PHYS-GA2000-PS6

Ahmet Koral Aykın

October 31, 2023

1 Introduction

In this problem set, we are interested in performing principal component analysis (PCA) on a real data set. The data consists of central optical spectra of 9713 near galaxies. The central optical spectra of the first five galaxies presented in Figure 1.

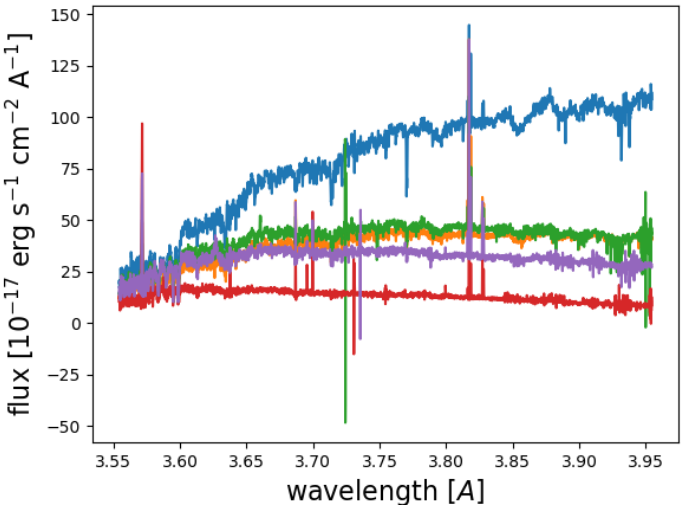

Figure 1: Central optical spectra of the first five galaxies. Each color corresponds to a different galaxy.

At first sight, the spectra reminds us the electron transitions in a hydrogen atom. The transition of an electron from one principal quantum number to another gives a peak at a particular wavelength that associated with the energy released or absorbed during the transition. Analogous to hydrogen atom spectrum, based on where the peak is located, we could infer useful information about the galaxies.

2 Methods

To make the PCA more meaningful, first we normalize the each spectrum such that the integral over the wavelength is equal to 1. After normalization, we subtract the mean from each spectral array. To check whether the spectrum of each galaxy is properly normalized, we can plot the sum the flux data over each wavelength (see Fig. 2). In Figure 3, normalized and zero mean spectrum of the 0th galaxy is presented.

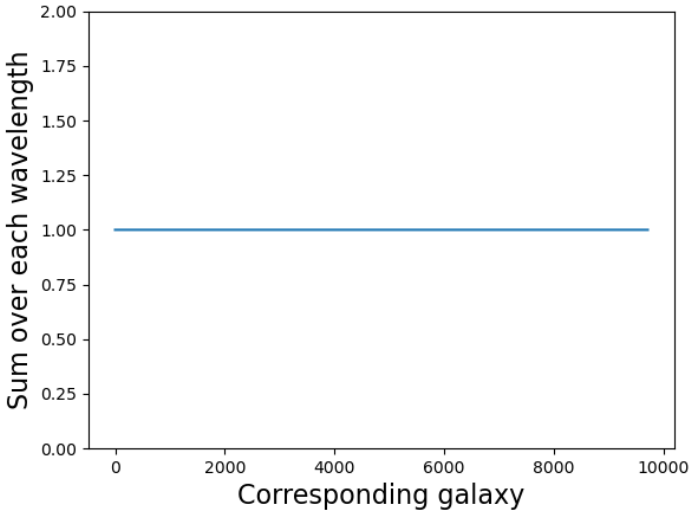

Figure 2: Sum of flux over each wavelength for each galaxy.

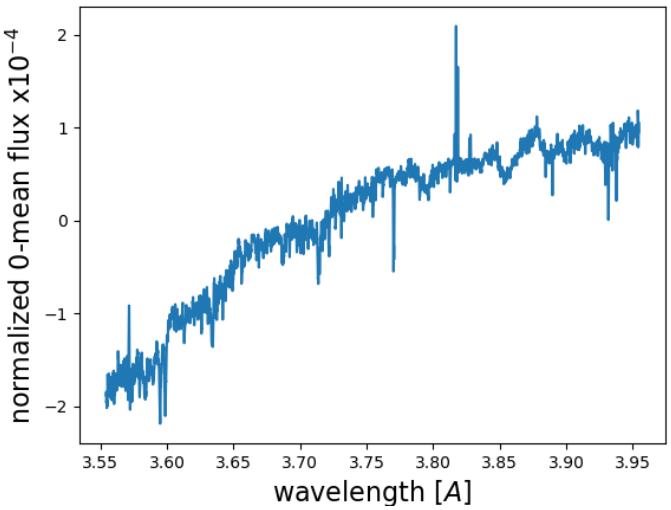

Figure 3: Normalized and zero mean flux plot of the 0th galaxy.

