

Liam Kolber Collaborators: Cameron Heppe

Mark Rast

ASTR 3800

14 December 2016

### Measuring Galaxy Redshift

Project four tasked students with analyzing the redshift of galaxies using each galaxy's electromagnetic spectrum. This information was then taken and used to determine an expansion rate of the universe. In this project, however, only a single galaxy is analyzed. Students take the recorded spectrum of this galaxy and then compare it to known spectral lines that aren't redshifted. By using various tools to align the spectrum of the galaxy and the known rest wavelengths, students can then estimate a redshift for the galaxy at hand. As with any scientific analysis, an extensive measure of uncertainties in the data is also analyzed in order to ensure accurate results.

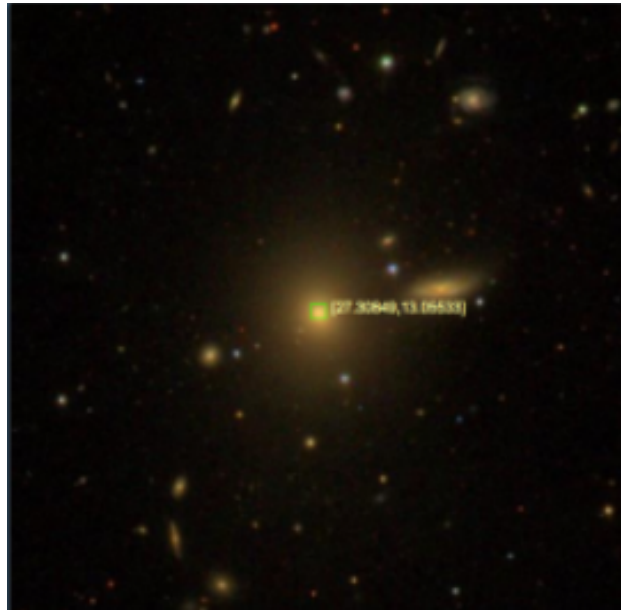


Figure 1: The Observed galaxy in which the spectrum data was received via the Sloan Digital Sky Survey (SDSS)

The data received from the SDSS shows the flux of the galaxy in various wavelengths. Anywhere there are spikes in the data is a good indicator of the existence of a spectral line in that region. Unfortunately this data has an increasing trend to it due to the fact that longer wavelengths are associated with higher fluxes in this data set. In order to perform proper analysis of this set to isolate those potential spectral lines, the data needs to be “detrended” first. This detrending is done using Fourier transforms and removing low frequencies from the data set. After the increasing trend has been removed from the data, the rest wavelengths are plotted on top of the data (Figure 2).

