Report file - Problem Set #6

Matteo Dell'Acqua
GitHub: MatteoDellAcqua6121

October 15, 2024

Abstract

This is the report for the problem set #6. The scripts and the raw files of the images are in this directory.

1 Introduction

In this problem, we will perform a simple Principal Components Analysis (PCA) on a real data set: the central optical spectra of $9,713$ nearby galaxies from the Sloan Digital Sky Survey.

In the following we are going to use N_g, N_w to denote respectively the number of galaxies in our dataset, and wavelengths measured for each galaxy.

Due to memory problems, we run the program only for the first $N_g = 500$ galaxies.

2 Methods

After importing the data with the `atropy` package:

```
hdu_list = astropy.io.fits.open('specgrid.fits')
logwave = hdu_list['LOGWAVE'].data
flux = hdu_list['FLUX'].data
```

we normalize it so their integrals over wavelength are the same (we use the rough estimate of the Riemann integral[1]):

```
for i in np.arange(Ng):
    for j in np.arange(Nw-1):
        integral[i]+=flux[i,j+1]*(wave[j+1]-wave[j])

for i in np.arange(Ng):
    flux_norm[i,:]=flux[i,:]/integral[i]
```

and we subtract off the mean of the normalized spectra:

```
for i in np.arange(Ng):
```

[1] Technically, in our implementation we are using the weights $w_i = \{0, 1, \ldots, 1, 1\}$ instead of $w_i = \{\frac{1}{2}, 1, \ldots, 1, \frac{1}{2}\}$, but due to the large size of the dataset, the error we are committing is negligible.

```
        flux_mean[i]=np.mean(flux_norm[i,:])
    for i in np.arange(Ng):
        flux_res[i,:]=flux_norm[i,:]-flux_mean[i]
```

We are now ready to start the actual PCA.

We compute the covariance matrix and its eigenvalue decomposition:

```
C=np.dot(np.transpose(flux_res),flux_res)
D, V = np.linalg.eig(C)
```

It is important to note that the eigenvector matrix is defined as $\vec{V^{(i)}} = V_{ji}\mathbf{e}_j$ (note the unusual sorting of the indexes).

Alternatively, we could have used the definition:

$$C = R \cdot R^T \tag{1}$$

and derive the eigenvectors of C starting from the singular value decomposition (SVD) of the residual matrix[2]:

```
(u, w, vt) = np.linalg.svd(flux_res, full_matrices=True)
```

Now that we have found the eigenvectors of the covariance matrix, we can use PCA by computing the coefficient matrix:

```
c=np.dot(flux_norm,V)
```

and increasingly approximate our data with its projection on the first N_c eigenvectors:

```
    for i in np.arange(Nc):
        for j in np.arange(Ng):
            for k in np.arange(Nw):
                if i==0:
                    flux_approximate[j][k][i]=c[j][i]*V[k][i]
                else:
                    flux_approximate[j][k][i]=flux_approximate[j][k][i-1]+c[j][i]*V[k][i]

#up to this point, the flux are normalized: let's undo the normalization!
    for i in np.arange(Ng):
        for j in np.arange(Nc):
            flux_approximate[i,:,j]=(flux_approximate[i,:,j]+flux_mean[i])*integral[i]
```

3 Results

We report the plot of the original data (realized using `matplotlip.pyplot`) in fig. 1. We note that the flux is highly peaked at the wavelengths corresponding to emission/absorption of the elements composing the galaxies. For example we can recognize some of the transitions in the hydrogen atom:

$$\lambda_1 = 4861\mathring{A}, \quad \lambda_2 = 6562\mathring{A}. \tag{2}$$

We report the plot of the normalized residuals in fig. 2.

[2]The option `full_matrices=True` is needed to obtain a square matrix V even when $N_g < N_w$.

