PS-6

Leen Alrawas

October 30, 2023

Abstract

This problem set deals with a real data set of galaxies spectra over a range of wavelengths. A simple implementation of Prinicpal Components Analysis was performed. Code scripts are available at `https://github.com/Leen-Alrawas/phys-ga2000/tree/main/ps-6`.

1 Methods and Results

1. Using the data of the central optical spectra of 9,713 nearby galaxies from the Sloan Digital Sky Survey, the first three spectra were plotted (fig 1). I also plot the same against the wavelength in nm units (fig 2). The presence of Hydrogen can be easily noticed in these galaxies from the spectra of Hydrogen. For example, one dominated spectrum line occurs at around 650 nm which coincides with the Balmer line of the Hydrogen spectrum.
2. The new spectrum is obtained by normalizing the original for each galaxy, then subtracting the mean. A plot of the new normalized spectra is provided for the first few galaxies (fig 3).
3. We find the eigenvectors of the associated eigenspectra using two methods. First, by constructing the covariance matrix and obtaining its eigenvectors, and second, directly from the residuals matrix using SVD. The results are the same as shown in figures 4 and 5.
4. The efficiency of the two methods is compared along with the condition numbers. The condition numbers are not exactly the same due to disagreement in the smallest eigenvalue (fig 6). SVD can be particularly useful when dealing with not-so-well-behaved matrices as it can deal with singular matrices by finding the best approximate eigenvalues.
5. Data can be approximated using PCA. Figures 7 and 8 show the approximated spectra for the first two galaxies using the first five coefficients. We can also map the new approximate spectra to the original fluxes range by doing the inverse of the transformation that was done in the first part (fig 9).
6. The plots of the first coefficient with the next two are shown in figures 10 and 11.
7. The error decreases using more coefficients in the approximation. In figures 12 and 13, the mean value of the residuals between the approximation and the spectra is plotted against the number of coefficients N_c for the first two galaxies. A snapshot of the fractional errors for $N_c = 20$ is demonstrated in figure 14.
8. The methods of finding the eigenvectors directly from the covariance matrix and using SVD are compared in terms of time consumption. SVD seems to be less efficient (fig 15).

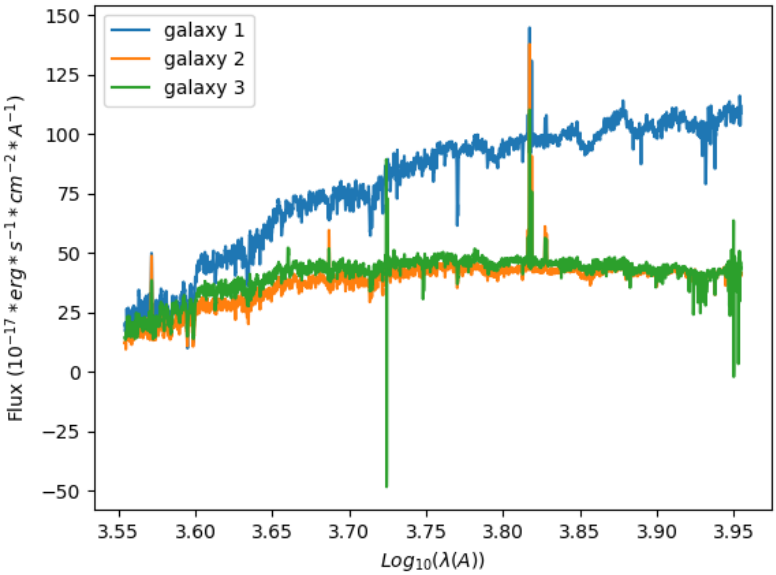

Figure 1: Central optical spectra of 3 nearby galaxies.

