PHYS-GA 2000 Computational Physics PS5

Kaixuan Zhou

November 1, 2023

1 a

We used the data of the central optical spectra of 9,713 nearby galaxies from the Sloan Digital Sky Survey with FITS(Flexible Image Transport System). Here we plotted the first three spectra.

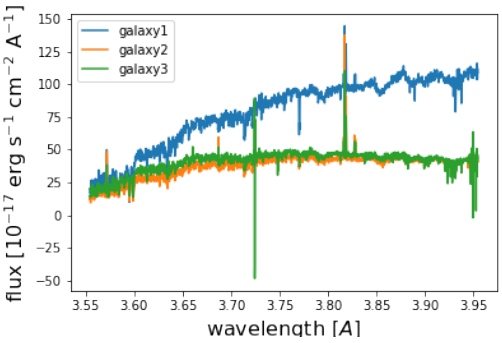

Figure 1: This is the figure for the first three galaxy spectra from the raw data.

These peaks, if we calculate the original wavelength, they are basically around 650nm(we will also see them in following data processing), which is Hydrogen spectrum peaks.

1.1 b

Here we have a big data set so we want to normalize the data. It is close enough to sum over the values to estimate the integral. The following figure shows it is funtioning well.

1.2 c

We subtract off the mean of the normalized data and got the new spectra as the figure 3.

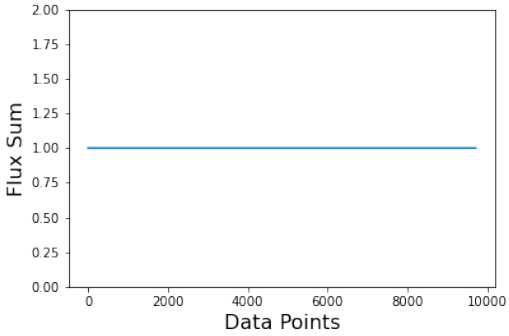

Figure 2: This is the figure for showing whether the normalization is working fine for galaxies.

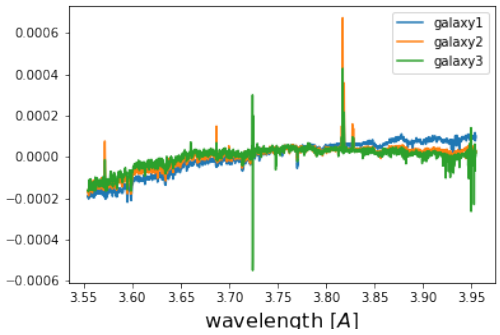

Figure 3: This is the figure of the normalized 0-mean flux. See the y axis is in really low value. That is exactly desired result.

1.3 d

This sectrion we performed the PCA. The idea of the PCA is to find the eigenvectors of the covariance matrix of the distribution. The covariance matrix calculated as:

$$\mathbf{C} = \frac{1}{N_{gal}} \sum_{ij} \vec{r_i} \vec{r_j} \tag{1}$$

Recast the residuals as:

$$\mathbf{C} = \mathbf{R} \cdot \mathbf{R}^T \tag{2}$$

Use this way we construct the matrix and plotted the eigenvectors:
Dimension $\mathbf{C}$ is checked to be $4001 * 4001$.

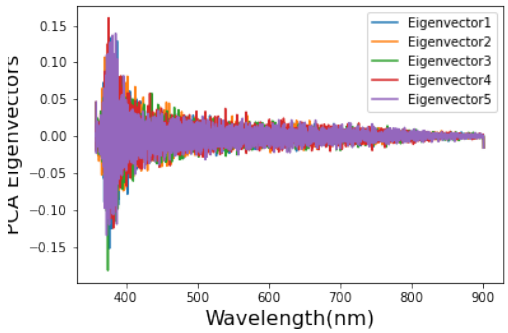

Figure 4: This is the figure of the eigenvectors with found vs wavelength by applying the PCA.

2 e

In this section we performed the SVD, which has been explained in former writing. We are using the similar ways doing it. So we construct the matrix and here is the figure we have for the eigenvector.

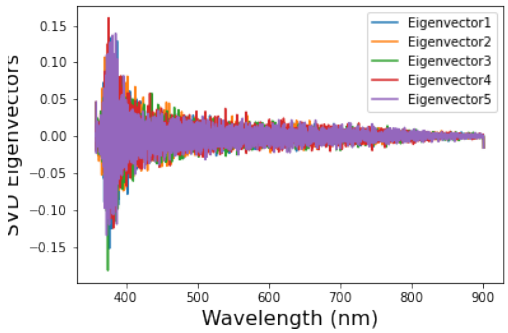

Figure 5: This is the figure of the eigenvectors with found vs wavelength by applying the SVD. See both of them works well.