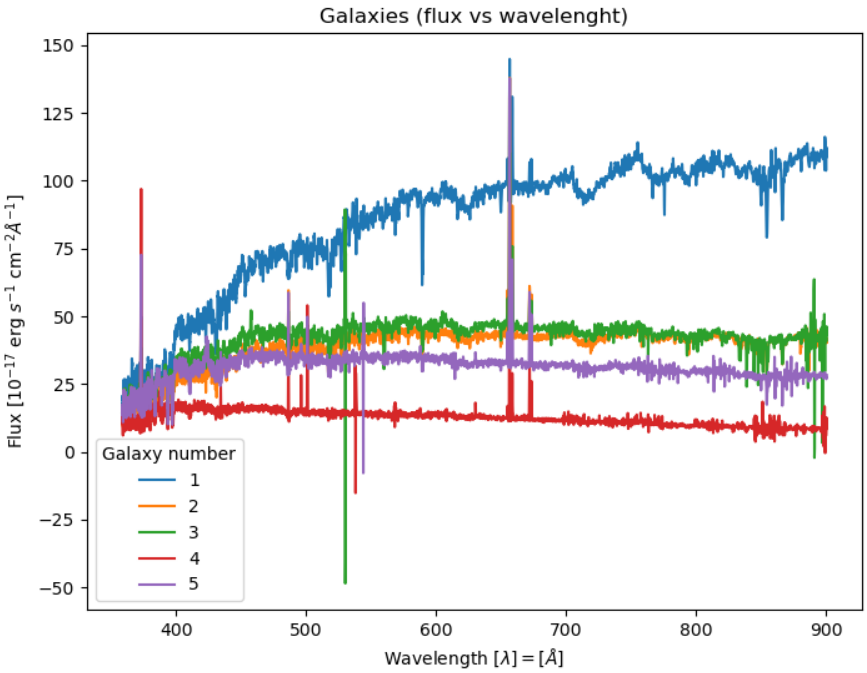

Figure 1: Plot of the flux vs the wavelength for a handful (five) of galaxies.

We report the plot of the firsts few eigenvectors of the covariance matrix and of the residual matrix respectively in figs. 3 and 4: from the graph you can visually see that the two methods are provide equivalent results. However, their computational cost is very different, as can be seen by the time it takes them to run[3]:

$$t_{\text{diag}} = 30.7247s, \quad t_{\text{SVD}} = 1.2998s. \tag{3}$$

The difference can be mainly attributed to the cost of multiplying the R and R^T matrices, since the SVD and diagonalization procedure are then applied to similar size matrices C and R. Another difference is given by the condition numbers of C and R are respectively given by:

$$c_C = 3789007400000.0, \quad c_R = 732.4101. \tag{4}$$

The first one is much larger, since $C \sim R^2$ implies that the eigenvalue (and thus the condition numbers) accordingly grows as a square!

Both the condition numbers and the computational costs suggest the SVD method is faster and more numerically stable.

We report the plots of first coefficients in the PCA in figs. 5 and 6.

We report the plot of the original data of the first galaxy, together with the result of its approximation with $N_c = 1, \ldots, 5$ principal components in fig. 8.

[3]Measured taking difference of global times provided by the funciton `time.time()`.

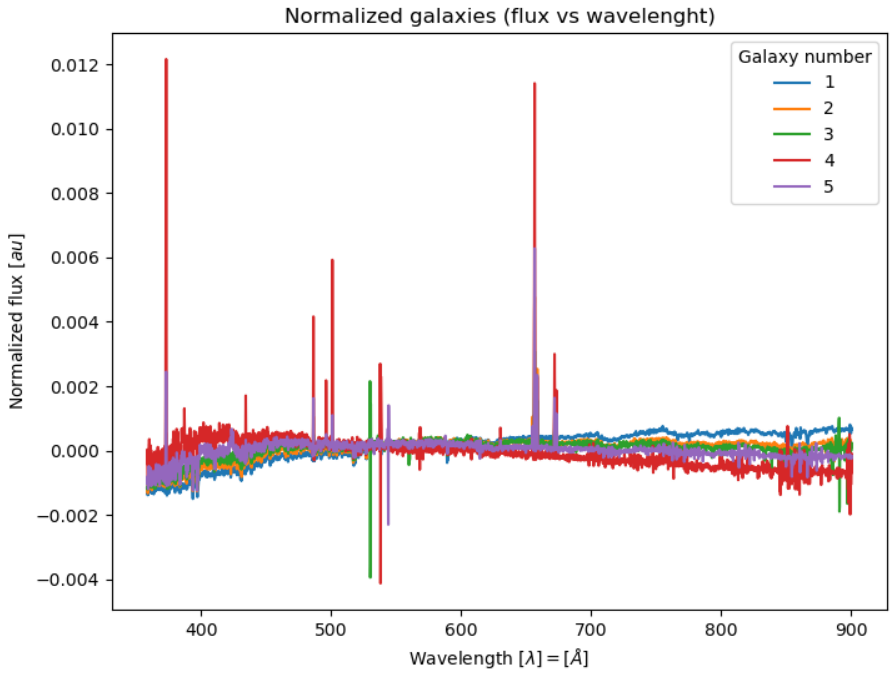

Figure 2: Plot of the normalized residual flux vs the wavelength for a handful (five) of galaxies.

