Quasar continuum component analysis of the WEAVE-QSO Survey

Antonio Navas Orozco

Tutor:

José Oñorbe Bernis

Universidad de Sevilla
Doble Grado en Física y Matemáticas
June 24, 2022

Contents

Resumen

Los cuásares son agujeros negros supermasivos rodeados de material muy caliente que, mientras cae en espiral, emite una gran cantidad de radiación electromagnética. Se encuentran en el centro de galaxias muy lejanas, y de las absorciones en sus espectros, se puede extraer información sobre la densidad, temperatura y composición del medio intergaláctico. Para ello, el continuo de estos espectros ha de ser estimado. Presento dos métodos para reconstruir el continuo de cuásares: el cálculo de la pendiente de la región UV de los espectros de cuásares y interpolación con splines cúbicos. Con el objetivo final de tener la técnica de análisis de datos lista cuando comience el catálogo de WEAVE-QSO a finales de 2022, uso el set de datos sintéticos WEAVE-QSO OpR3b para testear la aplicabilidad de estos métodos reconstruyendo el continuo de los de datos observados sintéticos y comparándolo con el continuo teórico incluido en el set de datos. Encontré la región más apropiada para el cálculo de las pendientes para este set de datos. La pendiente no mostró dependencia ni en la magnitud aparente en banda r ni en el redshift. Adicionalmente, encontré una cierta parcialidad (o 'bias') entre las pendientes observadas y teóricas, y para los cuásares con $S/N > 3$ propongo un método para corregirla en aplicaciones futuras. En cuanto a la interpolación por splines cúbicos, estudié los parámetros óptimos del algoritmo seguido, en primer lugar visualmente para un número reducido de cuásares y en segundo lugar minimizando ciertos residuos en varias situaciones incluyendo distintas regiones (todo el espectro, a la izquierda y a la derecha de Lyman-α) y posibles constantes de renormalización. Estos parámetros óptimos no dependieron del redshift, pero sí de la magnitud aparente en banda r. Usé la dependencia con la magnitud para proponer un algoritmo que lleva la magnitud aparente en banda r de un cuásar en los parámetros optimos para la interpolación por splines cúbicos. Por último, comprobé los resultados de este algoritmo haciendo un stack encontrando errores del 16 %, y solo del 2.2 % para $S/N > 3$.

Abstract

Quasars are supermassive black holes surrounded by very hot material that, as it falls spirally, emits huge amounts of electromagnetic radiation. They are located in the center of very distant galaxies, and from the absorption present in their spectra, information about the density, temperature and composition of intergalactic medium can be extracted. For that, the continuum from these spectra must be estimated. I present two methods to reconstruct quasar continua: computation of the slope of the UV region of QSO spectra and cubic spline interpolation. With the final objective of having the analysis pipeline ready when the WEAVE-QSO Survey starts in late 2022, I take advantage of the WEAVE-QSO OpR3b mock dataset to test the extent of the applicability of these methods by reconstructing the continuum with the mock observed data and comparing it with the theoretical continuum included in the data set. I found the most appropriate the region for the computation of the slopes in this dataset. The slope of mock and theoretical spectra did not depend on neither r-band apparent magnitude nor redshift. Additionally, I found certain bias between observed and theoretical slopes, and for $S/N > 3$ quasars I propose a method to correct it for future applications. Regarding the cubic spline interpolation, I studied the optimal parameters of the algorithm followed, first visually for a reduced number of quasars and secondly minimizing certain residuals in a number of situations including different regions (whole spectrum, bluewards and redwards of Lyman-α) and possible renormalization constants. These optimal parameters did not depend on redshift, but did depend on r-band apparent magnitude. I used the dependence on magnitude to propose an algorithm mapping QSO r-band apparent magnitude to optimal parameters for the cubic spline interpolation. Lastly, I checked the results of this algorithm by stacking finding errors of 16 %, and only 2.2 % for $S/N > 3$.

1. Introduction

The name 'quasar' is an abbreviation of 'quasi-stellar', for they were referred to as 'quasi-stellar radio sources' in the past (similarly to stars, they are point sources, therefore they are star-like or 'quasi-stellar' celestial objects). Nowadays, both the term quasar and quasi-stellar object (or QSO, for short) are used interchangeably. They are the most powerful and widely studied type of AGN or active galactic nuclei (see Kembhavi and Narlikar, 1999).

