SciDBoss Documentation

1. Quasar sample

I used a testing sample of 5000 quasars from the DR12 catalog (`http://www.sdss.org/dr12/algorithms/boss-dr12-quasar-catalog/`). These quasars were required to have $2.15 < z < 5$ (using Z_VI) and $BAL\_FLAG\_VI = 0$; that is, they were required to have no visually apparent broad absorption line troughs. From this set I randomly selected 5000 quasars and downloaded the spec files for each of these quasars, from `http://data.sdss3.org/sas/dr12/boss/spectro/redux/%d/spectra/%04d/spec-%04d-%05d-%04d.fits%(rerun,plate,plate,mjd)` (see `https://www.sdss3.org/dr9/spectro/pipeline.php` for a description, under "per object spec files"). For each quasar I used the coadded data combining each exposure. These data are stored in the fits file *boss_test_sample.fits* and in the directory *boss_test_sample*. The function loading all the data is *load_all_data* within *lya_functions.py*. I also mask the data by rejecting any pixel with the $AND\_MASK > 0$ (see `http://www.sdss.org/dr12/algorithms/bitmasks/#SPPIXMASK` for a description of the pixel maskbits; the AND_MASK means that the maskbit was set for any of the exposures within the coadd, not necessarily all). Since the SDSS sky subtraction is often inadequate around bright night sky lines, I additionally mask 6 Å on either side of the night sky lines at 5577 Å, 6300 Å, 6363 Å, and the ISM sodium D line at 5890 Å. The function doing this masking is *mask_data* in *lya_functions.py*.

2. Finding <F(z)>

As a first step, I fit a model of the following form to each quasar spectrum:

$$f(\lambda) = A_{qso}\bar{C}(\lambda_r) < F(z_{abs}) > \tag{1}$$

where $\bar{C}(\lambda_r)$ is the mean quasar continuum, A_{qso} is the amplitude of each individual quasar, and $< F(z_{abs}) >$ is the mean transmitted flux fraction through the Lyα forest. This model assumes that all quasars have the same shape and does not consider fluctuations from the mean transmission.

Rather than fitting this model in one go I proceed iteratively. In the first iteration I take the mean continuum to be 1:

$$\bar{C}_0(\lambda_r) = 1 \tag{2}$$

The mean continuum is defined over a specified grid; for now, the grid starts at 600 Å, ends at 3600 Å, and has spacing of 4 Å between points. I use the Faucher-Giguere 2008 model as a rough estimate for $< F(z) >$ (the analytic fit cited in the Font-Ribera DLA paper):

$$< F(z)_0 >= \exp\left[-0.0018(1+z)^{3.92}\right] \tag{3}$$

$< F(z) >$ is defined in a grid where each point has $\Delta z = 0.03$, and spans a total width of 1.5 in redshift. The beginning of $< F(z) >$ is the lowest measurable absorption redshift: the blue end SDSS spectroscopic limit, 3566.97 Å, divided by Lyα, 1215.24 Å, minus 1.

Last, I find the initial quasar amplitude by finding the weighted mean of the quasar flux between 1275 and 1285 Å.

$$A_{qso,0} = \frac{\sum f_i w_i}{\sum w_i} \tag{4}$$

w_i is the weight of each pixel, calculated from its total inverse variance:

$$w_i = \left(\sigma^2_{pipe} + f_i^2 \sigma_i^2\right)^{-1} \tag{5}$$

where σ^2_{pipe} is the variance outputted by the SDSS pipeline (from the spec file) and σ_i^2 is the intrinsic variance, a step function

$$\sigma_i^2 = \begin{cases} 0.1 & \text{if } \lambda_r < 1215.24 \\ 0.01 & \text{if } \lambda_r > 1215.24 \end{cases} \tag{6}$$

Note that σ_i^2 is dimensionless. We need σ_i^2 for two reasons. First, there is intrinsic variance associated with the transmission fraction in the Lymanα forest (due to cosmic structure) so even a high S/N quasar should have some floor to its variance in the Lymanα forest region. Second, we need an intrinsic variance to prevent high S/N quasars from being weighted too heavily; if we had no intrinsic variance, then the mean continuum would be heavily biased towards high S/N quasars. Note also the units of the weight: since the weights are inverse-variance weights, their units are flux^{-2}.

Now I begin to iterate. First I compute $\bar{C}$:

$$\bar{C}(\lambda_r)_i = \frac{\sum_j w_j \frac{f_j}{A_{qso,i-1} \bar{F}(z)_{i-1,j}}}{\sum_j w_j} \tag{7}$$

The weights are equal to the total inverse variance of the quantity being averaged

$$w = \left(\frac{\sigma^2_{pipe}}{A^2_{qso,i-1} < F(z) >^2_{i-1}} + \bar{C}_{i-1} \sigma_i^2 \right)^{-1} \tag{8}$$

Note that I use the $(i-1)$th estimates of A_{qso}, $< F(z) >$ and $\bar{C}$. For any pixel with $z_{abs} > z_{max} + \Delta z/2$, I have no estimate of $< F(z) >_{i-1}$, so I use the Faucher-Giguere estimate at these redshifts for all iterations.

