driver, S. P., Ferguson, A. M. N., Ferguson, H. C., Gawiser, E., Kaviraj, S., Knapen, J. H., Lintott, C., Lotz, J., Newman, J. A., Norman, D. J., Padilla, N., Schmidt, S. J., Smith, G. P., Tyson, J. A., Verma, A., Zehavi, I., Armus, L., Avestruz, C., Barrientos, L. F., Bowler, R. A. A., Bremer, M. N., Conselice, C. J., Davies, J., Demarco, R., Dickinson, M. E., Galaz, G., Grazian, A., Holwerda, B. W., Jarvis, M. J., Kasliwal, V., Lacerna, I., Loveday, J., Marshall, P., Merlin, E., Napolitano, N. R., Puzia, T. H., Robotham, A., Salim, S., Sereno, M., Snyder, G. F., Stott, J. P., Tissera, P. B., Werner, N., Yoachim, P., Borne, K. D. & Members of the LSST Galaxies Science Collaboration (2017), 'Large Synoptic Survey Telescope Galaxies Science Roadmap', *ArXiv e-prints* .

Ross, D. A., Lim, J., Lin, R.-S. & Yang, M.-H. (2008), 'Incremental learning for robust visual tracking', *International journal of computer vision* **77**(1-3), 125–141.

Rubin, D. B. & Thayer, D. T. (1982), 'Em algorithms for ml factor analysis', *Psychometrika* **47**(1), 69–76.

Sart, D., Mueen, A., Najjar, W., Keogh, E. & Niennattrakul, V. (2010), Accelerating dynamic time warping subsequence search with gpus and fpgas, *in* '2010 IEEE International Conference on Data Mining', IEEE, pp. 1001–1006.

Schawinski, K., Urry, C. M., Simmons, B. D., Fortson, L., Kaviraj, S., Keel, W. C., Lintott, C. J., Masters, K. L., Nichol, R. C., Sarzi, M., Skibba, R., Treister, E., Willett, K. W., Wong, O. I. & Yi, S. K. (2014), 'The green valley is a red herring: Galaxy Zoo reveals two evolutionary pathways towards quenching of star formation in early- and late-type galaxies', **440**, 889–907.

Sérsic, J. (1963), 'Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy', *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina* **6**, 41.

Simmons, B. D., Lintott, C., Willett, K. W., Masters, K. L., Kartaltepe, J. S., Häußler, B., Kaviraj, S., Krawczyk, C., Kruk, S. J., McIntosh, D. H., Smethurst, R. J., Nichol, R. C., Scarlata, C., Schawinski, K., Conselice, C. J., Almaini, O., Ferguson, H. C., Fortson, L., Hartley, W., Kocevski, D., Koekemoer, A. M., Mortlock, A., Newman, J. A., Bamford, S. P., Grogin, N. A., Lucas, R. A., Hathi, N. P., McGrath, E., Peth, M., Pforr, J., Rizer, Z., Wuyts, S., Barro, G., Bell, E. F., Castellano, M., Dahlen, T., Dekel, A., Ownsworth, J., Faber, S. M., Finkelstein, S. L., Fontana, A., Galametz, A., Grützbauch, R., Koo, D., Lotz, J., Mobasher, B., Mozena, M., Salvato, M. & Wiklind, T. (2017), 'Galaxy Zoo: quantitative visual morphological classifications for 48 000 galaxies from CANDELS', **464**, 4420–4447.

Simmons, B. D., Lintott, C., Willett, K. W., Masters, K. L., Kartaltepe, J. S., Häußler, B., Kaviraj, S., Krawczyk, C., Kruk, S., McIntosh, D. H. et al. (2016), 'Galaxy zoo: quantitative visual morphological classifications for 48,000 galaxies from candels', *Monthly Notices of the Royal Astronomical Society* p. stw2587.

Skibba, R. A., Bamford, S. P., Nichol, R. C., Lintott, C. J., Andreescu, D., Edmondson, E. M., Murray, P., Raddick, M. J., Schawinski, K., Slosar, A., Szalay, A. S., Thomas, D. & Vandenberg, J. (2009), 'Galaxy Zoo: disentangling the environmental dependence of morphology and colour', **399**, 966–982.

Skrutskie, M., Cutri, R., Stiening, R., Weinberg, M., Schneider, S., Carpenter, J., Beichman, C., Capps, R., Chester, T., Elias, J. et al. (2006), 'The two micron all sky survey (2mass)', *The Astronomical Journal* **131**(2), 1163.

Smethurst, R. J., Lintott, C. J., Simmons, B. D., Schawinski, K., Marshall, P. J., Bamford, S., Fortson, L., Kaviraj, S., Masters, K. L., Melvin, T., Nichol, R. C., Skibba, R. A. & Willett, K. W. (2015), 'Galaxy Zoo: evidence for diverse star formation histories through the green valley', **450**(1), 435–453.