A quasar consists, basically, on a supermassive black hole located in the center of a galaxy, surrounded by very hot material that emits huge amounts of 'light' (electromagnetic radiation). This allow us to observe them through telescopes in spite of the fact that they are very far away. The light they emit, as it travels towards the Earth, may suffer some absorption from the matter in the line between the quasar and us. As the universe is expanding, the wavelengths of this light travelling towards us stretch, and therefore so does the spectrum or flux against wavelength (see Condon and Matthews, 2018). As this occurs gradually, an absorption at a certain wavelength produces different changes in the spectra depending on the moment it occurred. Once the light reaches us, from these different absorptions, information about the density, temperature, composition and more about the clouds of gas that produced the absorptions can be obtained. In particular, with enough absorption features the density field can be reconstructed. However, in order to do so, the emitted, absorption-less spectra must be estimated, which is not a straightforward task (see Dall'Aglio et al., 2008 as an example of continuum reconstruction techniques). Different methods for this estimation will be presented and analyzed using a mock data set with the objective of testing their applicability in the WEAVE-QSO Survey, which is intended to start by September 2022.

I structured this work as follows. In section 2, I establish the theoretical framework strictly necessary to understand this work. In particular, I define magnitude and redshift, and introduce the typical QSO spectrum in the region of interest (basically UV and visible wavelengths). In section 3, I explain the procedure to generate mock data and, in particular, the characteristics of the OpR3b data I use in the following sections. In section 4, I explain the two methods I will be testing on the data set, namely the computation of the slope of the UV spectra, and the cubic spline interpolation. In section 5, I present the results obtained after applying them. In particular, for the slopes in the UV, a study of the slope distribution is performed, and whether obtaining this slope from real data is possible or not is discussed; and for the spline interpolation, a study of the distribution of the optimal parameters of these splines is included, and the method proposed is tested by stacking a particular group of quasars. Lastly, in section 6, I show the conclusions I could reach, along with some ideas for the further development of these methods.

2. Theoretical framework

In this section, I will introduce 'magnitudes' as the usual convention for the measuring of luminosities and fluxes, I will talk about redshift and why the main redshift component for QSO is the cosmological one, I will define quasars both observationally and theoretically via the unification theory for AGN, and I will explain what the typical QSO spectrum looks like, in particular in the UV region (the region of the data I will analyse in sections 4 and 5). For general astrophysical concepts, I recommend consulting Geller et al. (2019) and Carrol and Ostile (2017), and for quasar-specific concepts, Kembhavi and Narlikar (1999).

2.1. Magnitude

Let me begin by saying that the word 'flux' is used for different physical properties in astrophysics depending on the criteria followed by the writer. I will use it to refer to the monochromatic flux[1], F_λ, measured in $[\mathrm{J}/(\mathrm{sm}^2\text{Å})]$[2] (that is, energy per unit time, per unit surface and per unit wavelength). As I will only use F_λ and not F_ν ($|F_\lambda d\lambda| = |F_\nu d\nu|$), I will get rid of the sub-index from now on.

Nevertheless, for historical reasons, it is very common to use magnitudes instead of fluxes. In general, one may define the difference between apparent magnitudes m of two objects '1' and '2',

$$m_1 - m_2 = -2.5 \log \frac{F_1}{F_2} \tag{1}$$

Note that, analogously to potential energy, it is defined save an additive constant. It is common to set it to 0 for the star Vega, for example. Also note how more luminous objects have lower magnitude, and that $F_1 = 10\ F_2 \implies m_1 = m_2 - 2.5$. Absolute magnitude M, on the other hand, is defined as the apparent magnitude the object would have if it were at a distance of 10 parsec (becoming, thus, independent of the distance). These two equations can be easily derived:

$$M_1 - M_2 = -2.5 \log \frac{L_1}{L_2} \tag{2}$$

$$m - M = 5(\log d - 1) \tag{3}$$

Here, d is the distance, L is the luminosity ($L = 4\pi d^2 F$, constant in vacuum), and '$m - M$' is called the distance modulus (they obviously refer to the same object). Moreover, back to equation (1), it defines apparent magnitude in a specific wavelength. In the real

[1] In the literature, it can sometimes be seen as spectral flux density, using flux without adjectives to refer to luminosity.