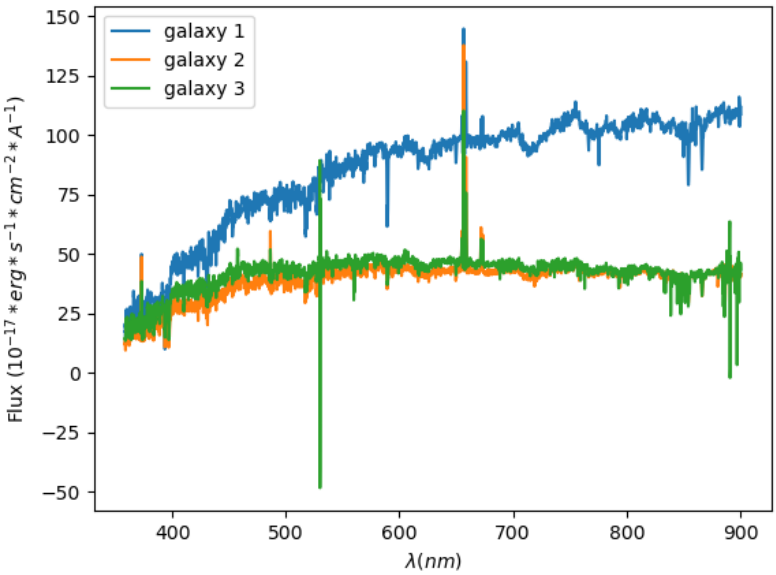

Figure 2: Central optical spectra of 3 nearby galaxies.

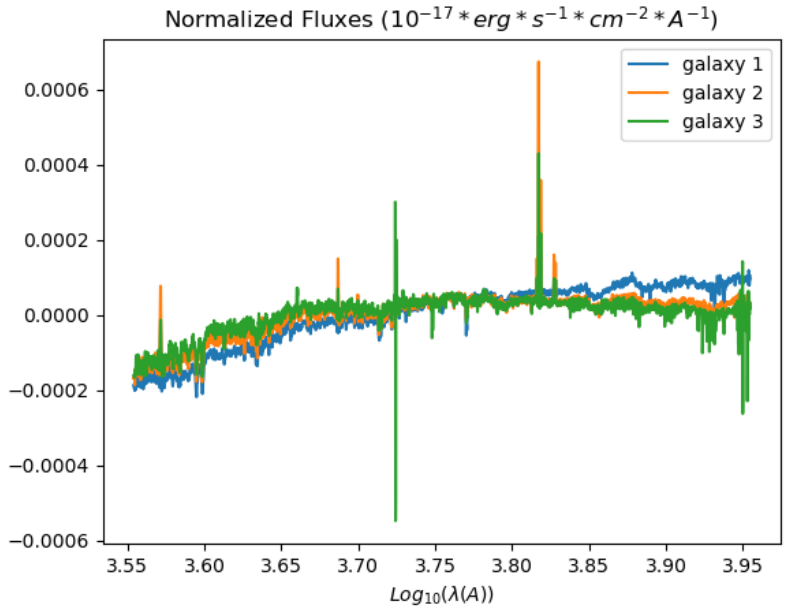

Figure 3: The normalized spectra of 3 nearby galaxies.

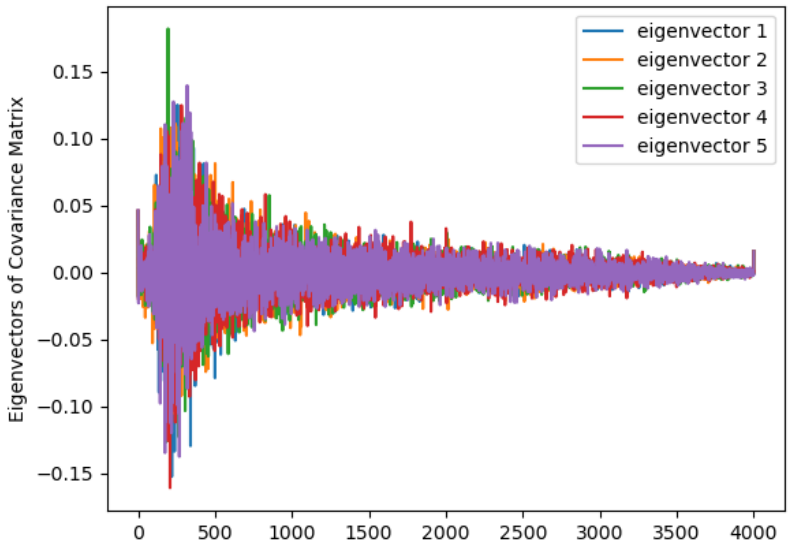

Figure 4: The first five eigenvectors using the covariance matrix.

