He II 4686 PROJECT OVERVIEW

Nathan J. Secrest[1]

[1]National Academy of Sciences NRC Research Associate, resident at Naval Research Laboratory

ABSTRACT

Some useful information to get this project started. This will be a living document that will be updated with each new development.

1. INTRODUCTION

The idea behind this project is to use machine learning techniques to distinguish between galaxies with He II 4686 (4686Å) emission due to extreme star formation and galaxies with He II 4686 emission due to possible active galactic nucleus (AGN) activity. The motivation for this is that we are looking for AGN activity in low metallicity galaxies, which are more representative of the earliest galaxies in the universe. This may help us constrain the nature of black holes in the early universe, an outstanding problem in astrophysics. Because metallicity (defined in astrophysics as elements heavier than helium, which are produced in supernovae) is a proxy for the history of star formation in a galaxy, it is also a proxy for the history of gas accretion, either into populations of stars or onto a massive black hole (MBH; mass $M_{\mathrm{BH}} > 100\ M_\odot$, where $M_\odot$ is 1 solar mass). If an MBH has accreted significant gas over the lifetime of the galaxy, its mass and other properties will not be representative of when it first formed. In other words, we're looking for 'fossil' black holes!

The advantage of the He II 4686 line is that, not only is it a proxy for the 'hardness' of the radiation field (and is thus sensitive to AGN activity), it is also metallicity-independent, since helium is one of the primordial elements. So, the idea is to produce a highly reliable and complete sample of galaxies with He II 4686. To this end we will be using the Sloan Digital Sky Survey (SDSS) Legacy survey,[1] which includes a magnitude (brightness, or flux)[2]-limited set of galaxies (the 'Main' sample) observed with the Sloan spectrograph. The advantage to this sample is that it gets everything out to a limiting flux in a homogeneous way (that is to say, using the same instrument), making it not only statistically useful but also a strong candidate for machine-learning methods, which benefit greatly from homogeneous data.

2. METHODOLOGY

The general outline of how we will accomplish our goals is given as follows:

1. Using CasJobs[3], create and account and obtain a manifest of every spectrum taken as part of the Legacy Main survey.

2. Remove anything classified by the pipeline as a "STAR" to reduce data load.

3. Depending on how many spectra that are left, make a local archive of the data. Apparently the best way to do this is by using Globus Online, as discussed in the SDSS documentation.[4]

4. a) With throughput limitations minimized, run a script to find the equivalent width (EW)[5] of each spectrum at the position of the He II 4686 line to pick out everything that *possibly* has the He II 4686 line. *This is ultimately the catalog of He II 4686 we want to create.* The catch is to flag the sources that have the WR bump, as outlined below.

b) Independently, using a training set of galaxies that have the Wolf-Rayet (WR) bump, classify the same galaxies to get a robust list of everything in the Legacy Main sample with WR activity. A good training set might be the sample of WR galaxies from Brinchmann et al. (2008). The ultimate goal is not only to recover everything from that sample that definitely does have the WR bump, but hopefully to improve upon it, so that everything with He II 4686 not flagged as WR using our method is reliable.

The end goal of this work is to produce a reliable and exhaustive (complete) sample of everything in the Legacy Main survey with He II 4686. This by itself will warrant a publication, likely in *The Astrophysical Journal Supplement Series* (ApJS)[6]. With this sample validated and passing peer-review, we can use it to pick out low-metallicity AGN candidates!

For computational resources, I think that the best thing to do would be to keep the database of spectra (and all software) on a (dedicated?) server at GMU. I also like the idea of making a website in which multiple people can help classify galaxies. Once we have (hopefully) shown that ML can classify galaxies with the WR feature, this project will lead to a more generalized classification of Sloan spectra.

REFERENCES

Brinchmann, J., Kunth, D., & Durret, F. 2008, A&A, 485, 657

[1] http://classic.sdss.org/legacy/
[2] https://en.wikipedia.org/wiki/Apparent_magnitude
[3] http://skyserver.sdss.org/casjobs

[4] http://www.sdss.org/dr13/data_access/bulk/
[5] https://en.wikipedia.org/wiki/Equivalent_width
[6] http://iopscience.iop.org/journal/0067-0049