[2] Dividing by $\hbar\nu = \hbar c/\lambda$, it becomes the photon density.

world, usually a continuous range of wavelengths is measured, and it is also necessary to take into account the filter used. Filters are accesories for telescopes or cameras used to improve the observation of celestial objects and the photon counting in a specific range of wavelengths. Their wavelength range is called the band-pass. In particular, taking into account the transmission function T_X of the band-pass filter used, 'X',

$$m_1 - m_2 = -2.5 \log \frac{\int F_1(\lambda) T_X(\lambda) d\lambda}{\int F_2(\lambda) T_X(\lambda) d\lambda} \tag{4}$$

In my case, I will use the AB magnitude system, defining the zero point as follows.

$$m = -2.5 \log \frac{\int F(\lambda) T_X(\lambda) d\lambda}{\int \frac{c}{\lambda^2} 3631 \ Jy T_X(\lambda) d\lambda} \tag{5}$$

Here, Jy stands for janskys, defined as 1 Jy$= 10^{-26}$ W m^2 Hz^{-1}. The c/λ^2 factor is due to the definition being originally for frecuencies, ν, not for wavelengths $\lambda = c/\nu$.

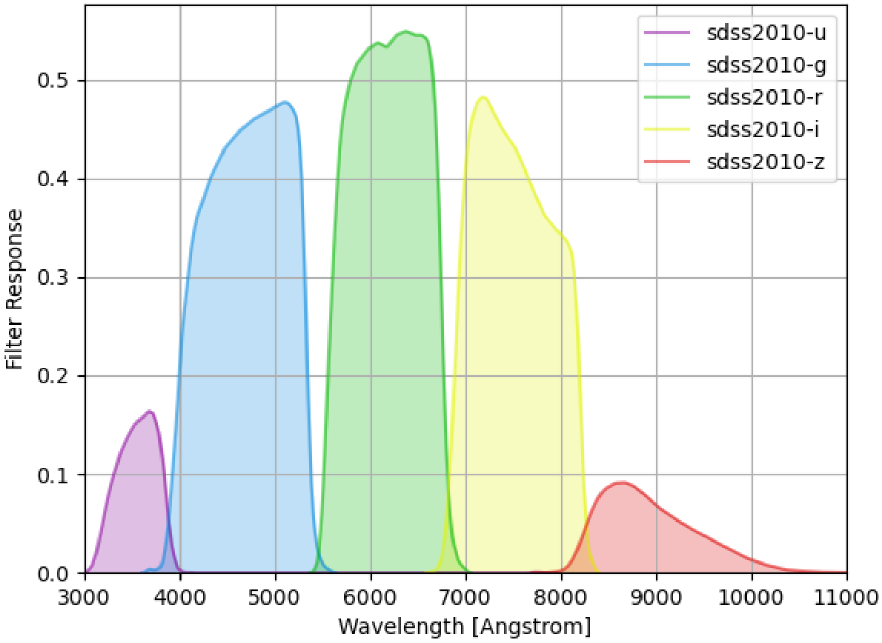


Figure 1: Filter response function for different filters in the 'ugrid' system. Through this work, only the r-band will be used. SDSS stands for Sloan Digital Sky Survey, see York et al. (2000).

Ideally, the filter transmission function, or filter response, would be rectangular-shaped, to measure only an interval of wavelengths, but uniformly well (no individual wavelength should be more important than another). However, real-life filters are not perfect, and have smooth limits, as is represented in figure 1. Note that the area below the curves is not 1, but it can be observed from equation (5) that the normalization of these curves is not very important, as it simply cancels out in the quotient. The letters that represent the five pass-bands stand for ‘ultraviolet’, ‘green’, ‘red’, ‘near infrared’ and ‘infrared’, in order. In particular, in this work I will use the apparent AB magnitude in the r-band only, which is part of the information the data I will work with provides.

