EIGENGALAXIES

A image space and its applications

https://arxiv.org/pdf/2004.06734.pdf

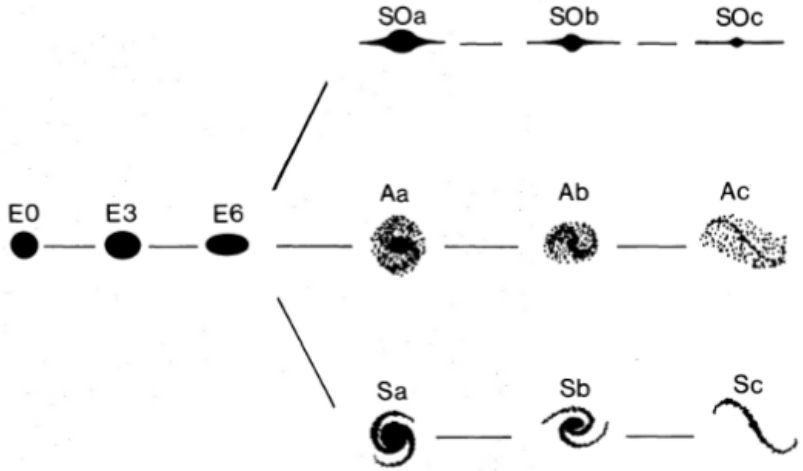
Emir Uzeirbegovic, James Geach, Sugata Kaviraj

CRUX

(how the sausage is made)

Galaxy morphology

- Hubble sequence 1926. Morphology as tied to galaxy formation history.
- Sets the scene to early 2000s: Sersic/Nuker profiles, CAS, M20, Gini, etc.
- Branched early 1990s; NNs, decision trees. Advent of empirical models.
- Galaxy Zoo biggest application of Hubble's classifications on survey datasets, discovered central implications of Hubble's model (eg. the relationship between galaxy bulge and spiral windings) could not be confirmed.
- Today; lots of empirical models.



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

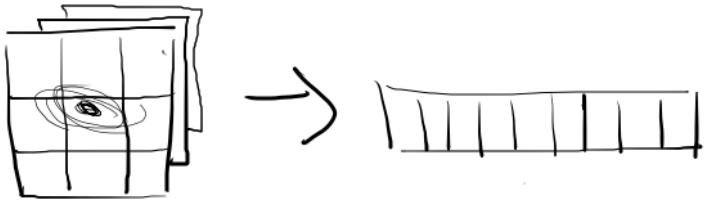
Galaxies as points in m-dimensional space

- I focused on spatial morphology: the patterns of pixels independent of band distribution or absolute flux densities.

- Abandon ex-ante classification: morphology as a continuous space.

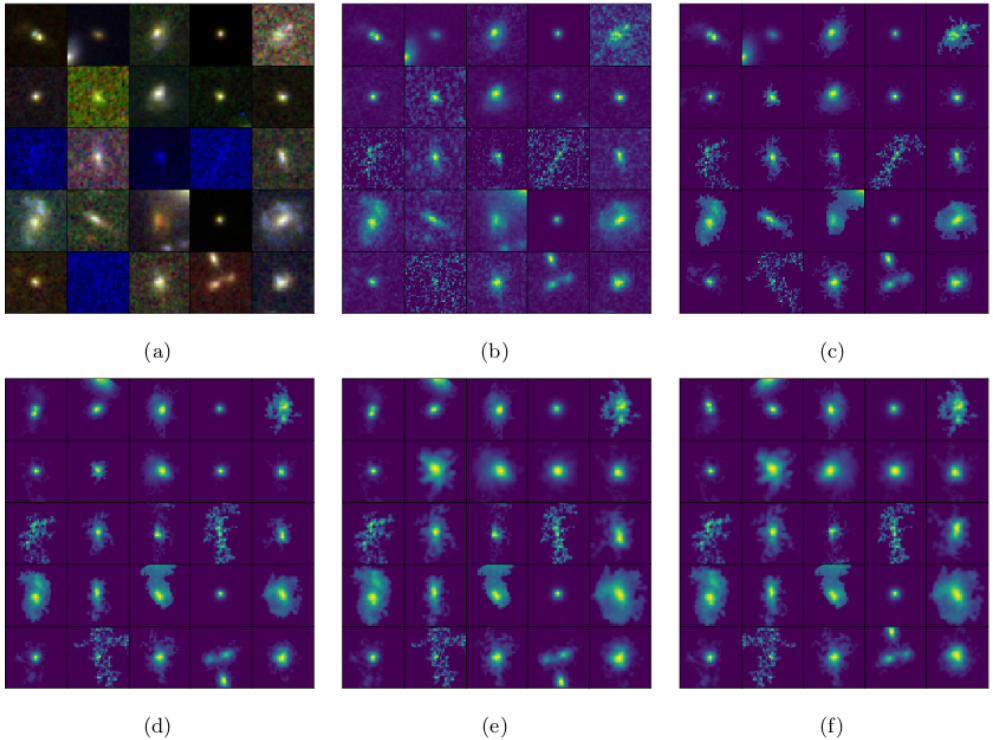
- Aim is to engineer a mapping from multi-band galaxy thumbnails to an m-dimensional vector space such that:
  - Close things in vector space are morphologically similar.
  - Space is robust against distortion.
  - Space is invariant to symmetries.
  - Space implies testable hypotheses.

