Data Science Lab Astrophysics 2024: Lab Worksheet 3: Low Rank Approximations

In this week's project, we will work with low rank representations of galaxy images.

Getting the data

Download the dataset of 256 images of spiral galaxies from the DESI Legacy Imaging Surveys (DECals)[1]. The images have all 256×256 pixels and store the luminosity in the g,r, and z bands. The values are normalised to arbitrary units ranging from 0 to 255. You can load them using the `h5py` library, which you might have to install via `pip install h5py` if you are using `pip`. Then you can load the images like this:

```
import numpy as np
import h5py

with h5py.File('DECals_galaxies.hdf5') as F:
images = np.array( F['images_spirals'] )
```

To display the galaxy number 27, you can generate an RGB image from the g,r,z bands (last index) via the command

```
import matplotlib.pyplot as plt
fig = plt.figure()
plt.imshow(images[27,...])
```

Exploring the data

Work through the tasks, document and report your results in a detailed and reproducible manner.

Task 1. Plot the images for a few galaxies, both as a composite RGB image (as above), and also separately as gray scale images for each band. Generate a new dataset that contains a gray scale image for each galaxy that averages the luminosity in the three bands. Generate a p times down-sampled version of the images by averaging over $p \times p$ pixel blocks.

From now on we will only work with grayscale images.

[1] `https://www.legacysurvey.org`

Task 2. Choose one galaxy of your liking and reproduce the wavelet low-rank approximation from the dataset using only 1/4th and 1/16th of the wavelet coefficients (i.e. keep the A part, discard all others).

Task 3. Use the SVD decomposition to compress the galaxy image. Use again 1/4 and 1/16 of the singular values. Compare the results to the wavelet decomposition. Implement the hard thresholding of the singular values as in the lecture notes. How does the result look? What fraction of singular values are kept in the hard thresholding case?

From now on we will only work with down-sampled grayscale images. Choose a down-sampling factor that still allows you to see some structure but that makes the problem tractable in terms of runtime on your computer.

Task 4. (option: choose either 4 or 5, do both for bonus) Let us look at finding a common basis for all images of spiral galaxies using the SVD. Compute the singular value decomposition (SVD) of the whole galaxy image data set (let one dimension represent a flattened image, the other the different images). Make a plot of the amplitude of the singular values vs. their rank. Show images of the first 4 singular vectors. Do you have an intuition what they represent?

Task 5. (option: choose either 4 or 5, do both for bonus) Use the convolutional autoencoder given in the example notebook on Moodle. Train it on the galaxy images using a 4-dimensional latent space. Compare the results to those given in the lecture notes. Plot histograms of the latent space values for the galaxy images. Use the decoder to generate images representing the latent space values (i.e. input a value to one of the channels of the decoder setting all others to zero) – show the images. What do the different latent space values represent? Can we interpret them in terms of galaxy properties?