2.2. Redshift

The parameter z, called ‘redshift’, represents the spectral shift from the wavelength emitted, λ_{em}, and the wavelength measured in another part of the universe, λ_{obs}, and its expression is:

$$z = \frac{\Delta\lambda}{\lambda_{em}} = \frac{\lambda_{obs}}{\lambda_{em}} - 1 \tag{6}$$

That is, $1 + z = \frac{\lambda_{obs}}{\lambda_{em}}$. As QSO have strong emission lines, corresponding to electronic transitions between states with tabulated energy, they can ultimately be identified, and thus redshift can easily be determined.

Now, the natural question is what the nature of this redshift actually is. There are three candidates: Doppler shift, which is due to the relative movement between the emitter and the observer; gravitational redshift, which is due to gravity, which in this case would be that of the emitter; and cosmological redshift, due to the expansion of the universe. I am going to argue why the latest is the only non-negligible one for quasars.

The first candidate is the Doppler shift. If a source of light moves with speed $\vec{v}$ away from the observer in their rest frame, the spectral shift of the light measured by the observer will be

$$1 + z = \frac{1 + (v/c)\cos(\theta)}{\sqrt{1 - (v/c)^2}} \tag{7}$$

θ is the angle made by $\vec{v}$ with respect to the radial vector between source and observer. This equation is relativistic, but in astrophysical terms, the Newtonian limit ($v \ll c$) may be more useful. Via Taylor expansion, its easy to prove that $z = v_r/c$, where $v_r = v\cos(\theta)$ is the (positive or negative) radial component of the relative velocity.

This Doppler shift has long been confirmed in stellar motion within our galaxy. Note that Doppler shift delinks redshift from distance: large redshifts do not imply large distances from us. One of the first theories suggested to explain quasars assuming Doppler shift as important stated that quasars were celestial objects ejected from violently active galactic centers. However, Kembhavi and Narlikar (1999) argue that this component of the redshift is negligible for quasars: the amount of energy required for such phenomena is problematic, and there is also the problem of blueshift: it can be proven that, under the asumption of isotropic ejection, the number of blueshifted quasars (those ‘heading towards us’, $v_r < 0$) would be around 81 times the number of redshifted quasars (moving away from us, $v_r > 0$). Observations discard this, serving as proof that Doppler shift is not the main component of the QSO redshift measured.

The second candidate is the gravitational redshift. Assuming spherical symmetry, from Schwarzschild's solution to Einstein field equations the following expression can be derived:

$$1 + z = \sqrt{\frac{1 - \frac{r_S}{R_{obs}}}{1 - \frac{r_S}{R_{em}}}} \tag{8}$$

Here, R_{obs} and R_{em} are the radius where the observer is and where the photon is emitted, respectively. The Schwarzschild radius, $r_S = \frac{2GM}{c^2}$ is the event horizon radius of a Schwarzschild black hole with mass M. In astronomical terms, Earth is at infinity, that is, $R_{obs} \gg r_S$, and equation (8) simply becomes

$$1 + z = \left(1 - \frac{r_S}{R_{em}}\right)^{-1/2} \tag{9}$$

It may seem possible, then, to obtain large redshifts by approaching R_{em} to r_S. However, as H. Bondi showed in 1964 (Kembhavi and Narlikar, 1999), any realistic equation of state (relation between the pressure, the density and the temperature as functions of r)[3] in a spherically symmetric object gives a 'surface redshift' (z from equation (9)) lower or equal than 0.62. Consequently, the gravitational redshift cannot be accounted for the high redshifts the QSO had.

The third alternative is due to the cosmological effects. There are multiple observational proofs that the universe is expanding, and redshift of distant objects, such as other galaxies or extragalactic objects, may be explained as a consequence of space and time dilation.

The Friedmann-Robertson-Walker (FRW) metric is the most general metric solution to Einstein's field equations that satisfies homogeneity and isotropy (Garcia-Bellido, 2005). It is given by:

$$ds^2 = c^2dt^2 - a^2(t)\left[\frac{dr^2}{1 - Kr^2} + r^2d\Omega^2\right] \tag{10}$$

Here, $a(t)$ is the scale factor, related to the physical size of the universe, $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$, and K characterizes the 'spatial curvature': $K = -1$ gives an open universe, $K = 0$ gives a flat universe and $K = +1$ gives a closed universe (conveying $r < 1$, while the other cases do not convey any such restriction).