- I used a the GOODS-S subset of the CANDELS survey with coverage in F814W,F125W,F160W bands for all examples.



Engineered invariance

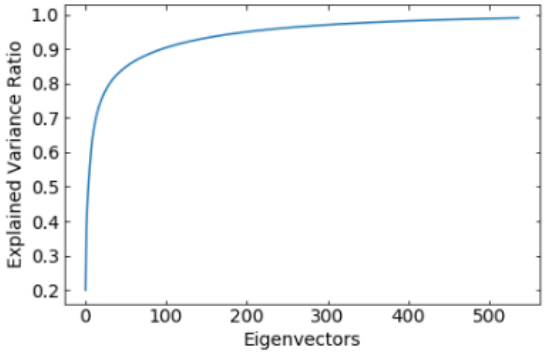
- **Max. image composites**: removes band information, retains spatial information.
- **Magnitude scaling**: relative flux densities only.
- **Rotation**: rotate brightness to vertical.
- **Flipping:** Reflect to top left.
- **Background clipping**: remove noise.
- **Image scaling**: compare like-for-like.

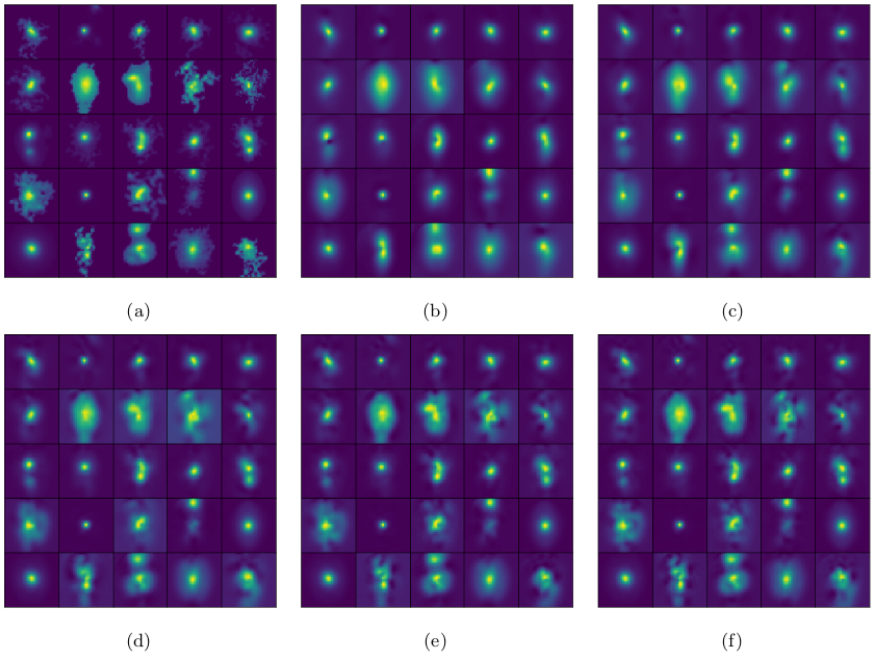


Stages of the image processing pipeline: (a) rgb composite of 25 random galaxies (b) max images created by taking max pixel values across bands to create a compound image (c) background clipping and biggest connected component extraction (d) rotation to vertical (e) image scaling (f) flipping to align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

