Ps 6 Problem 1

Adam Guerin

November 1, 2023

Abstract

This short paper will show some plots about and discuss using Principal Component Analyses on a data set for galaxies and their fluxes.

1 Introduction

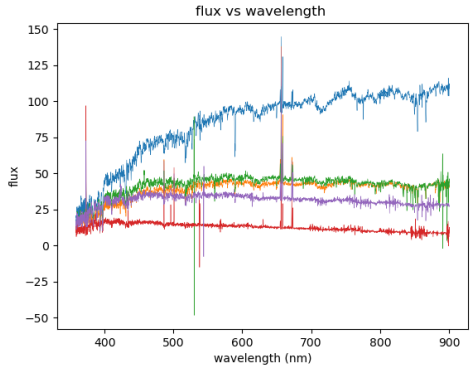
Principal Component Analysis is a technique for breaking up large data sets with many dimensions (in our case the fluxes of 9713 galaxies across 4001 wavelengths) into a series of eigenvectors that describe the system. When included in full the eigenvectors form a basis of your data set, but by only taking a subsection of those eigenvectors, you will have effectively reduced the dimensionality of your problem. If your data is correlated well, you will only need a few of the eigenvectors to describe your system to a high accuracy.

2 Methods

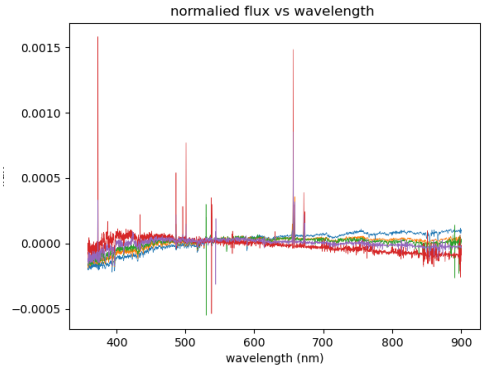
After importing the collection of galaxies and their fluxes, I then normalized the data and subtracted the mean so that every galaxy had the same flux and is centered around zero. With this I can look at each wavelength's relative importance across every galaxy at the same time. From there I did PCA and analyzed the data.

PCA was done in two different ways to compare: once by constructing the covariance matrix from the normalized data and then finding the eigenvectors of that matrix and then also by doing SVD decomposition on the residuals and using the eigenvectors from that.

Eigenvectors need not be unique, so these are slightly different between the two, but both form bases and can be used to reconstruct the data. See figure 3 for more information. There is a ratio of 1600 between the condition numbers for the two methods with the SVD method being the worse of the two. It also took nearly twice as long for me to to run SVD on R as it took to just calculate the Correlation. I wouldn't use SVD unless we had much much more frequencies we were going over. The correlation matrix is 4001 by 4001 so it's not that much different than the 9713 by 4001 matrix that is our data. If instead we had 500 galaxies and 100,000 frequencies, it would make more sense to do SVD on the 500 by 100,000 data than to try and calculate the eigenvectors for a 100,000 by 100,000 vector.

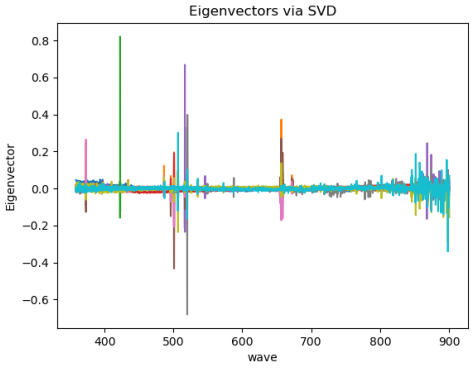


(a) Original Data

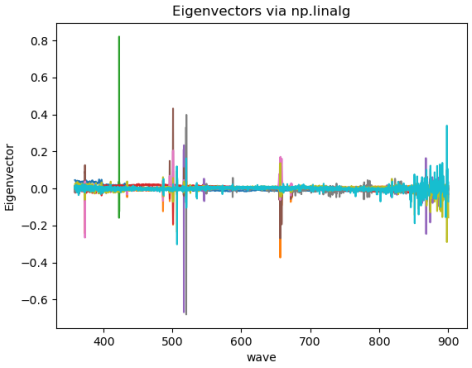


(b) Normalized Data

Figure 1: Because galaxies are all at different distances and angles in the sky and many other reasons, the actual flux that comes in will differ between each galaxy. By normalizing the data we can compare wavelength composition more easily. You can clearly see the first hydrogen spectrum transition line at 656 nanometers.



(a) Ten eigenvectors with the largest associated eigenvalues found via SVD on the residuals.



(b) Ten eigenvectors with the largest associated eigenvalues found by decomposing the entire correlation matrix

Figure 2: By finding the correct weights and summing up the eigenvectors with those weights, you can create every galaxy flux.

3 Results

All of this is to say that I was able to find the principal components of the data and reconstruct it either in part or in total. See figures for more details.