Soo, J., Moraes, B., Joachimi, B., Hartley, W., Lahav, O., Charbonnier, A., Makler, M., Pereira, M. E., Comparat, J., Erben, T., Leauthaud, A., Shan, H. & Van Waerbeke, L. (2018), 'Morpho-z: Improving photometric redshifts with galaxy morphology', *Monthly Notices of the Royal Astronomical Society* **475**, 3613–3632.

Storrie-Lombardi, M., Lahav, O., Sodre Jr, L. & Storrie-Lombardi, L. (1992), 'Morphological classification of galaxies by artificial neural networks', *Monthly Notices of the Royal Astronomical Society* **259**(1), 8P–12P.

Strateva, I., Ivezić, Ž., Knapp, G. R., Narayanan, V. K., Strauss, M. A., Gunn, J. E., Lupton, R. H., Schlegel, D., Bahcall, N. A., Brinkmann, J., Brunner, R. J., Budavári, T., Csabai, I., Castander, F. J., Doi, M., Fukugita, M., Győry, Z., Hamabe, M., Hennessy, G., Ichikawa, T., Kunszt, P. Z., Lamb, D. Q., McKay, T. A., Okamura, S., Racusin, J., Sekiguchi, M., Schneider, D. P., Shimasaku, K. & York, D. (2001), 'Color Separation of Galaxy Types in the Sloan Digital Sky Survey Imaging Data', **122**, 1861–1874.

Tipping, M. E. & Bishop, C. M. (1999), 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 611–622.

Tortora, R. D. (1978), 'A note on sample size estimation for multinomial populations', *The American Statistician* **32**(3), 100–102.

Tropp, J. A. (2012), 'User-friendly tail bounds for sums of random matrices', *Foundations of computational mathematics* **12**(4), 389–434.

Uzeirbegovic, E., Geach, J. E. & Kaviraj, S. (2020), 'Eigengalaxies: describing galaxy morphology using principal components in image space', *arXiv preprint arXiv:2004.06734* .

Valenzuela, L. & Pichara, K. (2018), 'Unsupervised classification of variable stars', *Monthly Notices of the Royal Astronomical Society* **474**(3), 3259–3272.

Van den Bergh, S. (1976), 'A new classification system for galaxies', *The Astrophysical Journal* **206**, 883–887.

Weir, N., Fayyad, U. M. & Djorgovski, S. (1995), 'Automated star/galaxy classification for digitized poss-ii', *The Astronomical Journal* **109**, 2401.

Weltman, A., Bull, P., Camera, S., Kelley, K., Padmanabhan, H., Pritchard, J., Raccanelli, A., Riemer-Sørensen, S., Shao, L., Andrianomena, S., Athanassoula, E., Bacon, D., Barkana, R., Bertone, G., Bœhm, C., Bonvin, C., Bosma, A., Brüggen, M., Burigana, C., Calore, F., Cembranos, J. A. R., Clarkson, C., Connors, R. M. T., Cruz-Dombriz, Á. d. l., Dunsby, P. K. S., Fonseca, J., Fornengo, N., Gaggero, D., Harrison, I., Larena, J., Ma, Y. Z., Maartens, R., Méndez-Isla, M., Mohanty, S. D., Murray, S., Parkinson, D., Pourtsidou, A., Quinn, P. J., Regis, M., Saha, P., Sahlén, M., Sakellariadou, M., Silk, J., Trombetti, T., Vazza, F., Venumadhav, T., Vidotto, F., Villaescusa-Navarro, F., Wang, Y., Weniger, C., Wolz, L., Zhang, F. & Gaensler, B. M. (2020), 'Fundamental physics with the Square Kilometre Array', **37**, e002.

Whittle, P. (1952), 'On principal components and least square methods of factor analysis', *Scandinavian Actuarial Journal* **1952**(3-4), 223–239.

Willett, K. W., Galloway, M. A., Bamford, S. P., Lintott, C. J., Masters, K. L., Scarlata, C., Simmons, B. D., Beck, M., Cardamone, C. N., Cheung, E., Edmondson, E. M., Fortson, L. F., Griffith, R. L., Häußler, B., Han, A., Hart, R., Melvin, T., Parrish, M., Schawinski, K., Smethurst, R. J. & Smith, A. M. (2017), 'Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging', **464**, 4176–4203.

Willett, K. W., Schawinski, K., Simmons, B. D., Masters, K. L., Skibba, R. A., Kaviraj, S., Melvin, T., Wong, O. I., Nichol, R. C., Cheung, E., Lintott, C. J. & Fortson, L. (2015), 'Galaxy Zoo: the dependence of the star formation-stellar mass relation on spiral disc morphology', **449**(1), 820–827.

Wollaeger, R. T., Korobkin, O., Fontes, C. J., Rosswog, S. K., Even, W. P., Fryer, C. L., Sollerman, J., Hungerford, A. L., van Rossum, D. R. & Wollaber, A. B. (2018), 'Impact of ejecta morphology and composition on the electromagnetic signatures of neutron star mergers', **478**, 3298–3334.