We report the plot of rms residual for $N_c = 1, 2, \ldots, 20$ for the first five galaxies in **??**. We can see that it quickly saturates to values (expressed in $[10^{-17}$ erg s^{-1} cm$^{-2}\mathring{A}^{-1}]$) :

$$\text{rms}_0 = 1.554, \quad \text{rms}_1 = 0.943, \quad \text{rms}_2 = 3.011, \quad \text{rms}_3 = 1.059, \quad \text{rms}_4 = 1.403. \tag{5}$$

which is small compared to the original values of the fluxes ~ 100.

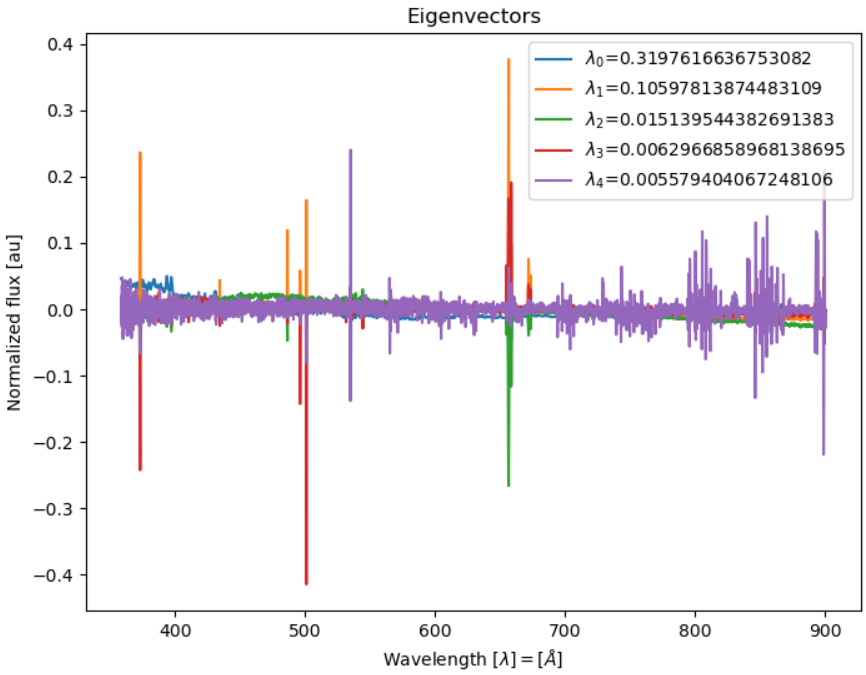

Figure 3: Plot of the normalized residual flux vs the wavelength for the first five eigenvectors of the covariance matrix.

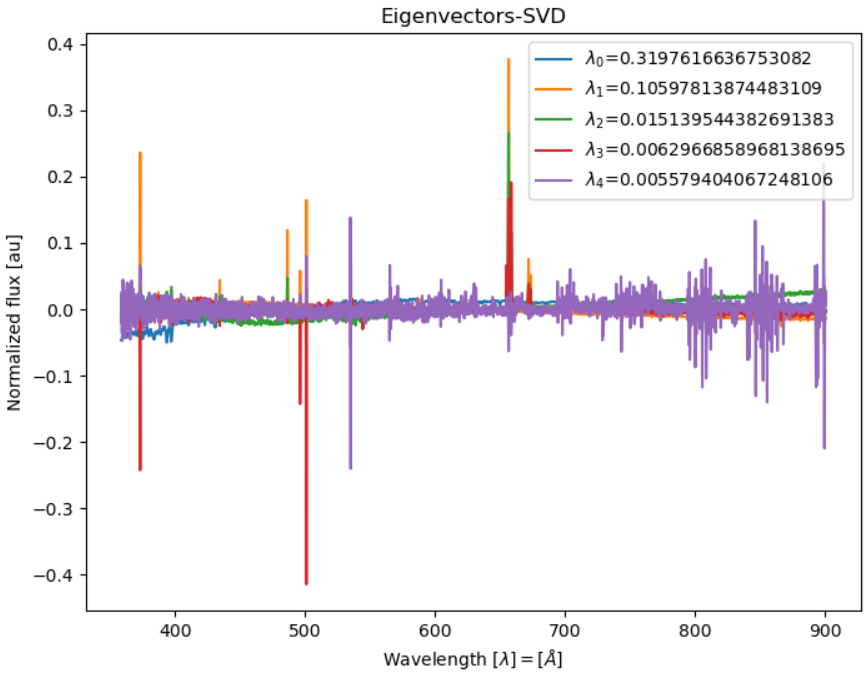

Figure 4: Plot of the normalized residual flux vs the wavelength for the first five eigenvectors of the covariance matrix.

