Problem Set 6

By Hayden Orth; GitHub: haydenorth

October 31, 2023

Abstract

This article contains my solutions to Problem Set 6 for the graduate Computational Physics course. Problem Set 6 investigates linear algebra and eigensystem methods for data analysis.

1 Principal Components Analysis

Problem 1 involves performing a Principle Components Analysis on a data set of optical spectra of 9,713 nearby galaxies.

1.1 Parts (a-c)

First, I read in the data from the fits file using the astropy method and the provided code snippet. I convert the wavelength data from a log scale in Angstroms to a linear scale in nm. I then plot the spectra of four galaxies on the lienar scale, as shown in Figure 1. The peaks in the galaxies' spectra occur at wavelengths similar to those of Hydrogen's spectral series. Particularly, a large peak occurs at a wavelength of around 650 nm.

I then normalize the galaxy flux data by weighting the data of each galaxy such that the galaxy's spectral data sums to one. Next, I calculate the residuals for each galaxy by subtracting the average normalized flux from the normalized flux. Now I have a matrix of residuals values that can be used to calculate the covariance matrix.

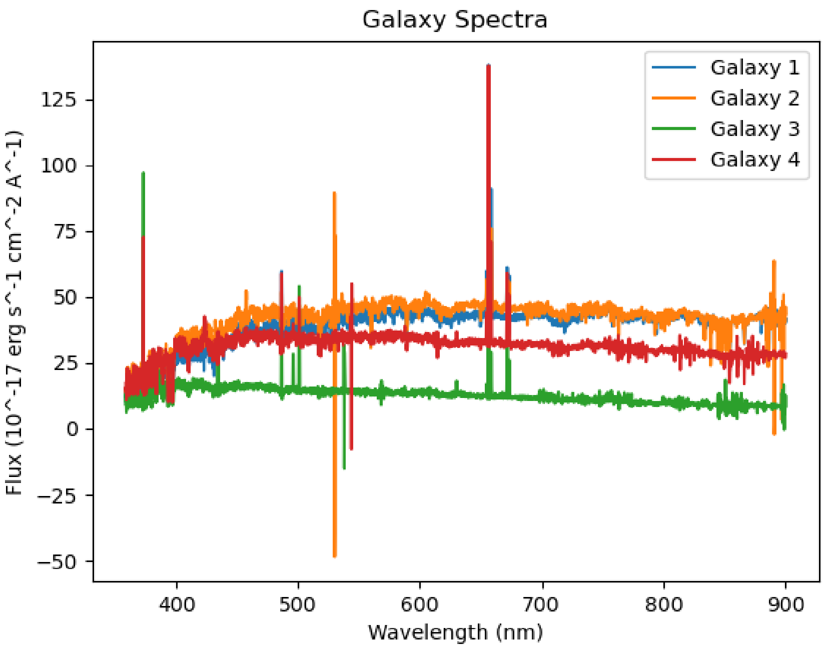

Figure 1: Plot of galaxy spectra for four of over 9,000 galaxies in the data set.

1.2 Part (d)

It is now time to perform the PCA with this matrix of residuals. I construct the covariance matrix C by multiplying the residual matrix R with its transpose. Then, I use NumPy's linalg library to compute the eigenvalues and eigenvectors of C. This operation took my computer 49.1 s to complete. The first 5 eigenvectors of C are plotted in Figure 2.