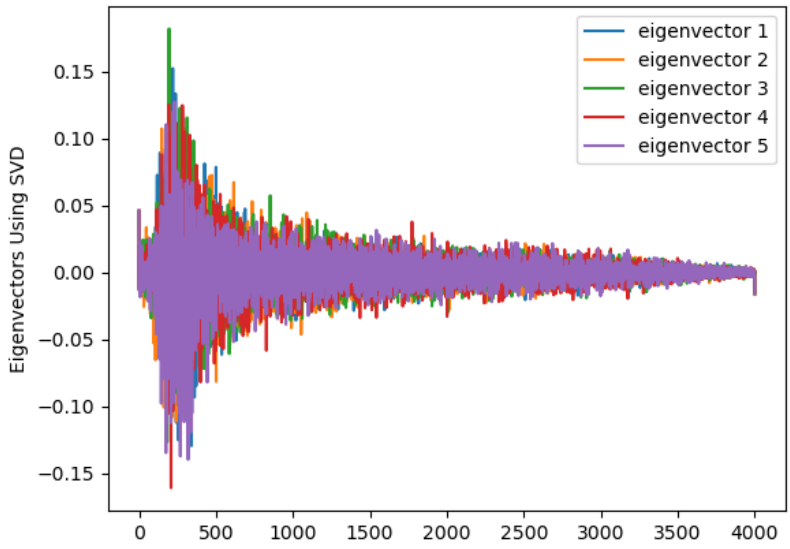

Figure 5: The first five eigenvectors using SVD.

```
Condition number: Using Covariance matrix = -37051940000.0
Condition number: Using SVD = 47130136000000.0
```

Figure 6: Condition numbers of the matrices in the two methods.

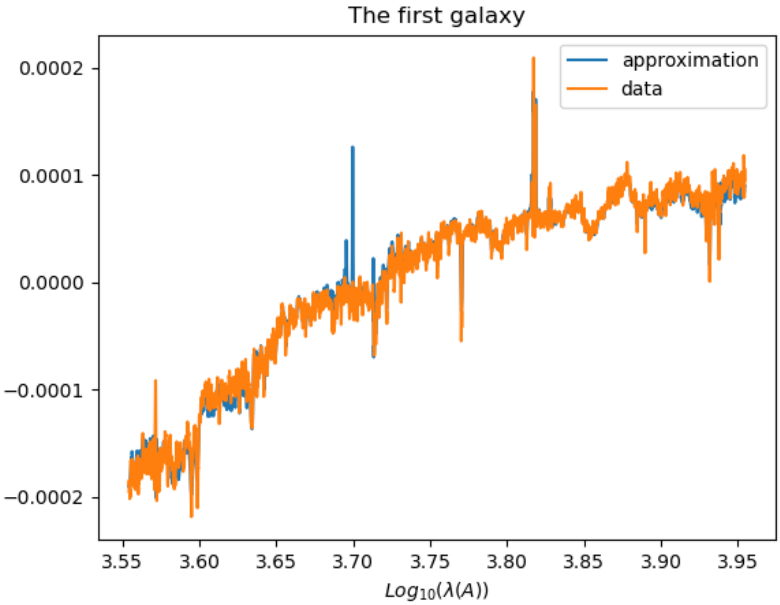

Figure 7: Approximated spectra using PCA for the first galaxy with 5 coefficients.

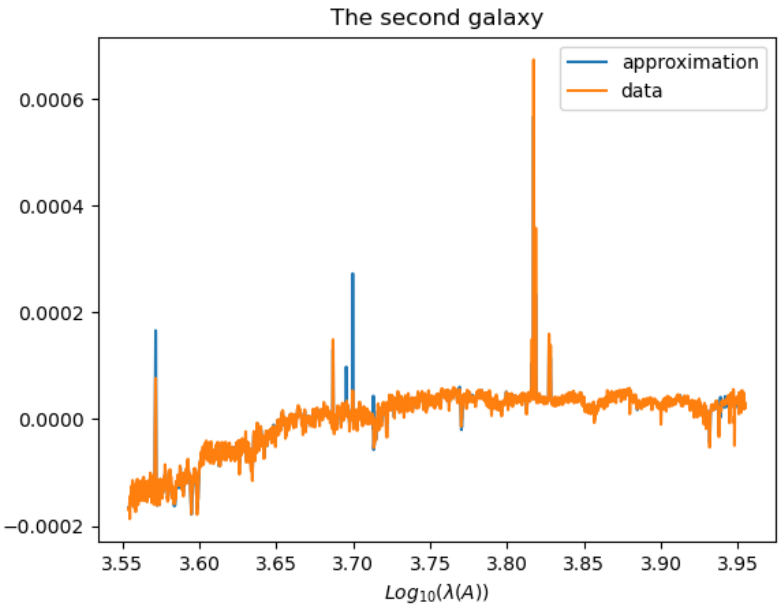

Figure 8: Approximated spectra using PCA for the second galaxy with 5 coefficients.