Next I compute A_{qso}. Here I use a chisq minimization that is formally identical to the weighted average but is easier to implement if I switch to using a multi-parameter model for the continuum. Specifically I only fit to the region red of Lyα, and I use weighted least

squares to properly account for the different variances between different pixels; the weights are given by:

$$w = \sqrt{\left(\frac{\sigma_{pipe}^2}{< F(z) >_{i-1}^2} + A_{qso,i-1}^2 \bar{C}_i^2 \sigma_i^2 \right)^{-1}} \tag{9}$$

in other words the inverse of the standard deviation of each pixel, in units of flux^{-1}. Note also that $< F(z) >_{i-1}$ will always be equal to 1 since we only fit red of Lyα. Written in matrix form we then solve

$$W\vec{T}A_{qso} = W\vec{f} \tag{10}$$

where W is the diagonal matrix of the weights, $\vec{T}$ is the template ($\bar{C}(\lambda_r)_i$ re-binned to the quasar wavelengths), $\vec{f}$ is the measured fluxes, and A_{qso} is the scalar amplitude that we're fitting for. This is a least squares problem of the form $\vec{a}X = \vec{b}$ so it can be solved quickly by already-existing libraries to find A_{qso}.

Now I find $< F(z) >$. I consider only the region between 1041 Å and 1185 Årest frame. I assign each pixel to the nearest z_{abs} and calculate $< F(z) >$ in the following way:

$$< F(z) >_i = \frac{\sum_j w_j \frac{f_j}{A_{qso,i}\bar{C}_i}}{\sum w_j} \tag{11}$$

where now w_j is given by

$$w_j = \left(\frac{\sigma_{pipe}^2}{(A_{qso,i}\bar{C}_i)^2} + < F(z) >_{i-1}^2 \sigma_i^2 \right)^{-1} \tag{12}$$

I also calculate the error on each measurement of $< F(z) >$:

$$\sigma_F^2 = \frac{1}{\sum w_i} \tag{13}$$

This follows from the variance of a weighted mean (see appendix). Also, $< F(z) >$ is fixed in order to break the degeneracy between $< F(z) >$ and $\bar{C}$ in the Lyα forest region. I fix the mean of $< F(z) >$ between $z = 2.2$ and $z = 2.6$ to equal the mean of Faucher-Giguere's measurement of $< F(z) >$ between 2.2 and 2.6.

To measure convergence, I calculate a "chisq" statistic:

$$\chi^2 = \sum \left(\frac{< F(z) >_i - < F(z) >_{i-1}}{\sigma_{Fi}} \right)^2 \tag{14}$$

where the sum runs over the redshift bins, and i refers to the ith iteration. Convergence is reached when this statistic is less than the number of redshift bins.

3. Appendix: variance of a weighted mean

Consider the estimation of the mean and standard error for a set of data x_i with weights w_i. These weights can be interpreted as non-integer "counts" of each data point. For instance, a weight of 1.3 means that a particular datapoint is counted 1.3 times.

Take the probability of obtaining a particular measurement x_i. Assume this distribution is Gaussian:

$$P_i(\mu') \propto \exp\,[-\frac{1}{2}(\frac{x_i - \mu'}{\sigma})^2] \tag{15}$$

The probability of observing the entire data set of N observations is given by

$$P(\mu') = \prod^{N} P_i(\mu')^{w_i} \tag{16}$$

We see that the interpretation of weights as effective counts means that we must raise each probability to the w_i power. We thus have

$$P(\mu') \propto \exp\,[-\frac{1}{2}\sum w_i(\frac{x_i - \mu'}{\sigma})^2] \tag{17}$$

According to the method of maximum likelihood, the most probable value for the mean is the value that maximizes the probability in equation 17. This is equivalent to minimizing the argument of the exponential:

$$X = -\frac{1}{2}\sum w_i(\frac{x_i - \mu'}{\sigma})^2 \tag{18}$$

The minimization condition is

$$\frac{\partial X}{\partial \mu'} = 0 \tag{19}$$

yielding (after algebraic simplification)

$$\sum w_i(x_i - \mu') = 0 \tag{20}$$

or

$$\mu' = \frac{\sum w_i x_i}{\sum w_i} \tag{21}$$

This checks with our intuition.

The standard error is given by

$$\sigma_\mu^2 = \sum [\sigma_i^2(\frac{\partial \mu}{\partial x_i})^2] \tag{22}$$

using error propagation. We have

$$\frac{\partial \mu}{\partial x_i} = \frac{\partial}{\partial x_i}(\frac{\sum w_i x_i}{\sum w_i}) = \frac{w_i}{\sum w_i} \tag{23}$$

Thus

$$\sigma_\mu^2 = \sum [\sigma_i^2 (\frac{w_i}{\sum w_i})^2] \tag{24}$$

In our case the weights are equal to the inverse of the variance. Therefore

$$\sigma_\mu^2 = \left(\frac{1}{\sum w_i}\right)^2 \sum \frac{w_i^2}{w_i} = \frac{1}{\sum w_i} \tag{25}$$