[PHYS-GA2000] Problem Set 6

Dylan Lane
Github: dl4729

October 31, 2023

Methods and Results

I began by loading the data with Astropy and plotting the flux against the wavelength for the first five galaxies, creating Figure 1. Figure 2 adds black vertical lines for the visible part of hydrogen's emission spectra scaled logarithmically; the peaks of the flux data align closely with the Balmer spectrum of hydrogen (3). The fluxes were then normalized such that the sum of data points for a given galaxy (the integral over wavelength, approximately) is 1. This can be accomplished simply by taking a sum over axis 1 and doing vector division on the transpose of the data matrix, then transposing back. Accordingly, adjusting the data to zero mean by calculating the per-galaxy mean, doing matrix-vector subtraction on the transposed data, then transposing back.

For Principal Component Analysis, begin by constructing the covariance matrix $C = R^T R$ (where R is the normalized adjusted flux matrix) and finding the eigenvectors and eigenvalues. Figure 3 shows the first five eigenvectors found this way sorted by their eigenvalues. The same computation can be achieved with SVD acting directly on R, where the eigenvectors can be extracted from the matrix V^T. Intuitively, eigendecomposition of the covariance matrix is more computationally efficient than SVD because it takes advantage of the fact that the covariance matrix is square, computing a single set of eigenvectors and eigenvalues instead of both a left and right set (one from $R^T R$, one from RR^T). Additionally, I found the covariance matrix technique to have a condition number on the order of 10^{-11} while SVD has a condition number of 10^{-14} for the galaxy data given on this problem set. In light of this, I cannot think of a reason to employ SVD when one has access to the square covariance matrix itself; the full eigendecomposition is faster and less prone to numerical instabilities because of a higher (although still rather poor) condition number. Figure 4 shows the first four covariance matrix eigenvectors. Figure 5 plots the eigenvectors generated by the two methods against one another; the sign flips are due to a choice of where to put the negative sign in SVD decomposition.

To finish the PCA, simply project onto the eigenbasis, pick the first N_c components, and reconstruct the approximate signal by summation over this subspace (3). The approximation of the second galaxy's spectrum with the first five principal components is shown in Figure 6; though some of the details are lost as expected, it captures most of the peaks and the general shape of the signal.

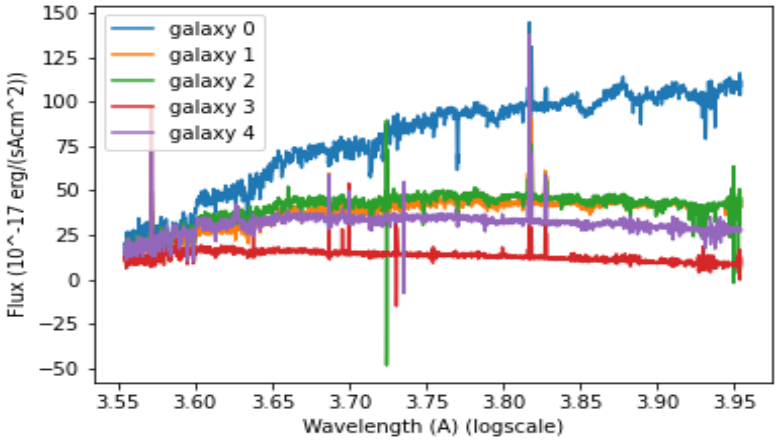

Figure 1: First Five Galaxy Spectra

Figures 7 and 8 show the first and second eigenbasis coefficients and the second and third coefficients respectively plotted against one another. $c0$ and $c2$ seem closely correlated, their plot lying in a nearly flat line (aside from one outlier, not shown in the plot) around $y = 0$. By contrast, c_0 and c_1 are highly variable.

The final step is to measure the residuals as a function of N_c ranging from 1 to 20. Figure 9 shows this result which affirms that the residuals quickly fall off as the number of components used in PCA increases. Due to the mean shifting and normalization, some of the spectra have zeros and the fractional residuals become infinite. It is unclear how to resolve this while preserving the quality of the data, so the absolute residuals are plotted in Figure 8.