Reducing the space from *m* to *k*-dimensions

- A 40x40 thumbnail has 1600 dimensions. High dimensional spaces can lead to degeneracies collectively known as the “curse of dimensionality”.
- But pixels tend to be highly correlated with their neighbourhoods. We can exploit that to reduce the number of dimensions.
- PCA used to optimally (and deterministically) reduce dimensions from m to k via low-rank approximation; in this instance from m=1600 to k=49 (in paper k=12 but it’s a different image space).

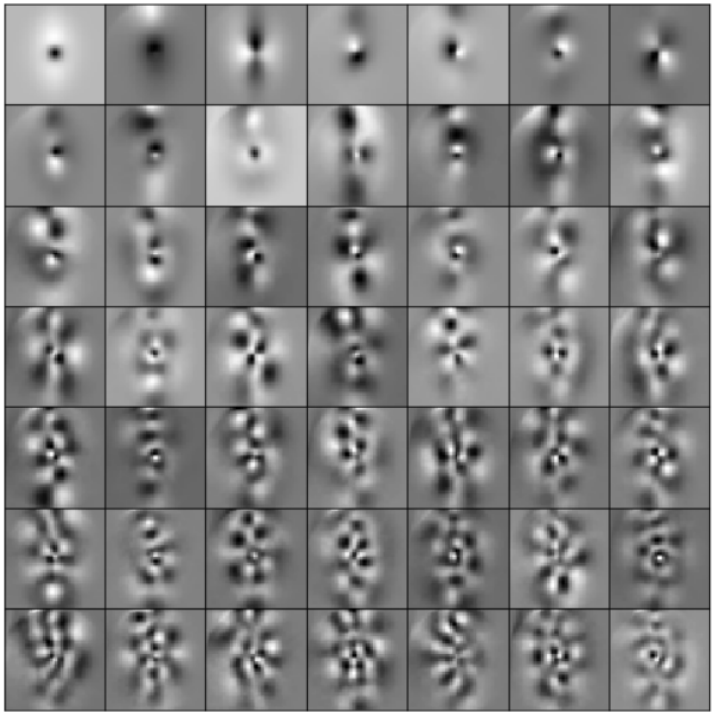




Panel (a) illustrates some random galaxy max images. The sequence from b-f shows the reconstruction of the galaxies in panel (a) using 2,10,20,40,50 and 60 eigenvectors respectively. Texture starts to meaningfully emerge at 40 eigenvectors i.e. panel (d).