Now assume, as a first approximation, that galaxies behave like point particles that follow space-time trajectories that are geodesics, with fixed comoving (that is, expanding along with space) coordinates (r,θ,ϕ) in the so-called cosmological rest frame[4]. Now

[3] I.e. the speed of sound cannot surpass the speed of light.

[4] In reality, galaxies show peculiar movement, that is, movement relative to the cosmological rest-frame. In the case of QSO, it is virtually negligible in most cases.

consider a beam of light travelling from a galaxy at $r = r_1$ to our galaxy at $r = 0$ (homogeneity of the universe allows us to choose an arbitrary origin of coordinates). For the light trajectories, $ds = 0$, and so

$$\frac{dr}{\sqrt{1-Kr^2}} = -\frac{cdt}{a(t)} \tag{11}$$

The negative sign indicates that as time passes ($dt > 0$), r decreases. Let's work, for now, with each monochromatic component of emitted wavelength λ_{em}, that will 'stretch' as the universe expands. Consider two consecutive crests emitted at t_{em} and $t_{em} + \lambda_{em}/c$ that reach us at $t_{obs} > t_{em}$ and $t_{obs} + \lambda_{obs}/c > t_{em} + \lambda_{em}/c$, respectively. Integrating (11),

$$\int_0^{r_1} \frac{dr}{\sqrt{1-Kr^2}} = \int_{t_{em}}^{t_{obs}} \frac{cdt}{a(t)} = \int_{t_{em}+\lambda_{em}/c}^{t_{obs}+\lambda_{obs}/c} \frac{cdt}{a(t)} \tag{12}$$

The left side, is a function of r_1, with analytical expression:

$$\int_0^{r_1} \frac{dr}{\sqrt{1-Kr^2}} = f(r_1) = \begin{cases} \arcsin(r_1), & \text{if } K = 1 \\ r_1, & \text{if } K = 0 \\ arcsinh(r_1), & \text{if } K = -1 \end{cases} \tag{13}$$

But what matters is that it remains constant, that is, the second and third integrals in equation (12) are equal. Dividing them in intervals,

$$\int_{t_{em}}^{t_{em}+\lambda_{em}/c} \frac{cdt}{a(t)} + \int_{t_{em}+\lambda_{em}/c}^{t_{obs}} \frac{cdt}{a(t)} = \int_{t_{em}+\lambda_{em}/c}^{t_{obs}} \frac{cdt}{a(t)} + \int_{t_{obs}}^{t_{obs}+\lambda_{obs}/c} \frac{cdt}{a(t)} \tag{14}$$

The second integral of the left-hand side cancels out with the first integral of the right-hand side. It can be assumed that the periods are small enough, $\lambda/c \ll$, and therefore $a(t)$ is approximately constant in the integration intervals. This way,

$$\frac{\lambda_{em}}{a(t_{em})} = \frac{\lambda_{obs}}{a(t_{obs})} \implies \frac{a(t_{obs})}{a(t_{em})} = \frac{\lambda_{obs}}{\lambda_{em}} \overset{def}{=} 1 + z \tag{15}$$

This is the definition of the cosmological redshift. Hubble's law can be easily derived from here. If $K = 0$, or $K \neq 0$ but $r_1 \ll 1$, and $a(t)$ varies slowly in the interval (t_{em},t_{obs}),

$$r_1 \approx f(r_1) \approx \frac{c(t_{obs} - t_{em})}{a(t_{obs})} \tag{16}$$

On the other hand, the slow variation of $a(t)$ allow us to expand its Taylor series and neglect $\mathcal{O}((t_{obs} - t_{em})^2)$.

$$\frac{a(t_{em})}{a(t_{obs})} \approx 1 - H_0(t_{obs} - t_{em}) \tag{17}$$

Here, $H_0 = \frac{\dot{a}(t_{obs})}{a(t_{obs})}$ is the 'Hubble constant', that must be calibrated observationally[5]. On the other hand, also assuming $z \ll 1$,

$$1 - z \approx \frac{1}{1+z} = \frac{a(t_{em})}{a(t_{obs})} \approx 1 - H_0(t_{obs} - t_{em}) \implies z \approx H_0(t_{obs} - t_{em}) \tag{18}$$