Figures

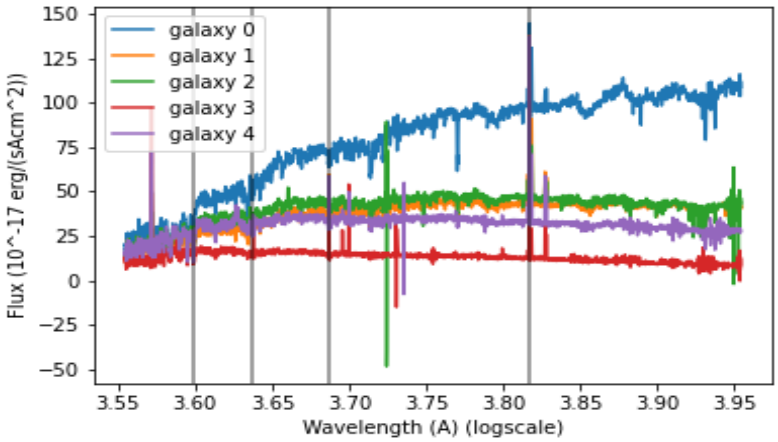

Figure 2: Part (a) with Balmer Spectrum (highlighted in black)

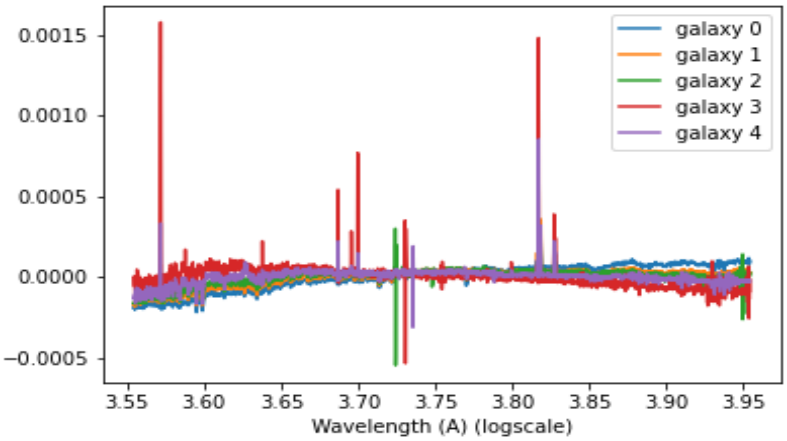

Figure 3: Renormalized Zero-Mean Data

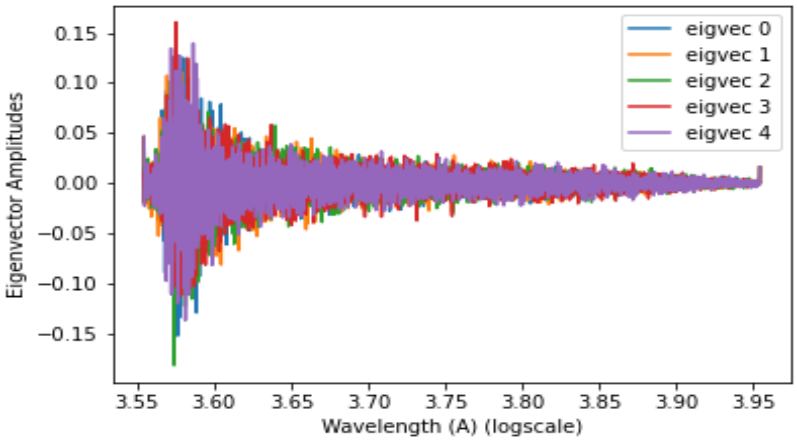

Figure 4: First Five Eigenvectors of the Covariance Matrix

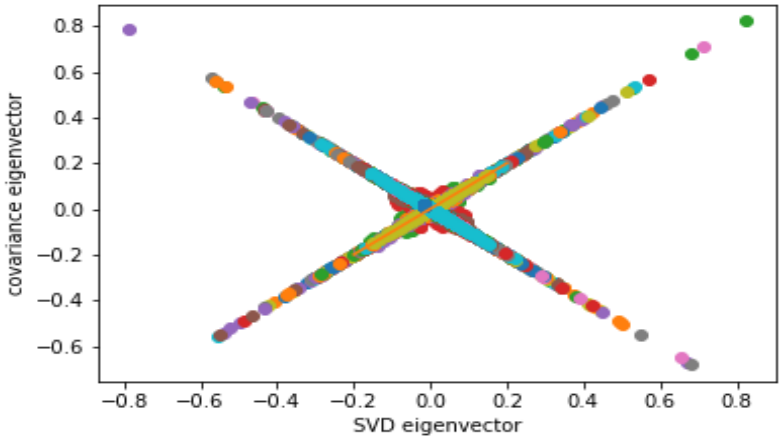

Figure 5: SVD Eigenvectors vs. Covariance Eigenvectors

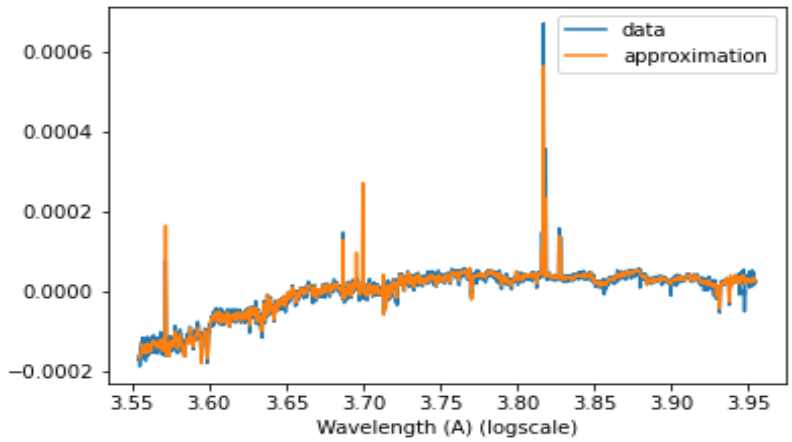