Eigengalaxies are just eigenvectors reshaped

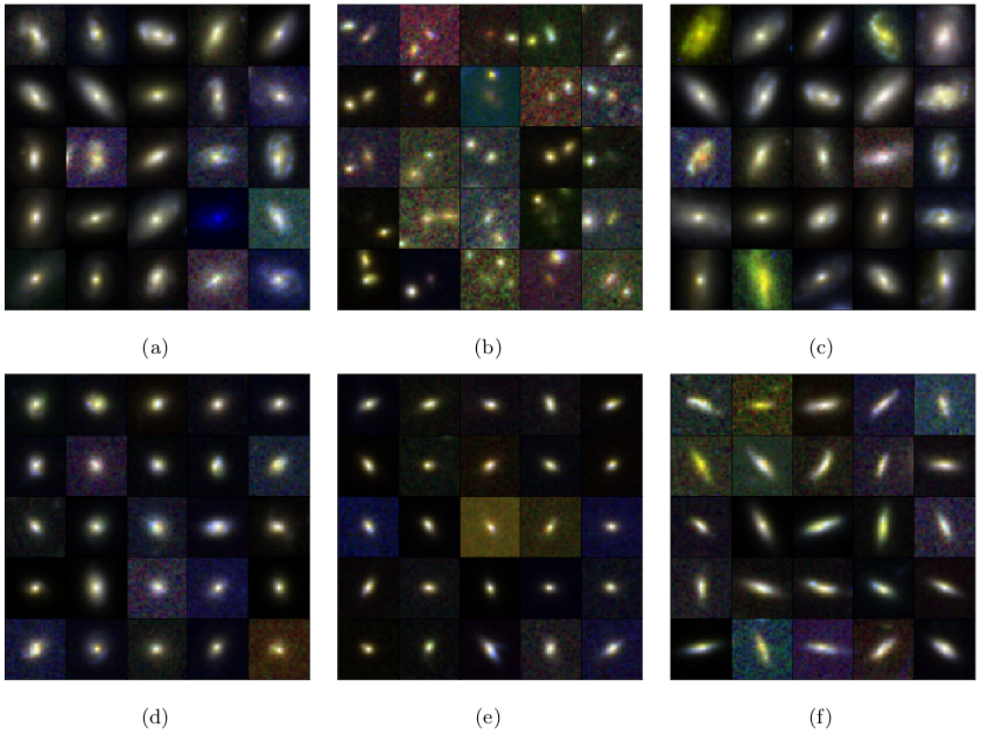
- First 49 eigengalaxies of the image space.
- They start bland and pick out gross features but get progressively more complex.
- Each eigengalaxy is orthogonal to every other thus the covariance of scores is always zero.



… and they’re pretty.

Testing implications: 1/3

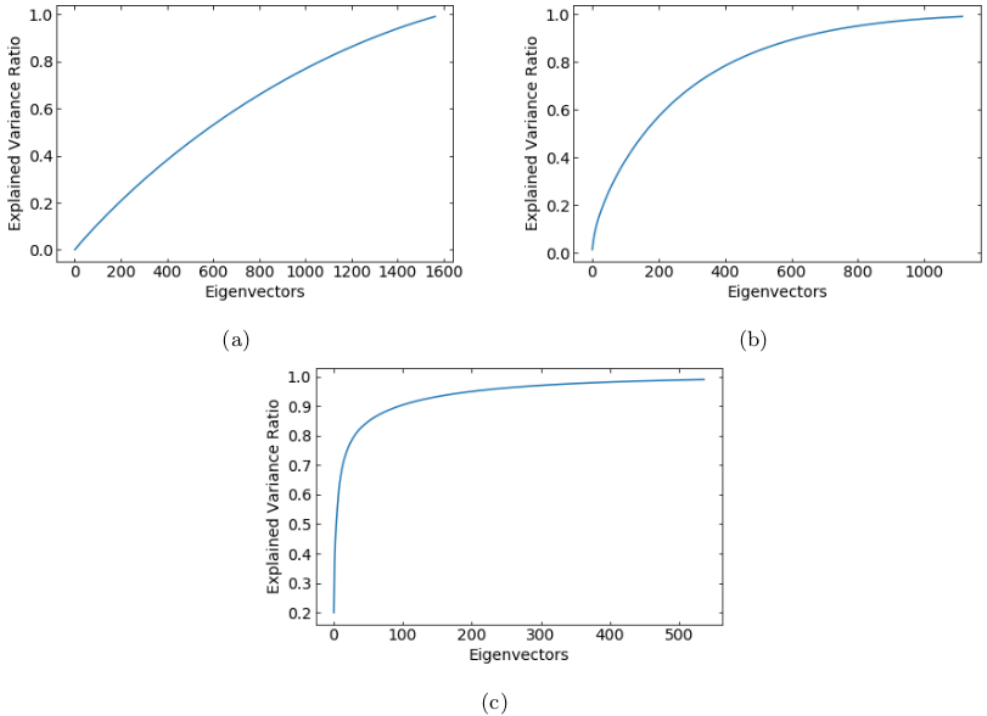
- In a meaningful morphological space, near things should be visually similar.
- We can test that by looking at the n-nearest neighbours.
- *Conversely, far away things should not be visually similar, but how do we test that?*



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.