Therefore, substituting $(t_{obs} - t_{em})$ in equation (16),

$$r_1 \approx \frac{cz}{H_0 a(t_{obs})} \implies H_0 D_L(r_1) \overset{def}{=} H_0 a(t_{obs}) r_1 = cz = v_H \tag{19}$$

Where D_L is called the luminosity distance, and represents how the 'original distance' r_1 has been enlarged by the expansion of the universe, quantified by $a(t_{obs})$[6]; and v_H is the recessional velocity of the galaxy that would explain the redshift as a non-relativistic Doppler effect. Note how, as z approaches and even surpasses 1, this approximation and the velocity assumption are manifestly false: the galaxy cannot travel faster than the speed of light. For higher redshifts, Hubble's law still applies (redshift is proportional to distance), but with a non-constant Hubble parameter $H(t)$, with $t = t(z)$ the look-back time (see Condon and Matthews, 2018 and Goswami et al., 2015 for more details on this).

A technical remark before going on. With high enough redshift, one must be careful with how the transformation of space and time affects the value of the physical quantities. By the very definition of redshift, we already know that $\lambda' = (1+z)\lambda$ is the relation between wavelenghts received on Earth (left side, λ') and actually emitted (right side, λ). A similar correction may be applied to the (density) flux in [J/(s m^{2}Å)]. Following the steps in Kembhavi and Narlikar (1999) and Condon and Matthews (2018), one can get that:

$$F'(\lambda') = \frac{F(\lambda)}{1+z} \tag{20}$$

This correction will be applied both ways throughout this work.

2.3. Quasars

The first definition for quasar was given in G. Burbidge and M. Burbidge (1967) and goes as follows: a quasar (or QSO) is a point source (not resolved) with radio source,

[5] According to Riess et al. (2019), the value is between 65.5 and 75.5 km/(s Mpc). Its units are velocity/distance, for it gives a linear relation between our distance to the object and the recessional velocity of the galaxy.

[6] Strictly speaking, both D_L and H_0 depend on time (r_1 did not), that is why H_0 is usually termed as $H(t)$, the Hubble parameter.

variable light, large ultraviolet flux of radiation, large redshift[7] and broad emission lines in their spectra, usually showing absorption lines as well. Later on, the radio source property became obsolete, as only a small percentage (around 10 %, according to Kembhavi and Narlikar, 1999) of QSO showed it. For my own purposes, the definition does not require the radio source property.

Quasars are part of the more generic group of active galactic nuclei (AGN), which include other stellar objects, such as Seyfert galaxies or blazars, usually distinguished by certain observational traits. A strong bias towards observational definitions can be noted through the literature, partly for its convenience in the analysis of great amounts of data, and partly for the lack of a complete and satisfactory theoretical framework for such objects until recently (as opposed to stars and their blackbody-like spectra)

One may then try to, if not define QSO by their physical properties, establish a plausible theory of QSO that supports the observational definition given. The most extended model of QSO postulates that QSO, and AGN in general, are spinning supermassive black holes with accretion disks that emit a huge amount of energy ('active') of gravitational source, located in the center of galaxies ('galactic nucleus'). The current unified models propose that all AGN are a single type of physical object, merely observed from a different relative orientation, as represented in figure 2.

In figure 2, the structure of AGN according to current unified models is shown. Note that they are (approximately) symmetrical around their axis and with respect to the plane that contains them: the AGN may have or not have the jets, but if it does, it has two. The representation of figure 2 simply includes radio-loud (those with jets, as they are the main radio source for AGN; upper half) and radio-quiet AGN in the same image (lower half). An AGN consists on a supermassive black hole, surrounded by an accretion disk and a torus-shaped dusty region, and sometimes two huge jets (even bigger than the diameter of an entire galaxy). The material closer to the black hole emits broad lines, as it spins rapidly and suffers a lot of Doppler effect, and the material farther away from the black hole emits narrow lines. This structure finds justification in the accordance between its predictions and the actual data measured.

Straightforward computations, either analytical or computational give us an estimate of the order of magnitude of some QSO properties. Using the period T of the variations of the flux over time and arguing that the size of the QSO is of the order cT, one gets that their size is approximately that of the Solar System. Their lifespan[8] is very short, astronomically speaking: it is of the order of ten million years. For a standard redshift value of $z = 2$ (see section 2.2 for more details on redshift), one gets that the order of

[7] According to Chapter 6 of Kembhavi and Narlikar (1999), z$\geq$ 0.1, although in the data I treat z$\gtrsim$ 2.