3 Results

The first five eigenvectors which are found via EIG method are presented in Figure 4. In addition to EIG, singular value decomposition method (SVD) is also utilized to find eigenvectors. The eigenvectors are plotted against each other for two different methods and presented in Figure 5. One can notice that some of the eigenvectors found via SVD reflected across x = 0. This is a result of the fact that the eigenvectors calculated this way are unique up to a sign-change. On the other hand, as seen in Figure 6, the eigenvalues are the same.

It is of great importance to state that SVD method is faster since it is easier to decompose the matrix. The condition number of the R found as 6561841.5 while that of C is 62384247000.0. There is a 4 order of magnitude difference between them, which implies SVD is advantageous over EIG method.

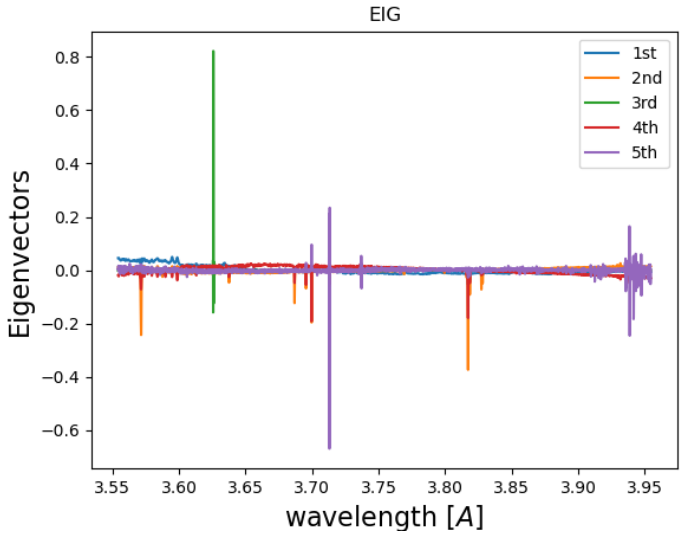

Figure 4: First five eigenvectors.

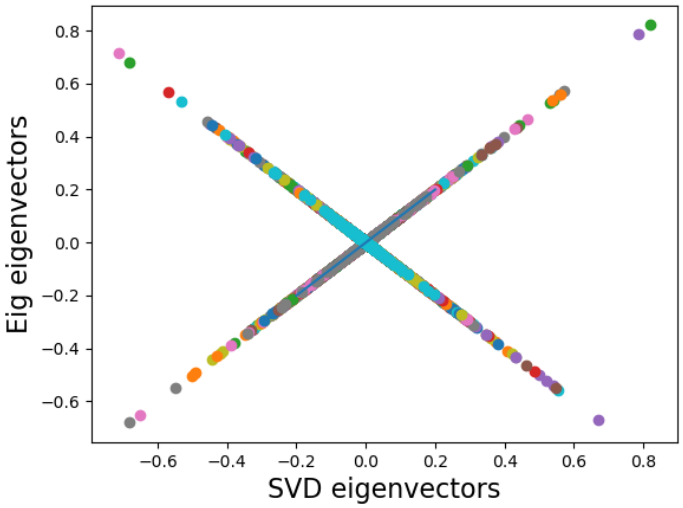

Figure 5: Eigenvector comparison of the two different methods.

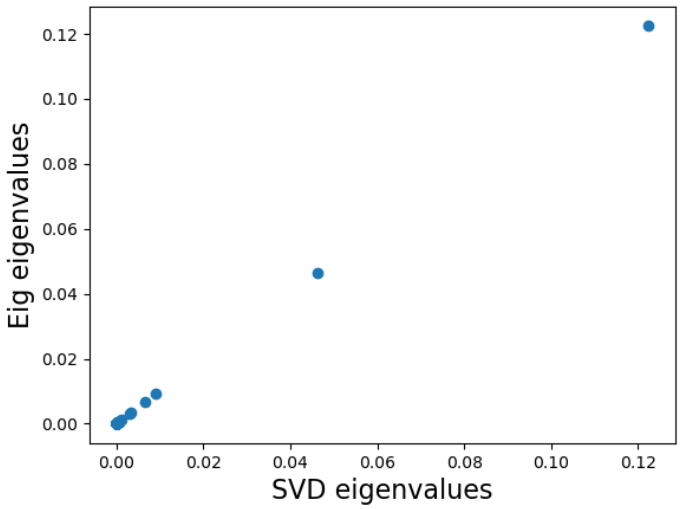

Figure 6: Eigenvalue comparison of the two different methods.

The reconstituted data based on keeping only 5 coefficents is presented in Figure 7 for the 0th galaxy.