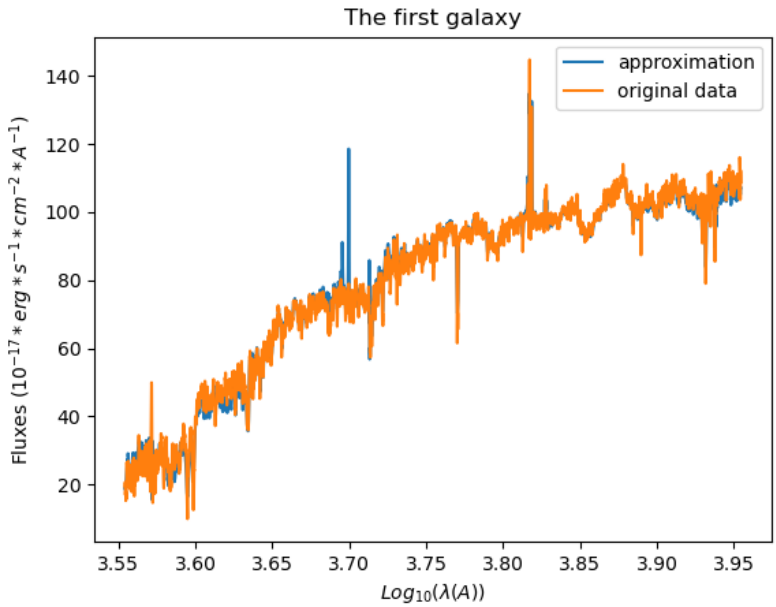

Figure 9: Approximated spectra using PCA for the first galaxy with 5 coefficients mapped back to the original data range.

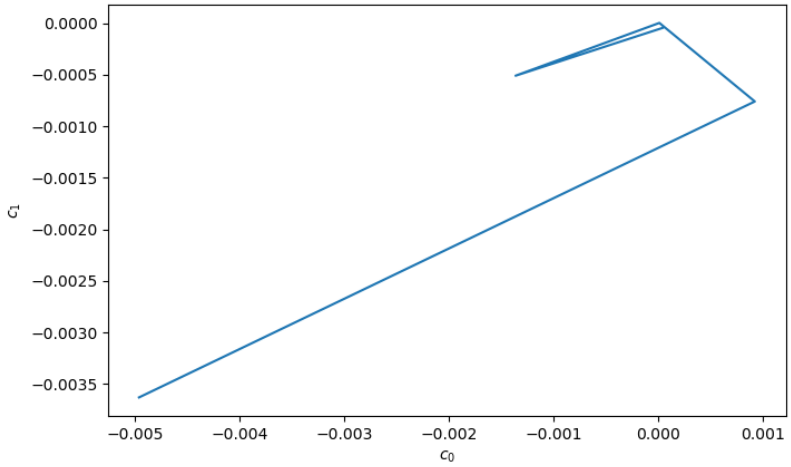

Figure 10: The coefficients in PCA approximation.

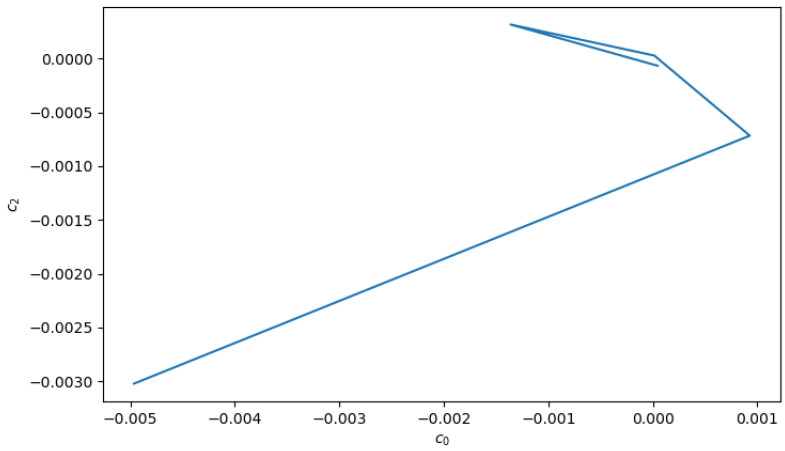

Figure 11: The coefficients in PCA approximation.