[8] Defined as the period of time where there is sufficient gas to form an accretion disc similar to that of figure 2.

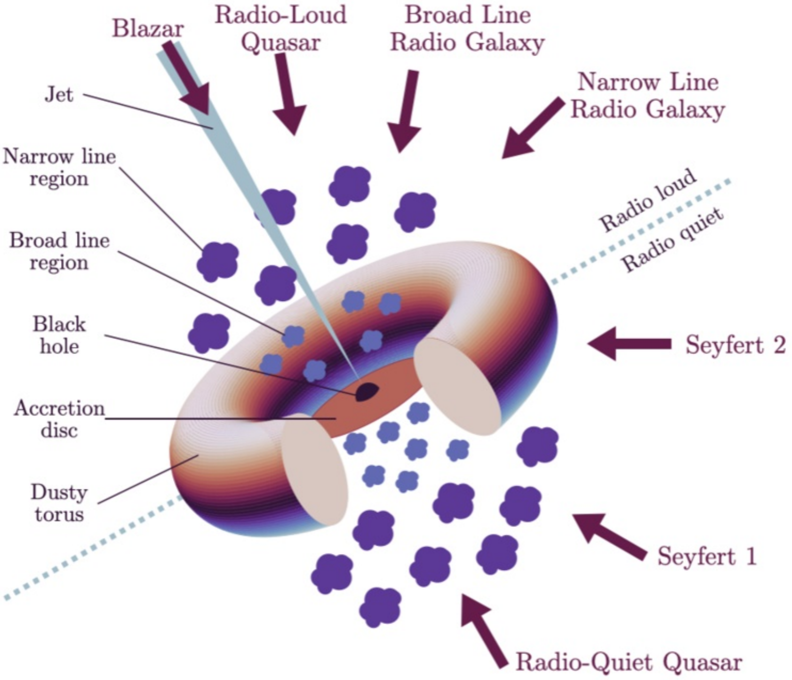


Figure 2: Structure of an AGN (left side, from jet to dusty torus) and different types of AGN (right side, from Radio-Loud Quasar to Radio-Quiet Quasar) as proposed by the unified theory. The types of AGN are merely differentiated by the angle of vision. Not to scale. Image from Urry and Padovani (1995).

magnitude of their distance from us is around 10^{26} m, and that they existed around $3 \cdot 10^9$ year since Big Bang (that is, $20 - 30$ % of the age of the universe). We are looking at the QSO as they were back in the past, for they are most likely extinct nowadays, as is suggested by the representations of QSO density vs age of the universe based on current surveys, which can be seen for example in Geller et al. (2019). The peak of this density is close to $z = 2$. Knowing their apparent magnitude and distance from us, I can also estimate the QSO luminosity (see subsection 2.1), getting that it is as high as that of an entire galaxy of around 10^{11} stars!). Obviously, for different z these numbers vary significantly, and the uncertainty of these numbers is high, but it gives an idea of the real numbers and therefore how astonishingly exotic these celestial bodies are.

The spectra of a QSO consists, basically, on a continuous spectrum and a series of emission lines of the elements it has, along with some absorption lines originated in the source or between the source and us.

The continuous component comes from the broadening of the emission lines of the material surrounding the black hole (in other words, the interactions between atoms, ions and molecules spread out the initially discrete emission lines of the QSO material, so they become no longer distinguishable). It is believed to be originated via relativistic processes like Doppler effect caused by movement in our line of sight, or via inverse

Compton scattering, which consists on a low-energy photon interacting with a high-energy, ultrarelativistic electron, gaining energy from it. This continuum can roughly be estimated by a rather simple power-law in the ultraviolet region (between 100 Å and 4000 Å):

$$F(\lambda) = A\lambda^{\alpha} \tag{21}$$

The power α is usually called the 'slope' of the QSO, as it becomes the slope of a line equation after applying logarithms on both sides (it is the slope of the line in logarithmic representation). Its values are usually between -2 and -1, as stated in Kembhavi and Narlikar (1999). The reasons why equation (21) is used are, mainly, that the continuum varies slowly for most QSO, and the power-law was experimentally proven to be valid at first order in this region of the spectrum. Additionally, as is showed of Kembhavi and Narlikar (1999), in some regions and under certain assumptions, power laws can in fact be derivated analytically. More details on the origin of the power law and other characteristics of QSO spectra are out of the scope of this work, and can be found in of Kembhavi and Narlikar (1999).