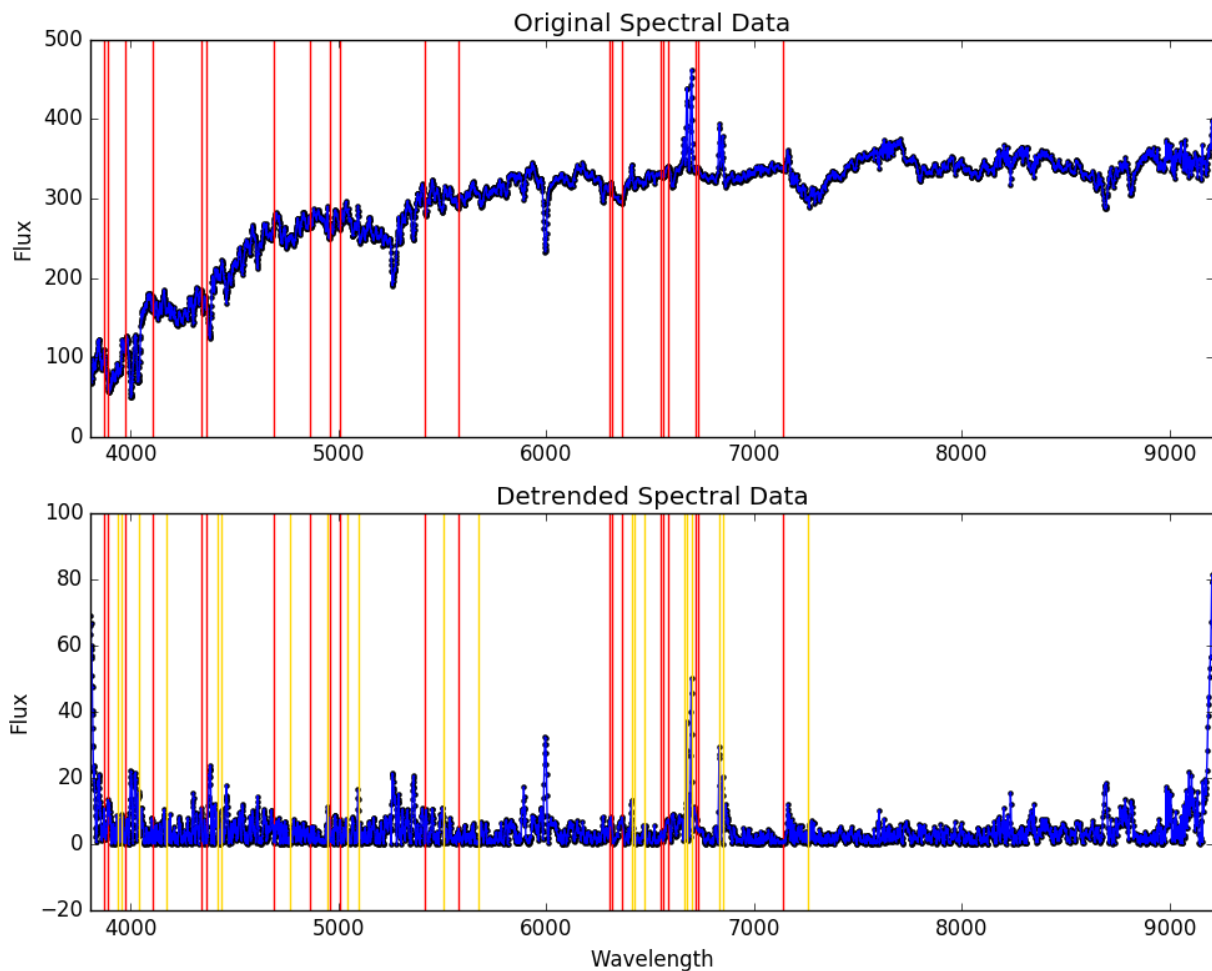


Figure 2: The original and detrended data. The red lines represent the original rest wavelengths and the yellow represents the best shift of the wavelengths estimated via cross correlation.

By performing a cross correlation between the rest wavelengths and the detrended data, it's possible to estimate how much the wavelengths have shifted as the most prominent spikes will line up the best with some of the rest wavelengths. The only part that needs to be considered carefully is that as the shift increases, the larger wavelengths shift more greatly than the shorter wavelengths. This is not difficult to implement though and is accounted for in the estimate. This method produced a redshift of about 0.017024 for the galaxy.

After this method produces a rough estimate of the redshift, each individual rest wavelength is then taken and shifted using said estimate. These locations are used as focus points for a much more accurate peak estimator: a Gaussian fit of the peaks. At each shifted wavelength, a gaussian fit is made of the data within a specified range of  $\pm 10\text{\AA}$  (determined through visual analysis of the most prominent peaks) using scipy's `curve_fit()` function.