Testing implications: 2/3

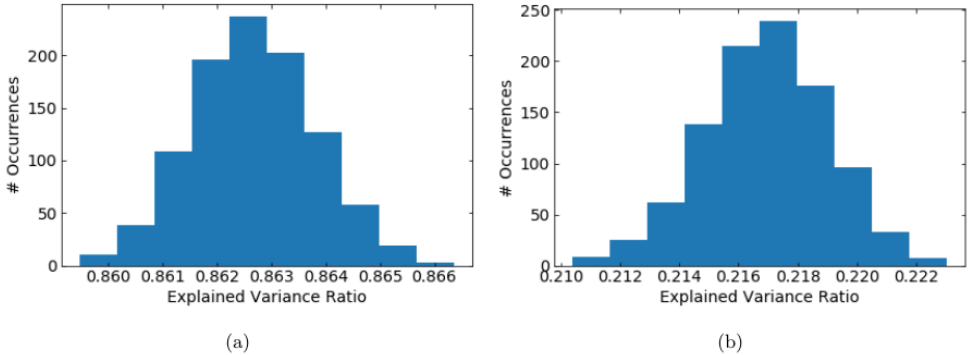
- Space should be sensitive to covariance since we are interested in patterns.
- We can test that assumption by seeing how eigenvectors are affected by randomising.
- Total randomisation keeps range of values but loses marginal distributions and covariance.
- Column-wise randomisation keeps marginal distributions but removes covariance.



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.

Testing implications: 3/3

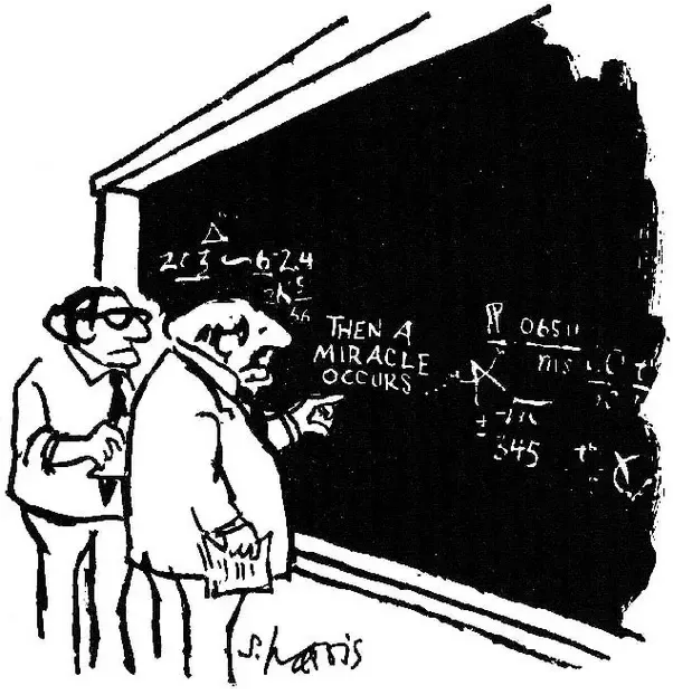
- Clustering and outliers can be a problem for dimensionality reduction.
- We can test if they are by doing the reduction many times on subsets of data.
- The inclusion or exclusion of clusters or outliers make the variance jump around for the bootstrap statistic.
- We can use explained variance at *k* as a statistic to check.



Both graphs show histograms of 1000 PCA fittings on 70% of the data random 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.

Probabilistic interpretation

- It turns out that PCA is equivalent to a certain type of factor model which is equivalent to certain type of multivariate Gaussian.
- … So PCA is equivalent to a certain type of multivariate Gaussian …
- … which can be parameterised by a mean and covariance.
- The parameterisation turns out to be very useful for data summarisation.
- The ability to assign likelihoods turns out to be useful for outlier detection and missing value prediction amongst other things.