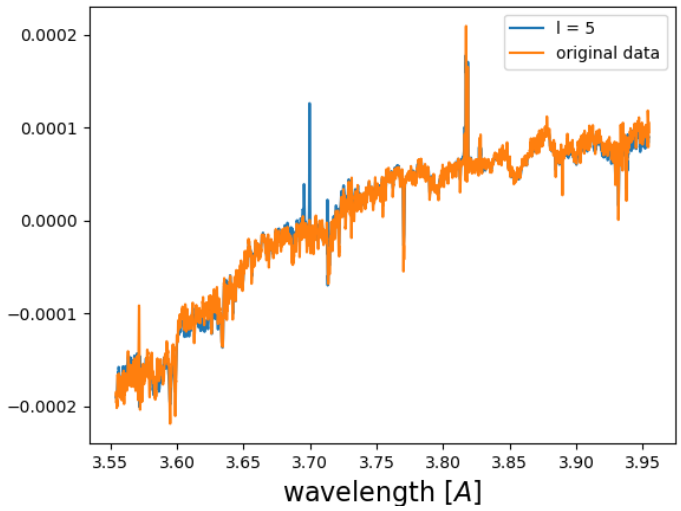

Figure 7: Reconstituted data using Nc (l) = 5 eigenvectors and original data.

The C0 vs. C1 and C2 are presented in Figure 8 and 9, respectively.

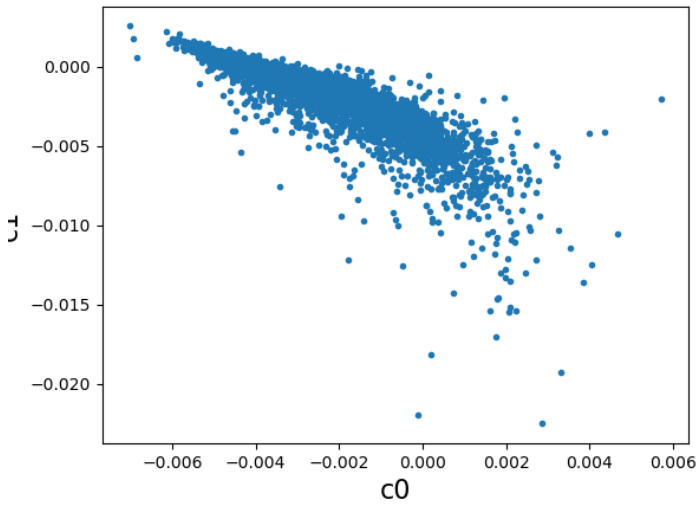

Figure 8: C1 vs. C0.

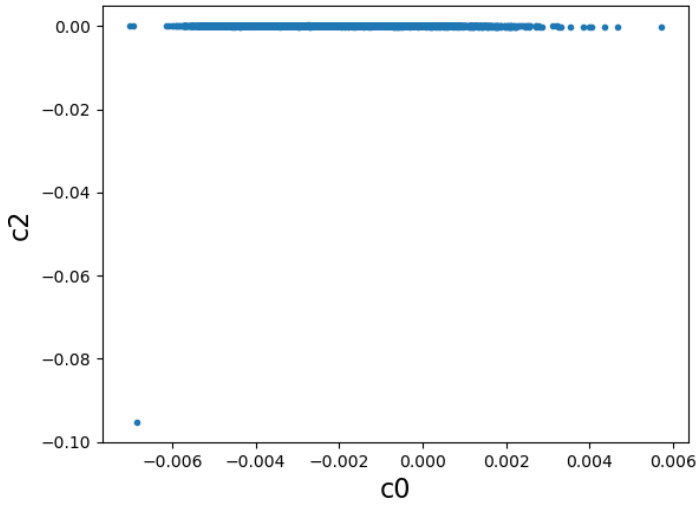

Figure 9: C2 vs. C0.

The residuals for the 0th galaxy over the wavelength for different number of coefficients (from $N_C = 1$ up to 20) are calculated. For a better visualization, the residuals for 3 different N_C are presented in Figure 10. Finally, the squared

fractional error for $N_C = 20$ is on the order of 10^{-4}. All residuals can be found in the code.

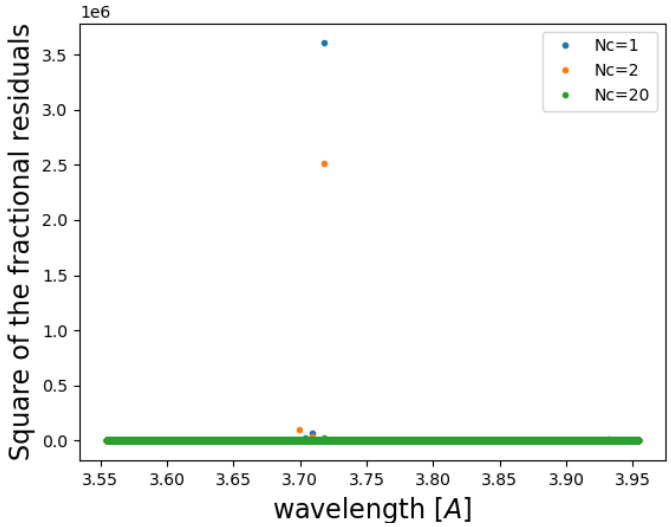

Figure 10: Square of the fractional residuals.