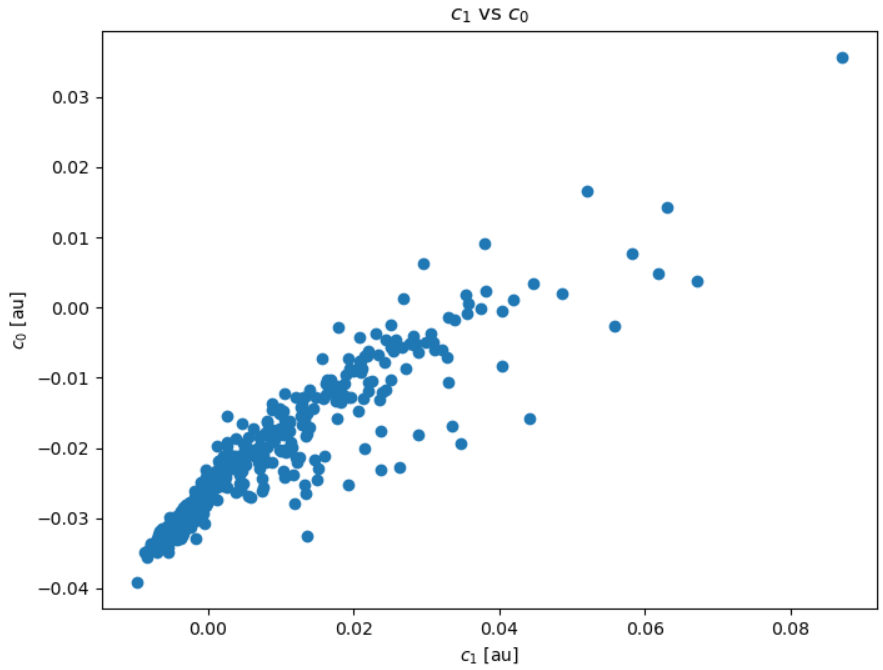

Figure 5: Plot of the first coefficient vs the second one in the PCA, for every galaxy.

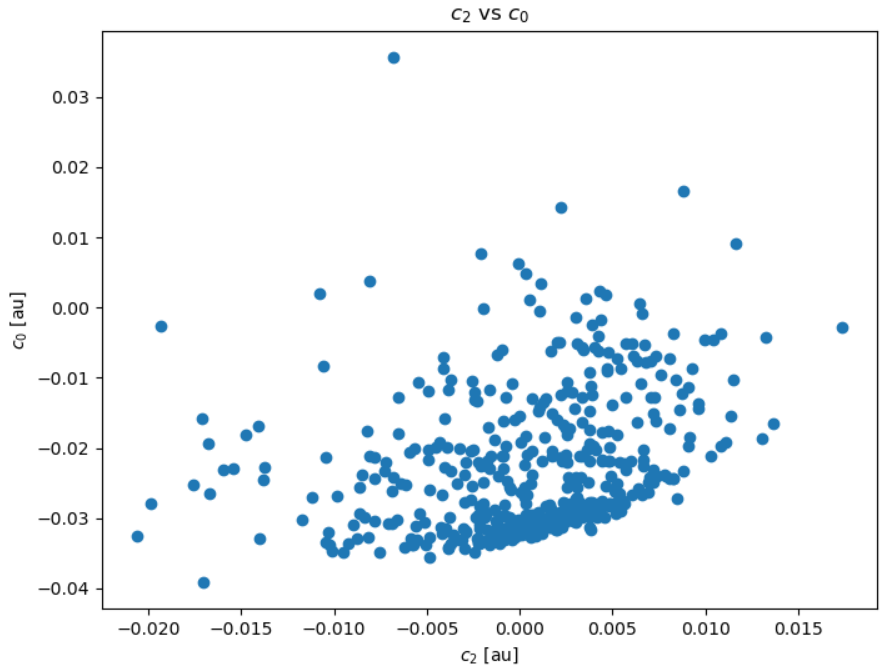

Figure 6: Plot of the first coefficient vs the third one in the PCA, for every galaxy.

