Problem Set 6

Edison Wang

Nov 3, 2023

Problem 1

(a) A plot of the first 5 galaxies is shown in Figure 1. I notice that there is a spike in the flux at the wavelength of $\lambda = 10^{3.82}$ angstrom, which is about 660 nm. This is a wavelength in the emission spectrum of hydrogen corresponding to red light. Therefore, a large part the the emission of these galaxies is due to electrons "dropping" from higher energy levels to lower energy levels inside hydrogen atoms (a process that emits photons).

(b) Each row, representing a galaxy, is divided by the sum of that row.

(c) The residuals are calculated for this part.

(d) The covariance matrix C is obtained by multiplying the transpose of the residual matrix with the residual matrix itself. The eigenvector matrix U is calculated by using the Python function `scipy.linalg.eig`. The first five eigenvectors are shown in Figure 2. The x-axis is chosen to be "logwave", but in fact it has just the same effect as choosing it to represent "i-th component of the eigenvector".

(e) Given the following SVD of the residual matrix R:

$$R = U \cdot W \cdot V^T,$$

the matrix V is composed of the right eigenvectors of $R^T R$. We note that for such a right eigenvector and a corresponding eigenvalue λ_i,

$$R^T R v_i = \lambda_i v_i$$
$$R R^T (R v_i) = \lambda_i (R v_i),$$

whereas for some right eigenvector u_i of $R \cdot R^T$, we have

$$R R^T u_i = \lambda_i u_i.$$

Therefore, given V, we can recover the eigenvector matrix U obtained in Part (d) by simply left multiplying V by R,

$$U' = RV,$$

and then normalizing the eigenvectors within U'. Thus, the two methods used in Parts (d) and (e) are equivalent. The eigenvectors in U recovered by the SVD method are shown in Figure 3, which are exactly the same as the eigenvectors shown in Figure 2.

The computation time for the method in Part (d) is about 34.9 s, while that for the SVD method is about 65.6 s. Therefore, the SVD method used for this part is slower than the former method.

(f) Even though the SVD method is slower, it is numerically more stable. The condition number of the covariance matrix C obtained in Part (d) is calculated to be about 17425996000.0, while that of the residual matrix used in Part (e) is 6561841.5. Clearly, the second (SVD) method deals with a matrix that is less close to singular and is thus numerically more stable. In fact, the numerical instability of the "covariance matrix" method manifests itself during the computations: a warning that says "casting complex values to real values discards the imaginary part" pops up, but this should not happen in theory since RR^T is a symmetric matrix and thus should have real-valued eigenvectors. The SVD method does not have this issue.

(g) The coefficients used for multiplying the eigenspectra come from the V matrix obtained in Part (e), while the eigenspectra (eigenvectors) themselves are given in the matrix U'. Since SVD guarantees $V^{-1} = V^T$, we can recover the transpose of the residual matrix used in Part (d) by simply multiplying V by the transpose of U'. Then, we can add the mean spectrum to the recovered residuals and multiply the results by the normalization values to recover the original data.

For this part, I approximate the original data by using only the first $N_c = 5$ coefficients (and thus 5 eigenvectors), which involves truncating the matrices V and U'. The approximated data are plotted and shown in Figure 4; the shapes of the curves are similar to those in Figure 1 (which are the original data).

(h) The coefficients c_i are actually sets of coefficients (numbers) that are represented by the columns of the matrix V, so each c_i can be taken as a vector. The component-to-component ratios c_0/c_1 and c_0/c_2 are plotted and shown in Figure 5. We can see that the components of c_0 are usually larger in magnitude than the corresponding components of c_1, which are in turn usually larger in magnitude than those of c_2. This makes sense since we expect the first eigenvector to have the most weight.

(i) The total squared fractional residuals are calculated for each value of N_e, by summing up the squared fractional residuals for each wavelength for each galaxy. These total squared fractional residuals are shown in Figure 6. There is a general trend for the residuals to decline.

My Github link: https://github.com/ziqiu-wang/phys-ua210.