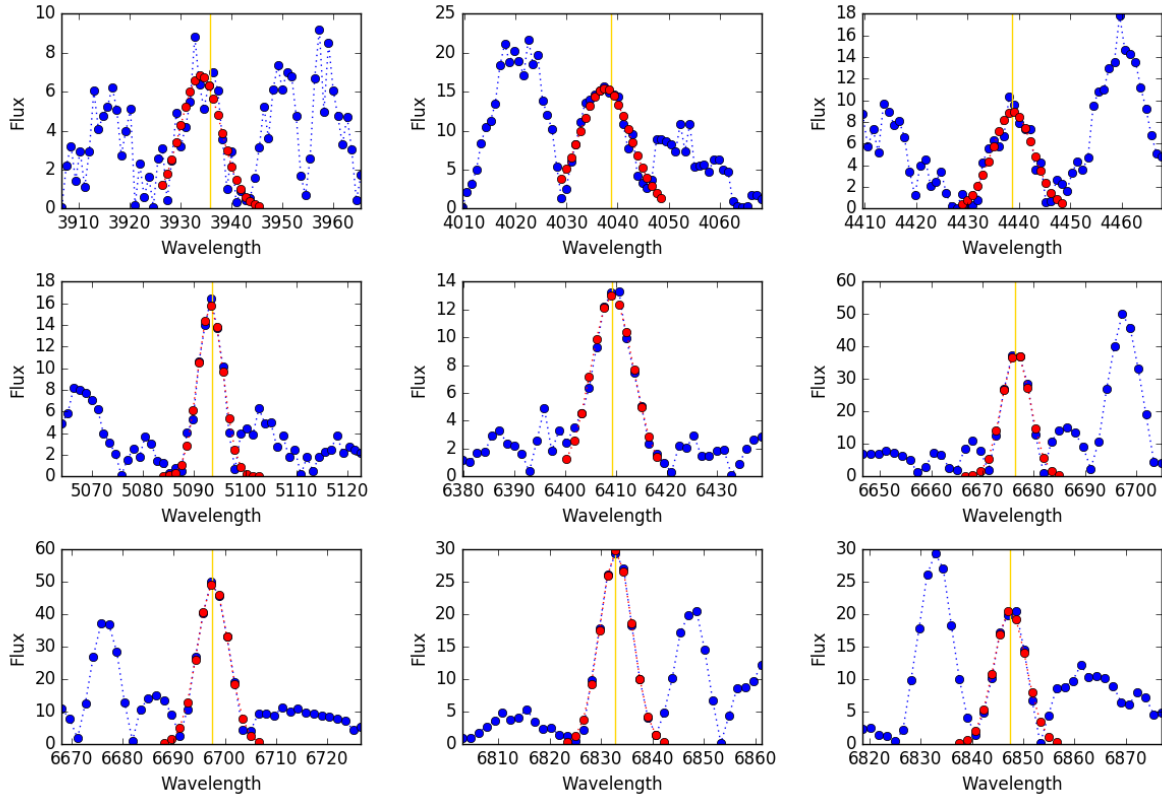


Figure 3: Gaussian fits (red) plotted on top of the data (blue) with the original estimate (yellow)

The location of the peak of the fit is the more accurate measurement of the exact location of the peak than through cross correlation alone, so these new wavelengths are then used to calculate more accurate redshifts of the galaxy along with corresponding uncertainties. In order to ignore any poor fits, any fit that produces a sigma larger than 1% of the its respective redshift is considered to be not useful. Once the individual redshifts and uncertainties are calculated, the average redshift and corresponding uncertainty is calculated via inverse variance weighting. This process is designed so that results with larger variances hold less weight in the ending result, so that they're not ignored, but won't significantly skew the data. This method produced a redshift of  $0.017045 \pm 1.23 \times 10^{-5}$  which puts the original estimate just outside the uncertainty.

A side method of analysis was added to the project concerned with how the uncertainty scales with the number of spectral lines used. In order to perform this, a random wavelength from the estimated redshift array is selected and then ran through the necessary functions to produce a redshift and uncertainty as explained in the previous paragraph (Gaussian fit). The process then continues by selecting another random line and analyzes the two for a redshift. It repeats again for a third random line, fourth, fifth, and so on until each line has been analyzed. This overall process is repeated again but starting with a different random line and continues until the full process has been achieved by starting with each line. The lines are chosen at random so as not to add any biases to the scaling of the data. After each full process is analyzed, it becomes apparent that the uncertainty decreases as more samples are added (Figure 4). This is expected because each iteration reduces the influence of any outliers.