The computing time for PCA and SVD on my computer are 18.518712043762207 s and 19.78552222251892 s, which is rather close.

To double check the PCA and SVD result we have this figure:

3 f

Since the running time is close, we want to check for the condition numbers.

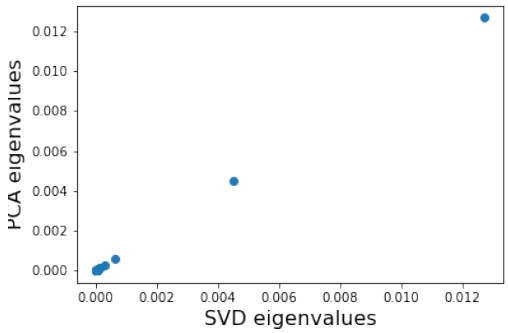

Figure 6: This is a figure for SVD and PCA eigenvalues check. It turned out to be they are highly coincide here.

```
Condition number C: 15052990000.0
Condition number R: 6561841.5
```

Figure 7: This is the conditional number for C and R.

The conditional number for C is obviously much higher than that of R. R should be a more reliable way here.

4 g

We create approximate spectra based on keeping only the first five coefficients in this section, see:

5 h

In this section we plotted c_0 versus c_1 and c_2

Notice the y axis of second figure is much smaller than the first one.

6 i

We noticed that as N_c grows, the squared fractional residuals is decreasing. And the result of $N_c = 20$ is 9.111186e-08, which is a rather small number.

GitHub account:luminousxuan

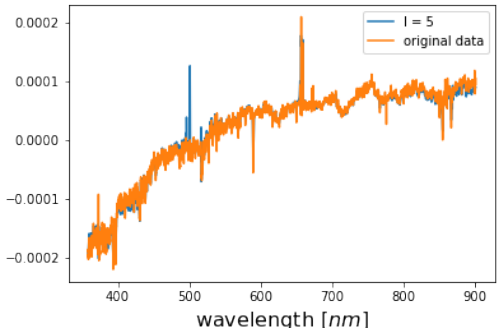

Figure 8: This is the approximated spectra using PCA for the first galaxy and we keep first $N_c = 5$ coefficients.

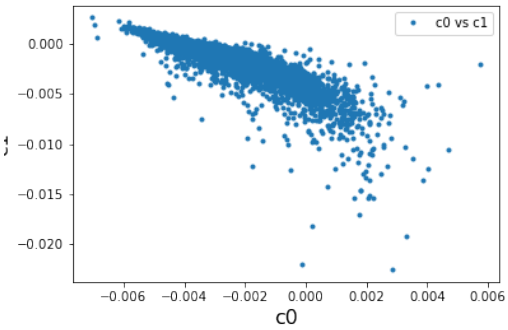

Figure 9: This is the figure for c_0 versus c_1.

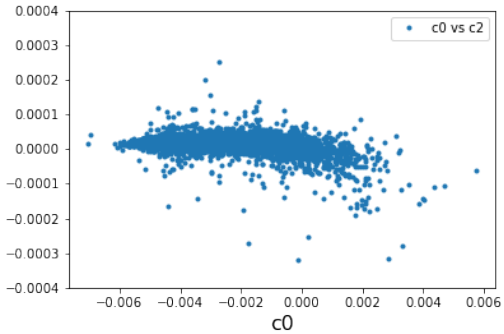

Figure 10: This is the figure for c_0 versus c_2.

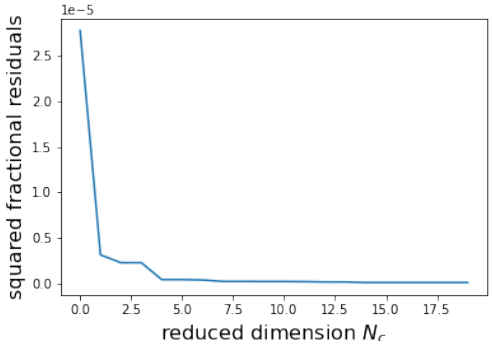

Figure 11: This is the figure of different N_c values instead of 5 we used above