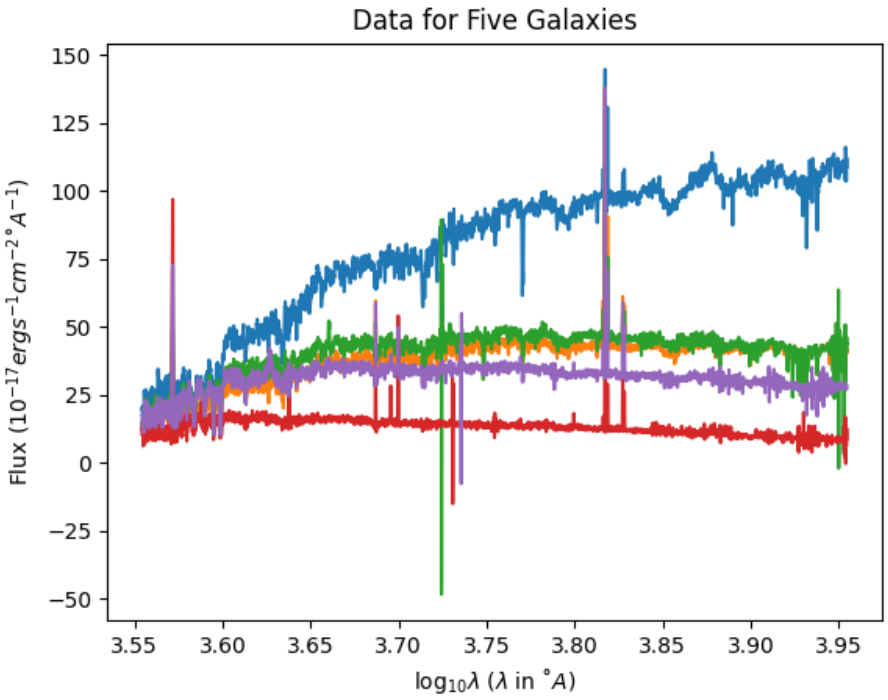

Figure 1: Plot of the flux for the first five galaxies.

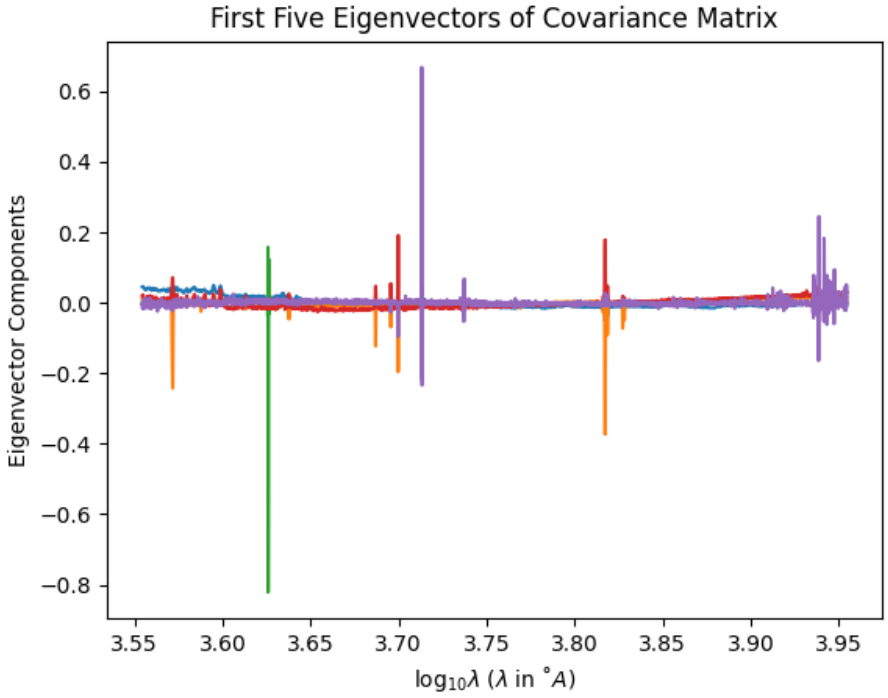

Figure 2: The first five eigenvectors of the covariance matrix, with individual components plotted along the horizontal axis.