Figure 6: Approximation of the 2nd (index 1) Galaxy Spectrum with 5 PCA Components

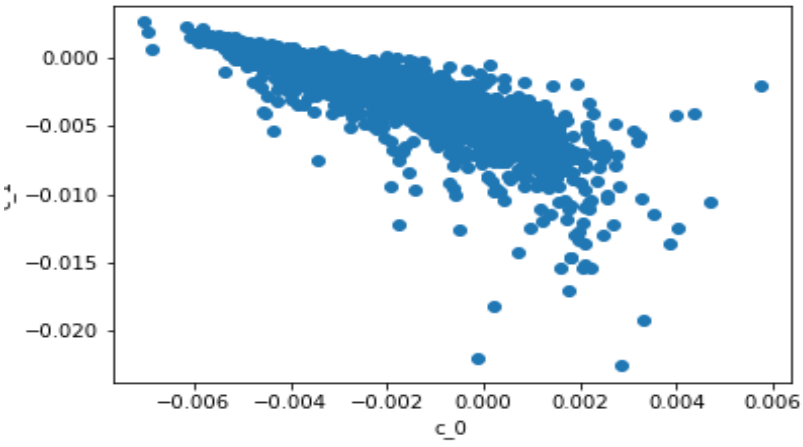

Figure 7: c_0 vs. c_1

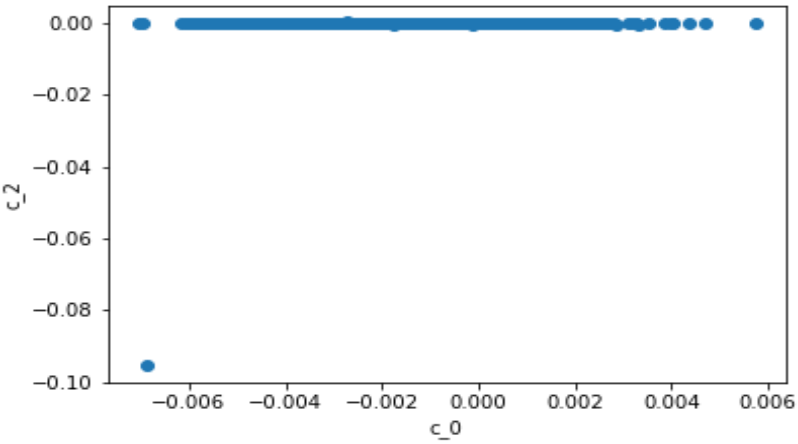

Figure 8: c_0 vs. c_2

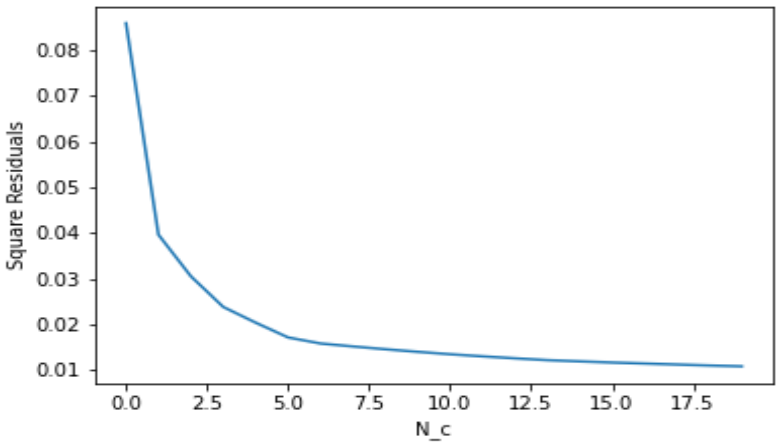

Figure 9: Absolute Squared Residuals

References

[1] Newman, M. 2012, Computational Physics (Createspace Independent Pub)

[2] https://blanton144.github.io/computational-grad/

[3] https://github.com/mcmorre/computational_TA/blob/main/PS_6_hints.ipynb