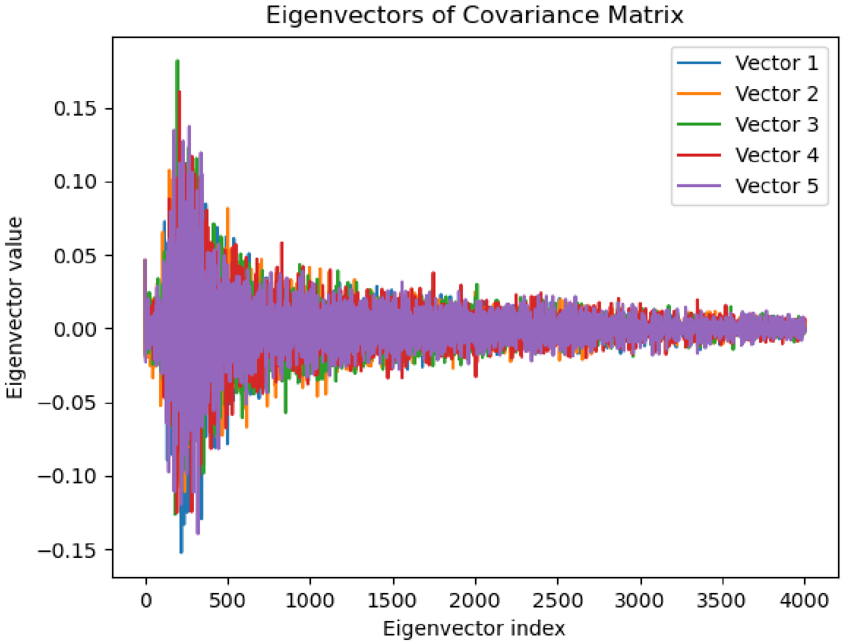

Figure 2: Plot of the first five eigenvectors of the Covariance Matrix C.

1.3 Parts (e-f)

Then, I find these eigenvectors directly from R by performing a SVD on R and its transpose. I used NumPy's SVD method to perfrom the calculation, which took 102.6 s to complete. Thus the PCA method from part (d) is around twice as fast as this SVD method. The eigenvectors are the same as those found in part (d), as shown in figures 3 and 4.

I found the condition number of C to be around three orders of magnitude larger than that of R (C=7.98E18, R=4.70E15). Both have very large condition numbers, meaning that the matrices are bordering on singular. R has a smaller condition number, so its results may be more reliable. However, performing the SVD on R took around twice as long as it took to perform the PCA with C.

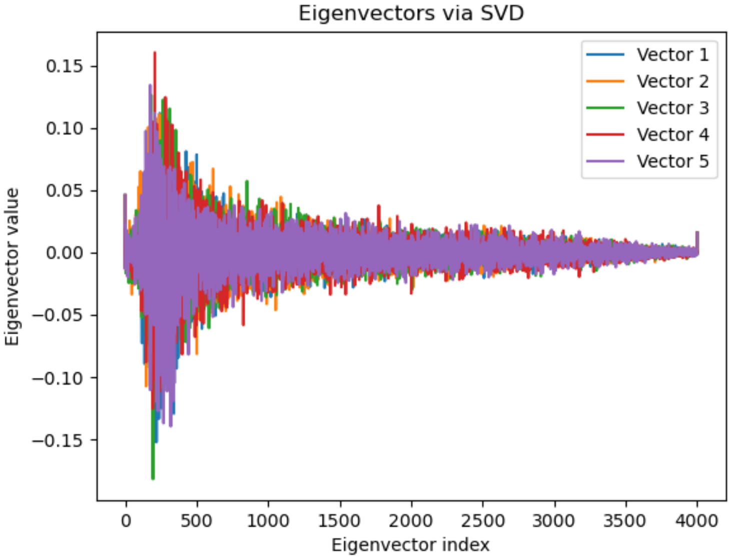

Figure 3: Plot of the first five eigenvectors calculated via the SVD method.