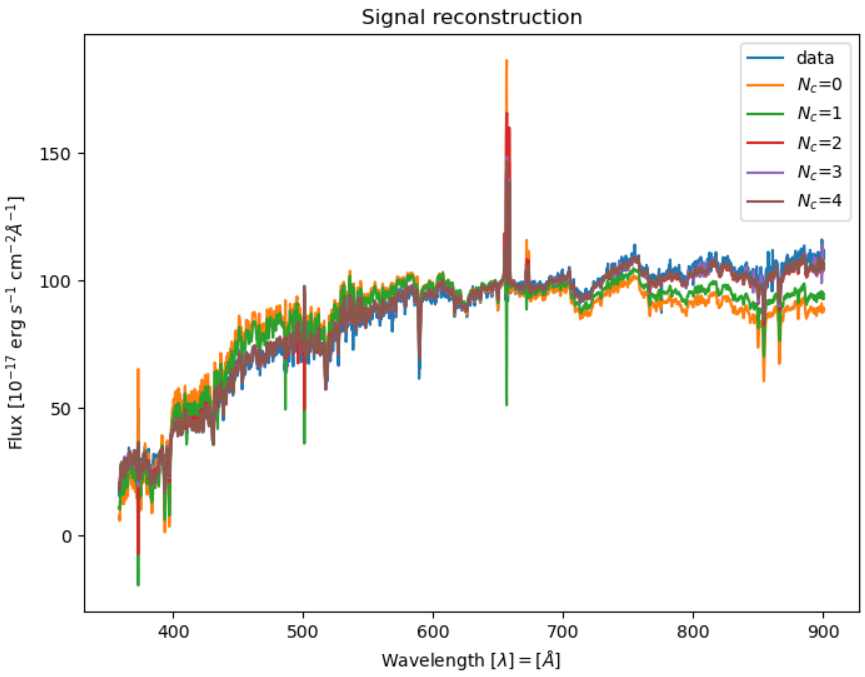

Figure 7: Plot of measured flux vs wavelength of first galaxy, together with the result of its approximation with $N_c = 1, \ldots, 5$ principal components.

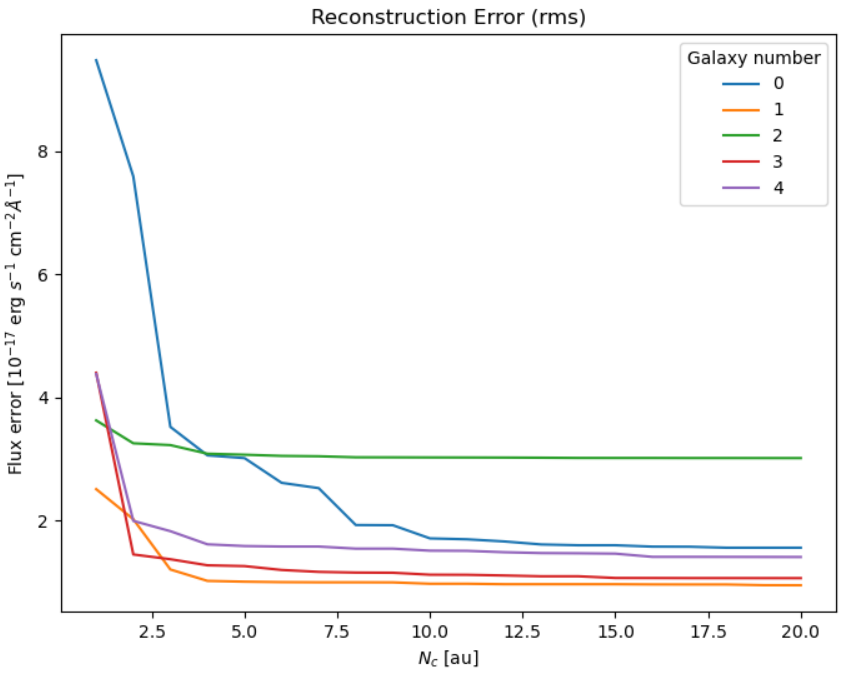

Figure 8: Plot of the rms error committed by approximating the first five galaxies via their first N_c principal components.