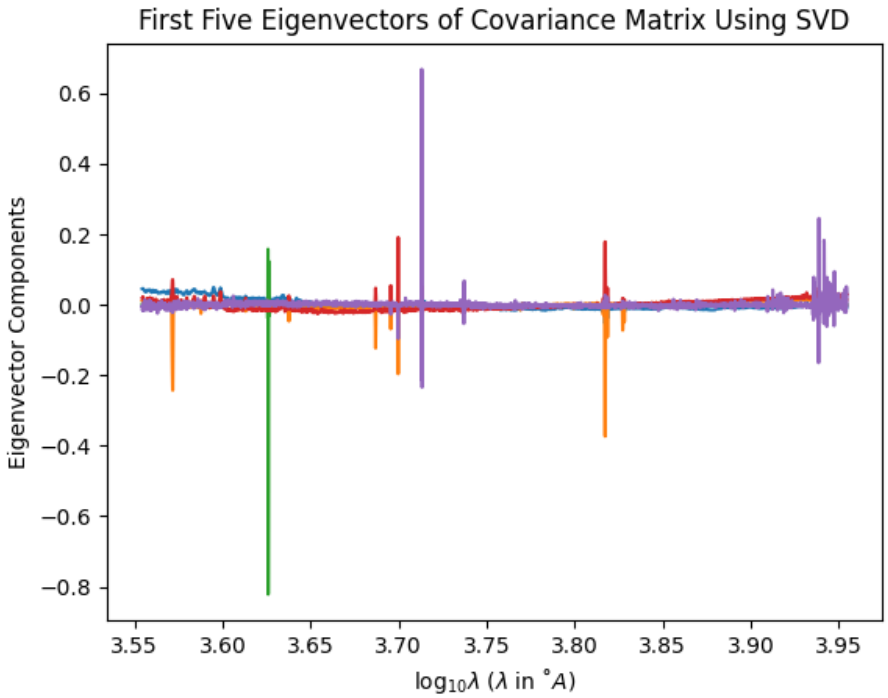

Figure 3: The first five eigenvectors recovered using the SVD method. Note that they look identical to those in Figure 2, as expected.

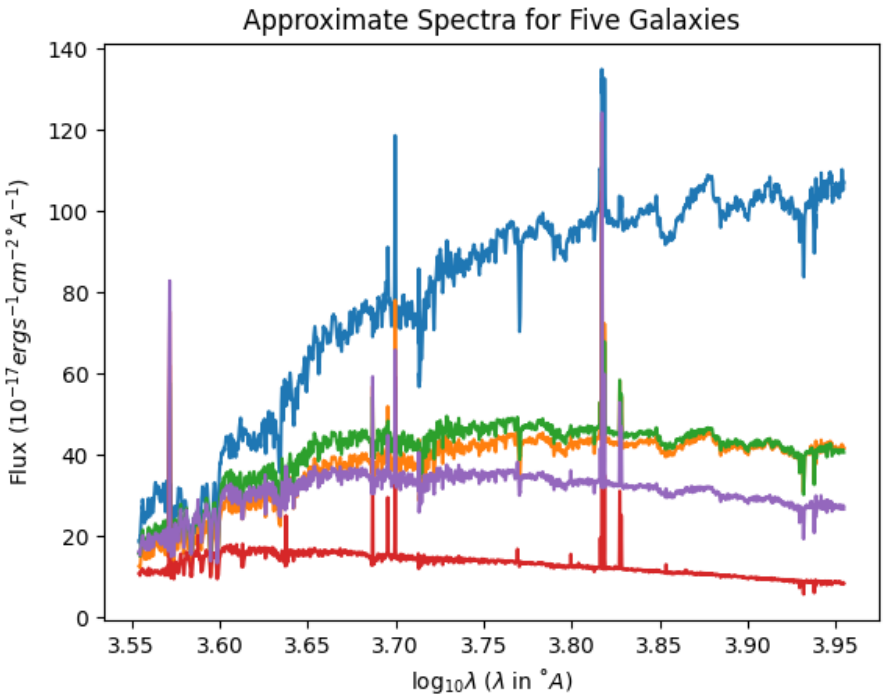

Figure 4: Approximations of the original data by using only the first five eigenvectors. Note the curves' similarity to those in Figure 1.