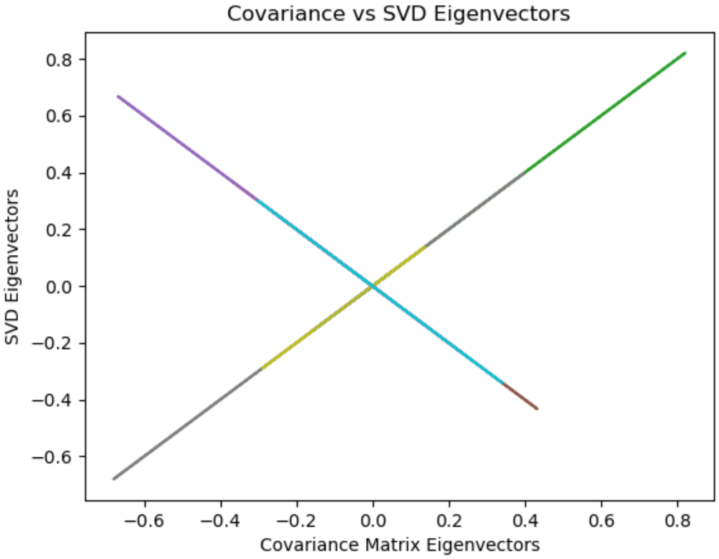

Figure 4: Plot of the first ten eigenvectors of the Covariance Matrix C versus the first ten SVD eigenvectors.

1.4 Parts (g-h)

Next, I rotate the spectra onto the eigenspectrum basis in order to get the necessary weights (coefficients) to approximate the data using its eigenvectors. The approximate spectra found using only the first five eigenvectors is shown in Figure 5. It looks pretty good. Figure 6 is a plot of the weights of the first eigenvector (c0) versus the weights of the second and third eigenvectors (c1 and c2). We can see that the weights of c1 and c2 both hover around 0, with c2's weights being almost exactly equal to zero.

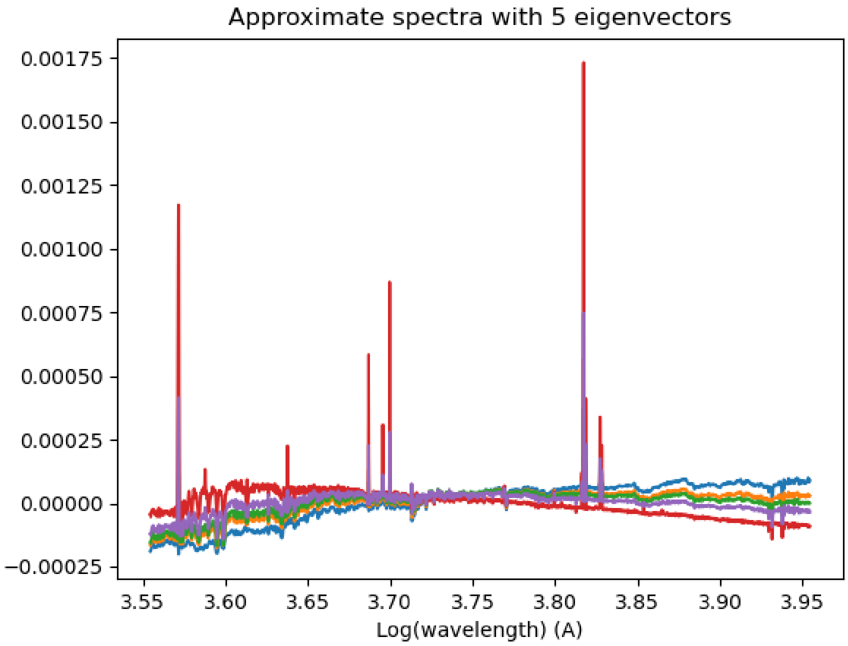

Figure 5: Galaxy Spectra approximated by its first five eigenvectors.

