Detection of the DLAs in Quasar Spectra with Machine Learning Algorithms

ML-IAP 2021

Ting Tan, LPNHE, Sorbonne University
ting.tan@lpnhe.in2p3.fr
Collaborators: Christophe Balland, Yuankang Liu, J. Rich, J-M. Le Goff

Lyman-Alpha Forests

The Lyman-alpha forest is a series of absorption lines in the spectra of distant quasars. Those absorptions are caused by the Lyman-alpha electron transitions when photons traverse through the neutral hydrogen regions in the universe.

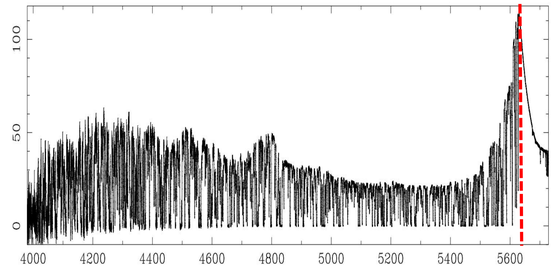

Figure 1. A quasar spectrum as a function of observed wavelength, the red line on the right side shows the Lyman-α peak (121.6 nm in rest frame).

Source: *Womble et al (1996)*

Lyman-alpha forests can be used as biased matter tracers to detect the Baryon Acoustic Oscillations (BAO) signal, which can be seen as a peak in the two-point correlation function. DLAs play an important role in the Lyman-alpha BAO analysis, in which a perfect DLAs catalog is needed.

Damped Lyman-α Systems

DLAs are strong absorption regions in Lyman-alpha forests caused by neutral hydrogen with extreme high column densities, usually log(NHI) ≥ 20. Figure 2 shows a voigt profile fitting for the modeling of a DLA.

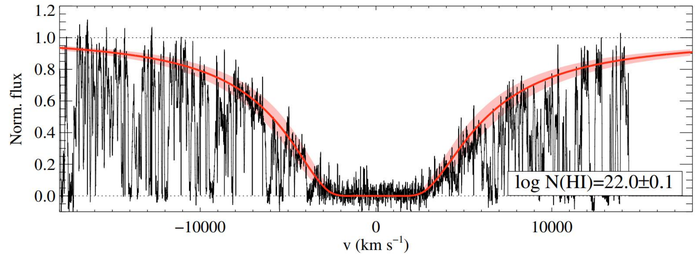

Figure 2. Voigt profile fitting for a DLA with column density log N(HI) =22±0.1 in velocity space.

Source: *Noterdaeme et al. 2009*

DLA finder

Different machine learning algorithms have been used to estblish the DLAs catalogs in SDSS, such as voigt profile fitting, CNN, and Gaussian processes. However, these approaches are milted: they have insufficient accuracy for DLAs with low flux and low column densities, and there exists significant disagreements between their finding catalogs.

CNN / GP	0 DLA	1 DLA	2 DLAs	3 DLAs
0 DLA	142759	5686	93	2
1 DLA	2397	8007	208	1
2 DLAs	117	234	333	5
3 DLAs	8	6	11	4

Figure 3. disagreements between Gaussian Processes and CNN for SDSS DR16 data.

Source: *M.-F. Ho et al. (2021)*

In our study, we use several CNNs to do DLA classification, redshift and column density estimations seperately. Each input quasar spectrum is cut into pieces by a sliding window, and the classification is made by an ensemble decision. In order to estimate the statistical uncertainties of the CNN output, we are adapting Bayesian CNNs to estimate the redshifts and NHI.

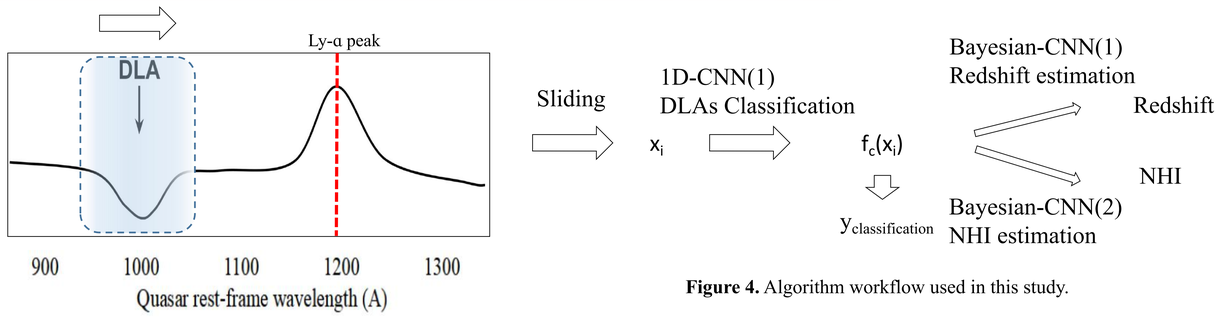

Figure 4. Algorithm workflow used in this study.

In this case, the weights transfered between layers are no longer numbers, but estimated distributions. As a result, the outputs are also estimated distributions including variances.

Results

The classification results based on the Saclay Lyman-α mocks (Etourneau et al. 2021) are shown in Figure 5. The results show encouraging performance on 6 data samples, depending on Signal/noise and NHI. The overall purity and Completeness for all DLAs with log(NHI)>20.3 are also above 90%.

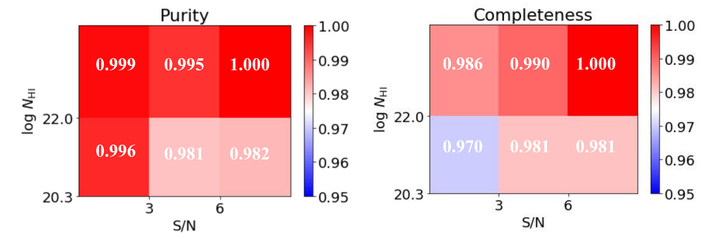

Figure 5. Purity and Completeness of our model based on Saclay Lyman-α mocks. 3 signal/noise bins and 2 NHI bins are applied for the test. The results show an overall 97% performance on these 6 data samples.