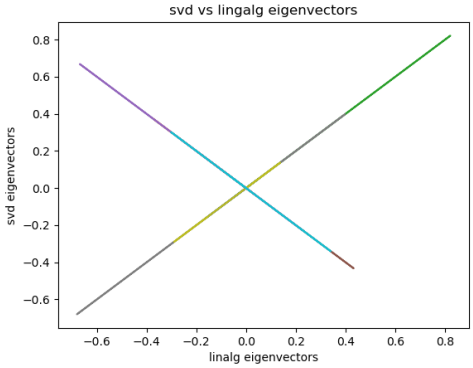
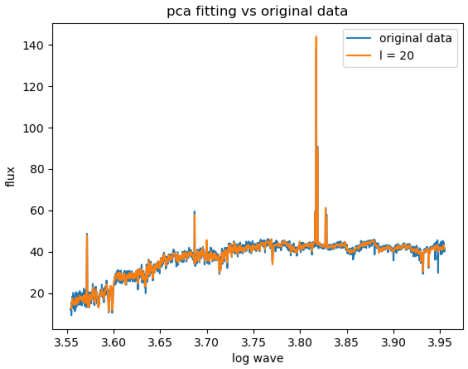
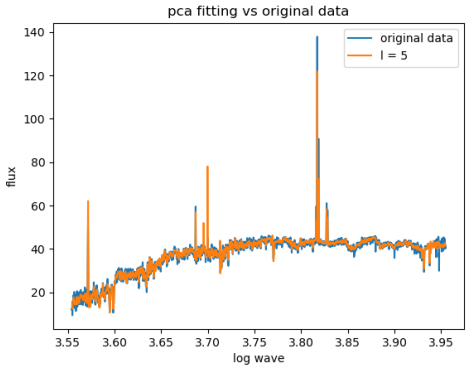


Figure 3: You can see that the eigenvectors for each of the two methods do line up, except that the ones found via SVD are sometimes negative when the ones founbd via numpy linalg on the correlation matrix aren't. But they line up linearly so its fine.



(a) Data reconstructed with 20 principal components

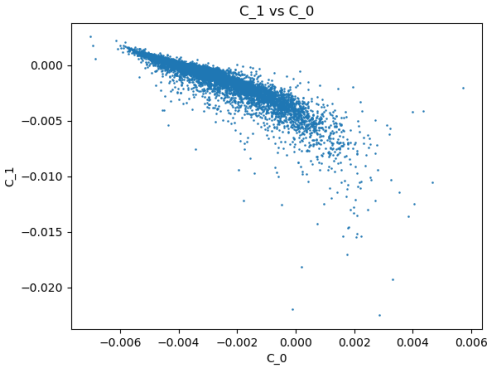


(b) Data reconstructed with 5 principal components

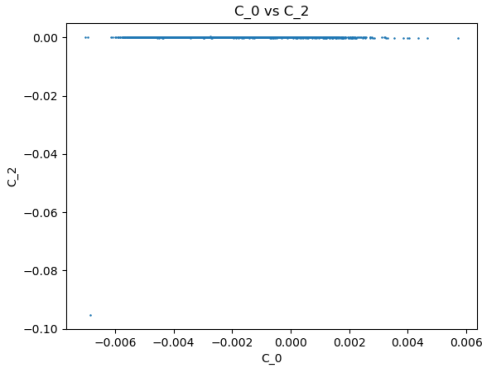
Figure 4: You can see how with only 20 out of 4001 components we are still pretty accurate and capture all of the major structures of the data. There is a lot of noise being left out, but including all principal components gets all of the fine details.

4 Discussion

PCA is a very useful method for storing and working with large data sets, if done well, you would only need to work with your entire data set a few times, and then you can simply work with the equivalent of a 'zipped' version of your data by using principal components and normalization factors.



(a) Weights for first principal component vs second principal component



(b) Weights for first principal component vs third principal component

Figure 5: These are the weights plotted against eachother for differing principal components. If you were to zoom in on the C0 vs C2 plot it would look similar to the C0 vs C1 plot, but the former has a very heavy weight for a single galaxy that makes the plot look less nice and pretty.

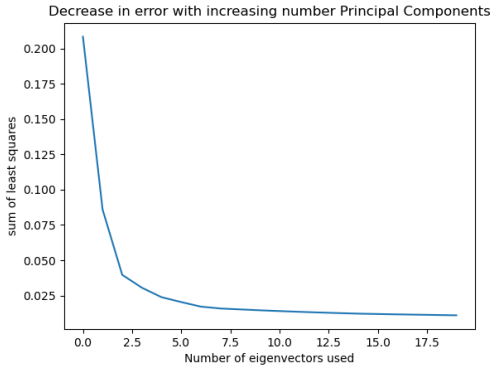


Figure 6: Relative error decreases very quickly at first and then levels off. The fractional error at 20 Principal components was ten percent, very good for using only .5 percent of the available components.