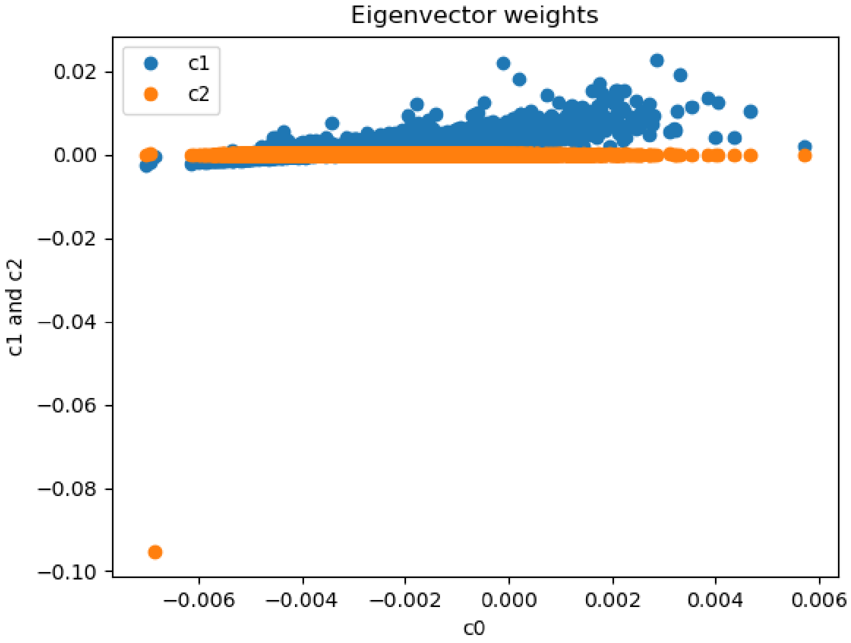

Figure 6: Plot of the weights of the first eigenvector versus the weights of the second and third eigenvectors.

1.5 Part (i)

Finally, in order to see how the approximation improves as the number of eigenvectors used is increased, I plot the squared residuals as a function of N, the number of eigenvalues used in the approximation. For each number of eigenvalues, N, I calculate an an approximated data set and calculate the squared residuals of this approximate set. Then, I sum the residuals to get an overall residual error for the approximation. I then plot these total residuals as a function of N, and the result is shown in Figure 7. We can see that the residual is falling off exponentially as N is increased. The fractional error for Nc = 20 is on the order of 10^{-8} (the total residual is what I plotted, which is a sum of the fractional error across the whole spectrum).

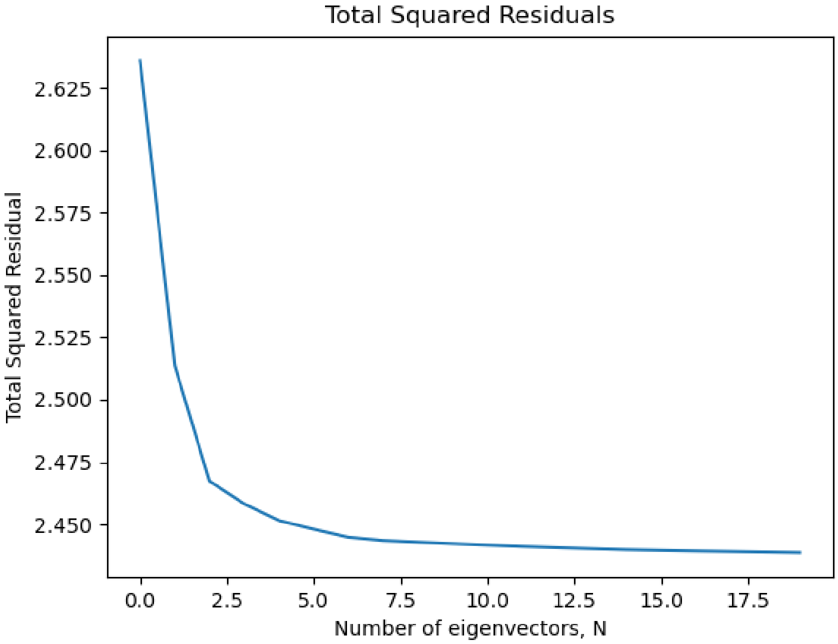

Figure 7: Plot of the square residuals of the approximated data set as a function of the number of eigenvectors used in the approximation.