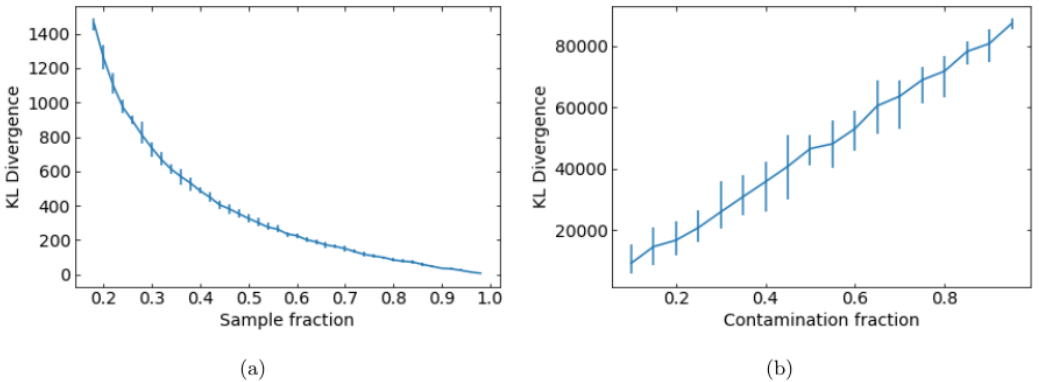
"I think you should be more explicit here in step two."

Summarising & comparing datasets

- Mean and covariance are informative summaries for data.
- We can compare two image spaces by using the mean and covariance parameterisation of their multivariate Gaussian interpretations and plugging those into the Kullback-Leibler divergence.
- Divergence ranges from zero to infinity but we can get a sense of what it means in this context by benchmarking it on random samples.

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

(has a closed form when P and Q are multivariate Gaussian)



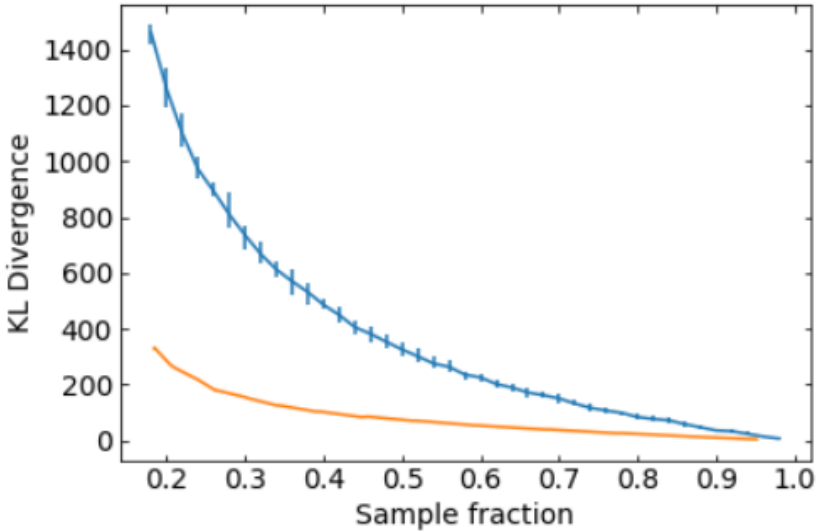
Panel (a) shows the KL divergence of the factor model at increasing random sample sizes, relative to a factor model calculated on the whole data. At each size, 10 random samples were taken, the vertical lines represent the interquartile range. There is a diminishing return to larger samples. Panel (b) 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.

APPLICATIONS

(sampling, clustering, searching, outlier detection, missing data prediction)

