Hw2

Wenyan Zhou

I tried three measurements to look for the SDSS spectrum closest to cb58, euclidean distance, mahalanobis distance, and correlation coefficient. The algorithm is quite simple. First, normalize cb58. Then, loop through 100 spectra and normalize them, for each spectrum, if its length is greater than or equal to cb58 (because I found an extremely short spectrum that is quite noisy), shift cb58 from its beginning to the end and calculate the three measurements between them. Finally, get the best results. Among the three measurements, correlation coefficient is the one I recommend.

Table1: Top 3 spectra

	Euclidean distance	Mahalanobis distance	Correlation coefficient
Top	5324-55947-0886	5324-55947-0886	5328-55982-0218
3	5328-55982-0218	5047-55833-0214	5324-55947-0886
Spectra	4242-55476-0780	7257-56658-0239	1353-53083-0579

From Table1, we can see that only the correlation coefficient measurement gives the noisy spectrum of cb58, and the other two spectra also interested Christy as they fit well. That is why I choose correlation coefficient as the best measurement.

Here I show the spectrum of cb58, and 3 SDSS spectra closest to cb58 separately in Fig1, Fig2, Fig3, and Fig4.

As to difficulties with data irregularities, I choose to normalize them so that they can be on the same scale, and I also exclude those with nonzero and_mask to filter good observations. However, since the data is continuous, I am not sure whether delete those nonzero and_mask is a good choice (i.e. peaks in cb58 and the candidate spectrum are expected to happen at the same place. If there is a bad observation before a peak, once deleted, it may cause a shift).