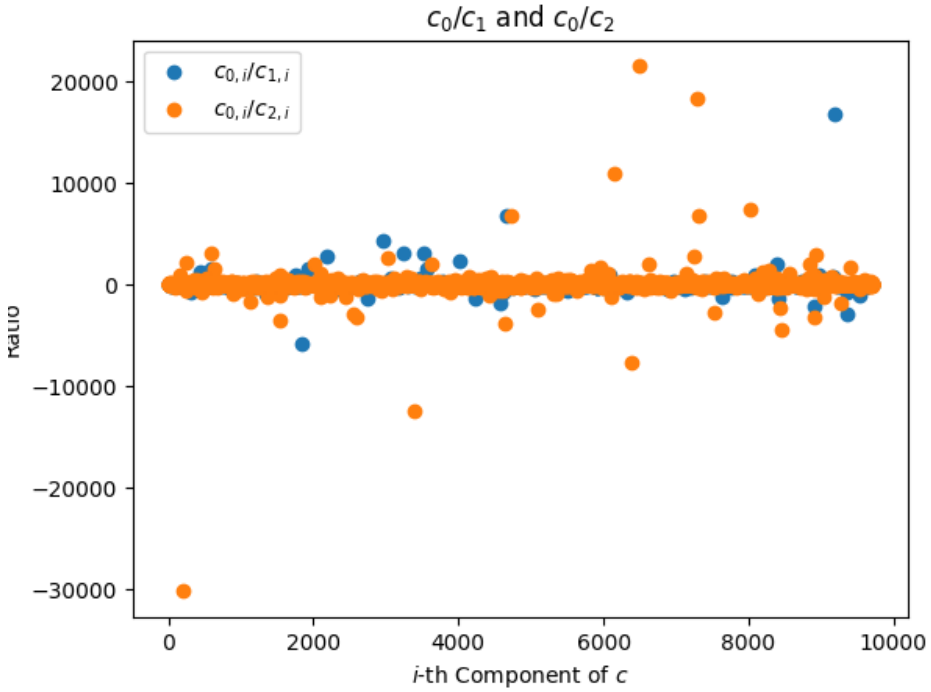

Figure 5: Component-wise ratios between the coefficients.

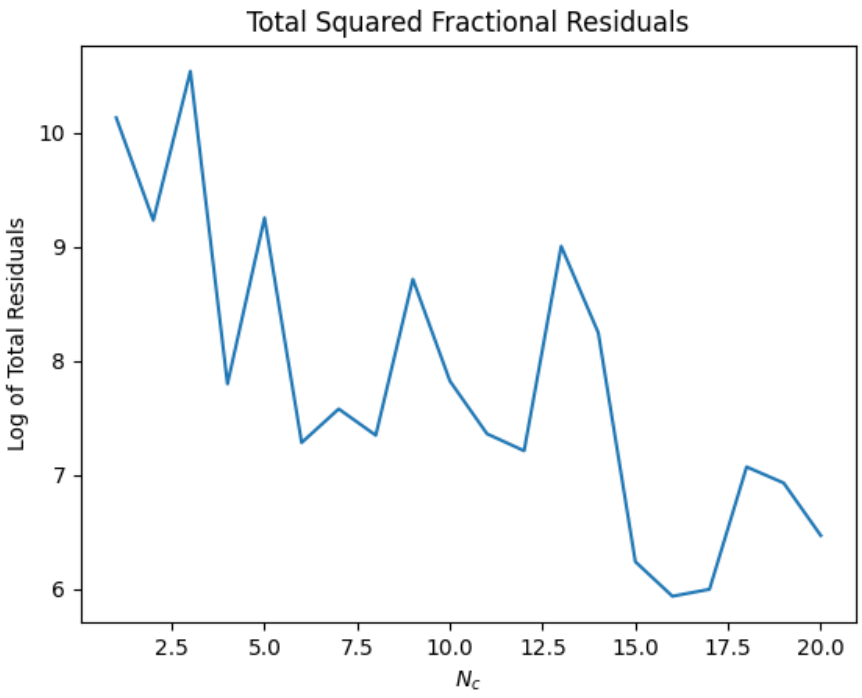

Figure 6: Total squared fractional residuals for different N_c.