Sampling

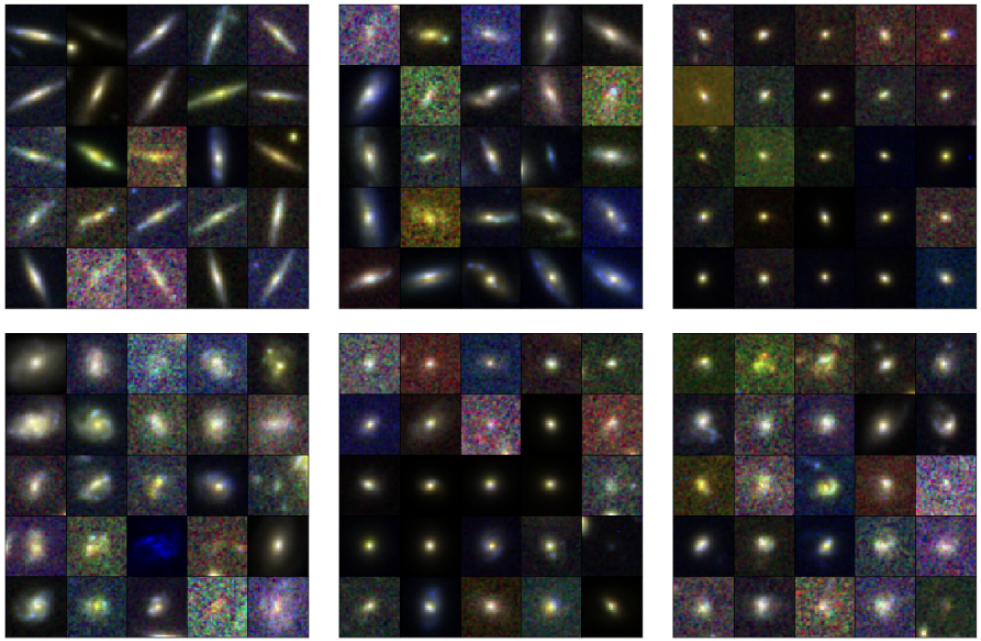
- Random sampling are par for the course, but for big surveys (e.g. LSST, >10B objects) a big enough sample is still a huge dataset.
- We can do better than random sampling. One example is “leverage scores” sampling in which the probability of picking a point is weighted according to its impact.
- Some schemes enable sampling with an error which is independent of the number of points.



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

Clustering

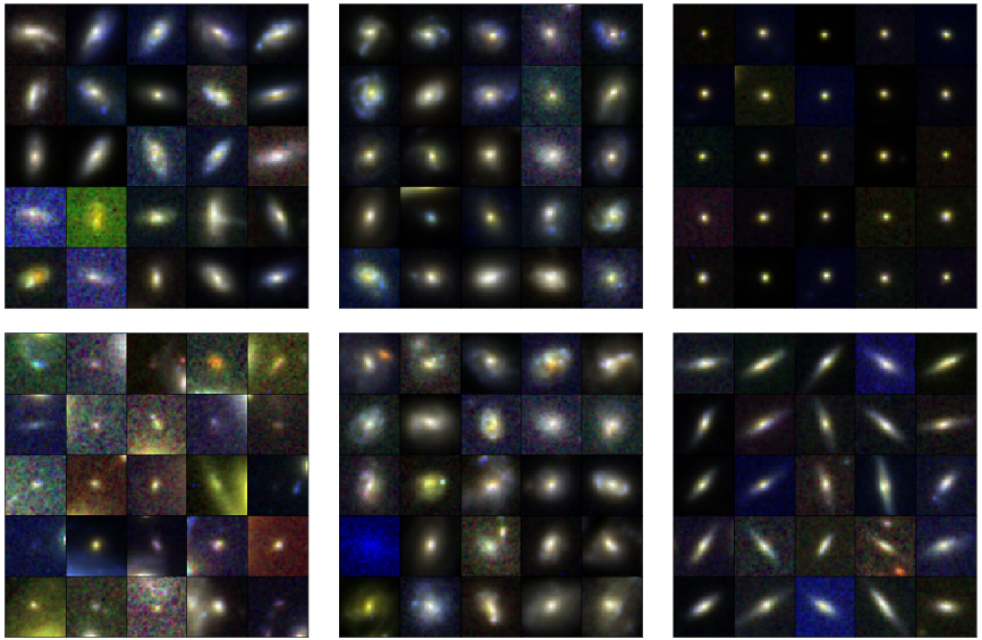
- The linear space makes it easy to create a “distance matrix”: how far away everything is from everything else.
- Lots of clustering algorithms take the distance matrix as a starting point: k-mediods, dbscan, hclust, etc.
- A great one is affinity propagation. Approximately optimal exemplar clustering. No need to specify the number of clusters.



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 thumbnail image is an RGB composite of the F160W, F125W, and F814W bands.