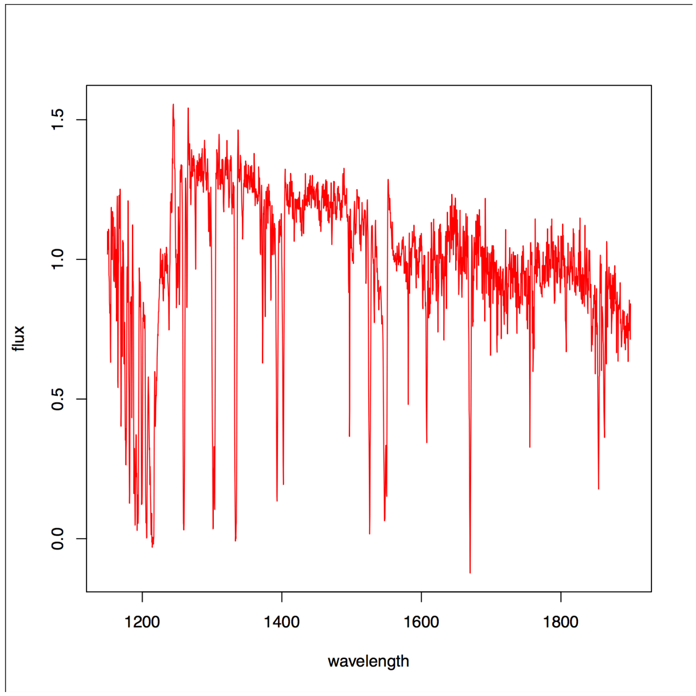

Fig1: cb58

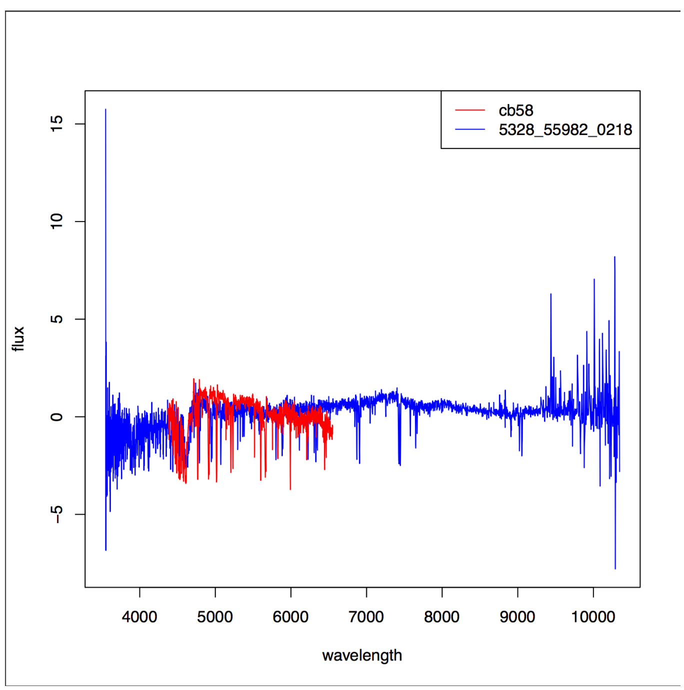

Fig2: cb58 vs. 5328-55982-0218

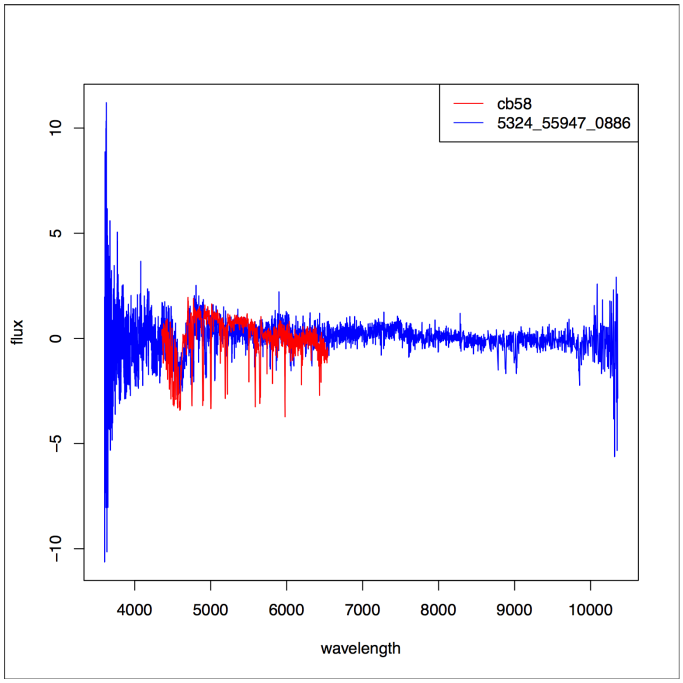

Fig3: cb58 vs. 5324-55947-0886

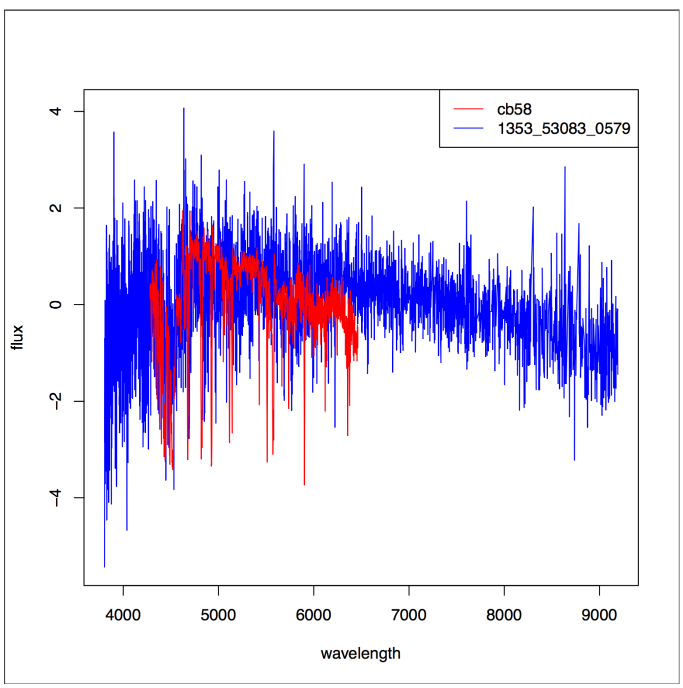

Fig4: cb58 vs. 1353-53083-0579