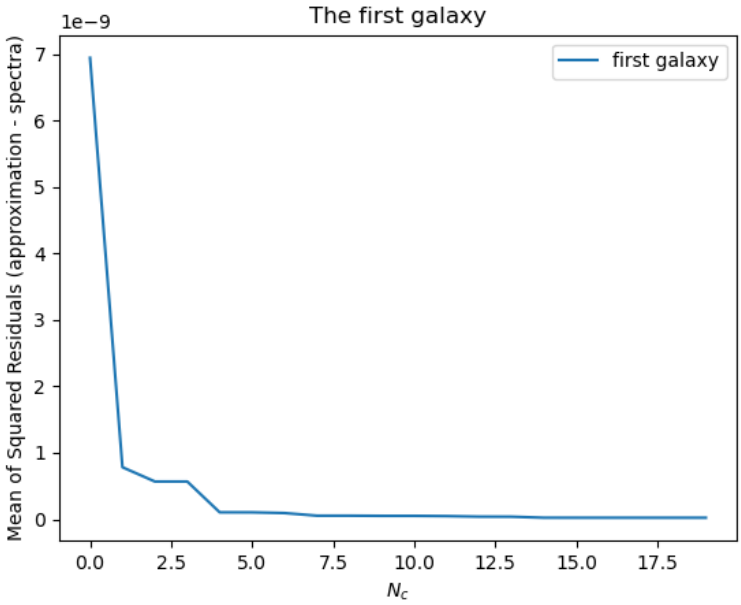

Figure 12: The mean value of the residuals between the approximation and the spectra against the number of coefficients N_c.

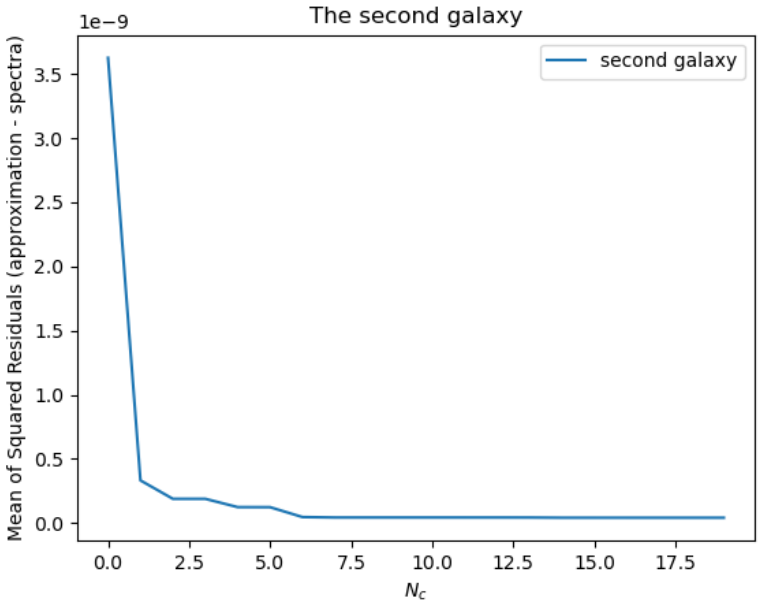

Figure 13: The mean value of the residuals between the approximation and the spectra against the number of coefficients N_c.

```
fractional_error [[8.3263174e-11 1.7970823e-10 3.1402463e-12 ... 1.1417308e-10
 9.6903326e-11 8.2505114e-11]
[3.8786009e-11 3.4694470e-12 4.7438380e-11 ... 5.2486847e-12
 8.3699478e-11 1.5746071e-12]
[3.8687286e-11 4.0300756e-11 1.4549208e-10 ... 2.1829413e-11
 8.3712132e-10 1.9575691e-10]
...
[9.6183228e-10 5.5317528e-09 4.9735105e-09 ... 3.5200485e-11
 1.4553246e-12 2.8048148e-12]
[9.2956047e-13 2.2927752e-11 6.9866286e-11 ... 1.0005174e-12
 1.6298085e-14 2.6693146e-11]
[1.7863380e-10 9.2841033e-11 6.3135226e-11 ... 2.8050848e-11
 9.4219507e-15 7.4208938e-11]]
```

Figure 14: Fractional errors using $N_c = 20$.

```
Time for Covariance matrix:  56.99385450000045
Time for SVD:  81.5372967000003
```

Figure 15: Efficiency comparison between the two methods implemented to find the eigenvectors.