Similarity search

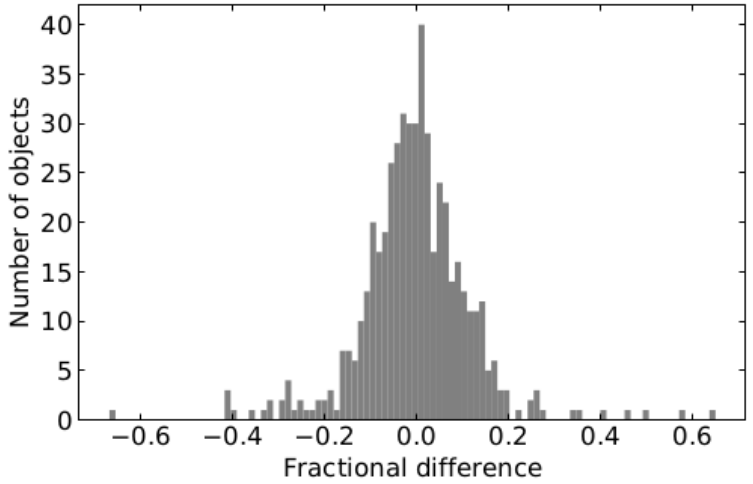
- Given some reference galaxy, a similarity search is just an ordering of the rest of the galaxies by distance to the reference.

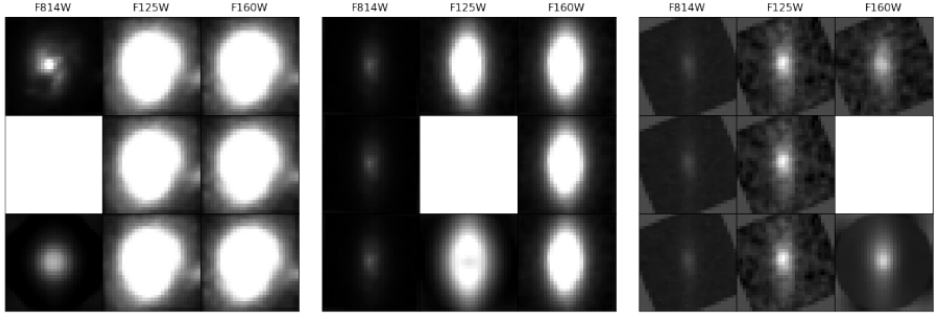


Examples of similarity searches. In each 40x40 thumbnail image (an RGB composite of the F160W, F125W, and F814W bands), the exemplar galaxy is given in the top left, followed by 24 of its nearest neighbours by Euclidean distance in the 49D space.

Missing data prediction

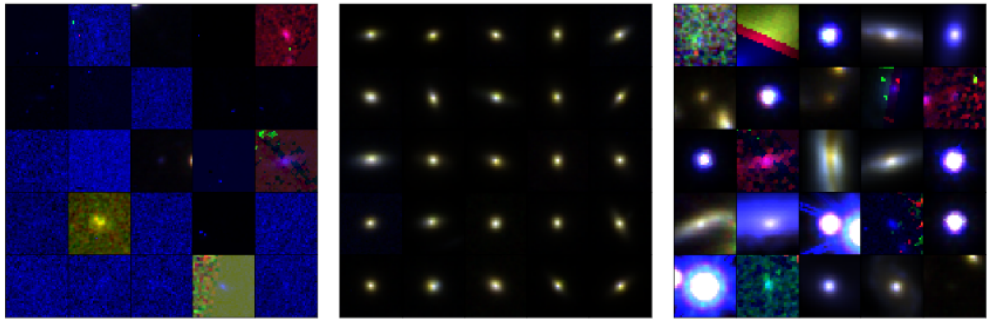
- There is an expectation maximisation algorithm for fitting the probabilistic version of PCA.
- It can be modified to maximise over latent and missing values simultaneously.
- This enables a natural way of imputing missing values whilst fitting PCA.
- Benefit is high fidelity predictions which take into account a lot of structure; difficult to match by other methods.





Outlier detection

- We can think of outliers as “rare” thumbnails.
- ...and “rare” as unlikely.
- The probabilistic interpretation allows us to assign likelihood to all galaxy points, hence sorting by likelihood is a simple way to identify outliers.



(a) (b) (c)

Panel (a) shows the 25 least likely galaxies. Noise or extremely sparse signal is predominent. This may be expected since the image space is only sensitive to visual features. 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.

WORK IN PROGRESS

(simulations, z-morphology,better sampling)

FIN.