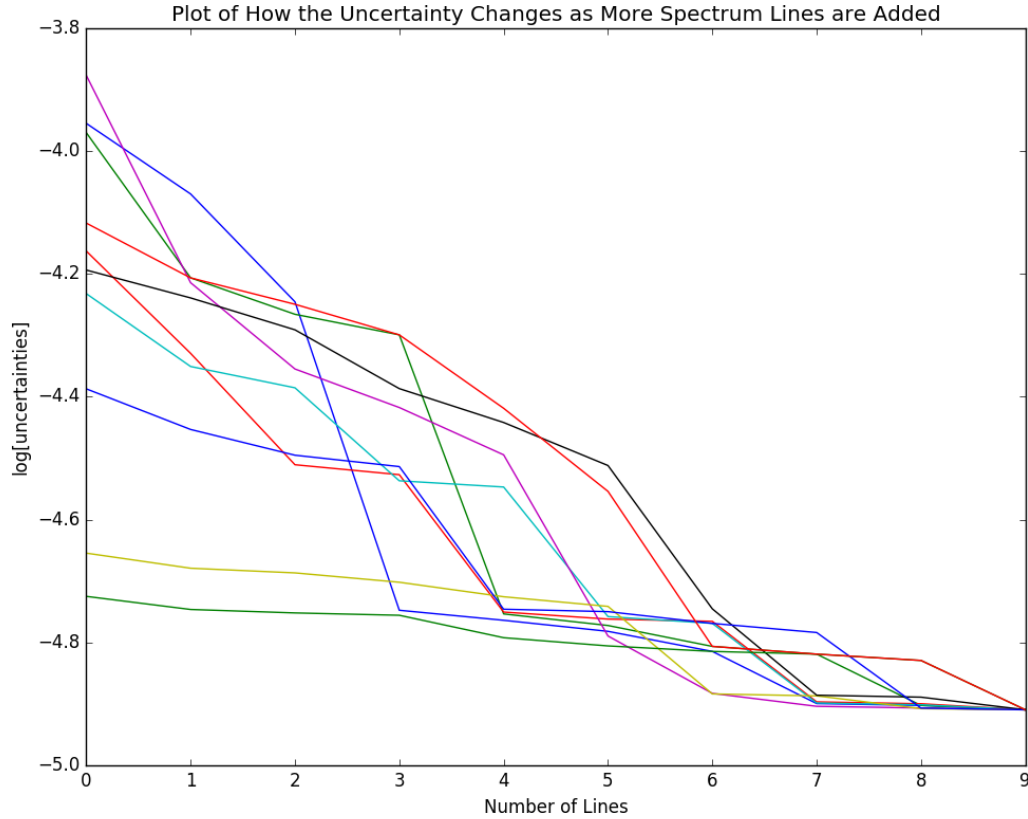


Figure 4: The change in uncertainties as the number of lines increases. Logarithmic scaling used to make differences more prominent.

The last required task of the project questions the existence of any significant systematic difference or trends between the measurements in each individual line vs all the lines as a whole. This is done by performing a  $\chi^2$  test of the redshift calculated from the sets analyzed in Figure 4 with the resultant redshift from the gaussian estimate (Table 1).

The calculated  $\chi^2$  all have probabilities to exceed of about 0.99 which correspond to rather uniform characteristics of each calculated redshift. This means that there is no apparent systematic differences measurements using different lines leading to the conclusion that the methods of calculation used are consistent adding confidence to the results.

In order to reduce uncertainty in these measurements for future attempts at this type of analysis, a key focus would be on reducing the noise of the original data. This would allow the

$\chi^2$	Probability to Exceed
0.140839321779	0.999999031439
0.0884977741039	0.999999845817
0.465344823976	0.999898551104
1.19541165865	0.996687049911
1.72225277068	0.988375341889
0.753767730862	0.99937695923
0.11366861634	0.999999584565
0.191861170574	0.999996731556
1.5090441513	0.992556619912

Table 4:  $\chi^2$  of resultant gaussian redshift vs. the redshifts of each number of spectral lines used.

emission line peaks to be more easily isolated and increasing the accuracy of the fits needed to estimate the shift. This can be done either through the removal of a larger number of frequencies than used in this analysis attempt. Also another method (though outside of the control of the of the students) is to improve upon the instruments used to collect the data so as to reduce noise. Another method in which this analysis could be more accurate is to use a variable range for the gaussian fit. The absence of a variable range made some gaussian fits inaccurate as their estimated centers weren't very accurate adding a bias to the data closer to the original estimated redshift.

### Comment

I very much enjoyed the less “cookie cutter” aspect of this project as it required more thinking and originality on my part. Thanks for an interesting and fun semester.