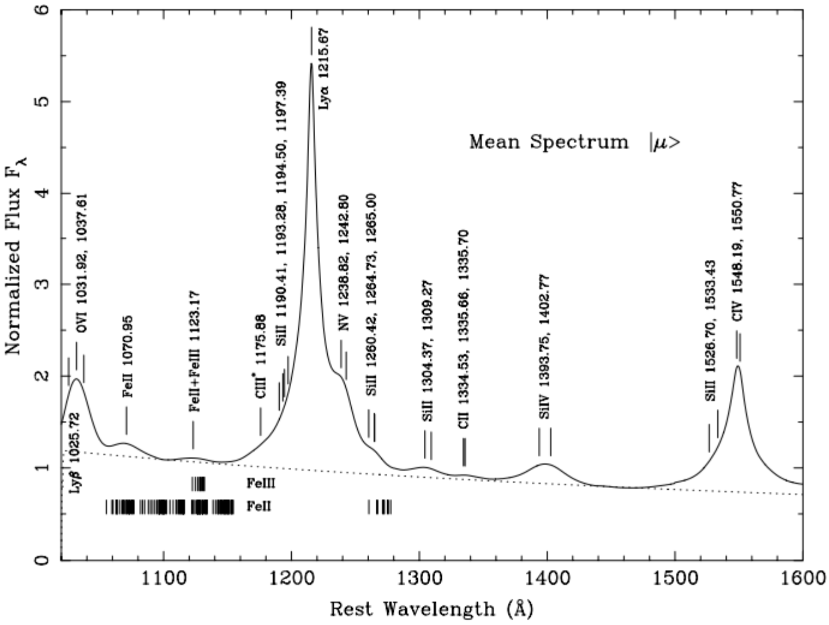


Figure 3: Main emission lines around the Lyman-α peak. This figure is from Suzuki (2005). Arbitrarily normalized flux is represented against wavelength, for the mean of 50 QSO spectra (the author uses Dirac notation for the spectra). The FE II and Fe III multiplets are indicated below the spectrum. The '*' in 'CIII*' indicates that there are several emission lines close to that wavelength. These lines will be present in the OpR3b data I will treat, their relative height varying for each QSO. To consult a broader range of wavelengths, if needed, check figure 2 of Francis et al. (1991). The dotted line is the power-law regression of the continuum component.

The emission lines are originated from the desexcitation of atoms which are accelerated to relativistic velocities and acquire very high temperatures due to collisions and friction as they fall spirally towards the black hole. The line emission is very strong in QSO, causing that part of them stay as broad emission lines instead of completely blending in with the continuum, as would happen with most stars.

As hydrogen is the main component of the orbiting and falling material surrounding the black hole of a QSO (which causes the absorption and emission of radiation), I may remind here Bohr's formula for the wavelengths of the emitted photons after electric transitions inside a hydrogen atom:

$$\frac{1}{\lambda} = R\left(\frac{1}{N^2} - \frac{1}{n^2}\right) \tag{22}$$

Here, $R = 1.097 \cdot 10^7\ m^{-1}$ is Rydberg constant, and N and n are the numbers of the inner and outer orbit, respectively. In particular, for $N = 1$, it gives the Lyman series, which lays in the ultraviolet region. Starting from $n = 2$, the spectral lines are called L_α (short for 'Lyman-α', at 1215.67 Å), L_β (at 1025.72 Å)... For the rest of the elements, values showed in figure 3 are empirically measured and have long been tabulated.

Lastly, as chapter 13 of Kembhavi and Narlikar (1999) points out, absorption lines can be classified into three categories:

- Broad absorption lines, trough-shaped. These are believed to be caused by the disk of accretion and the gas close to the black hole.

- Heavy element (carbon, magnesium, oxygen...) absorption lines. If the redshift of this absorption line is comparable to that of the quasar, then it may be attributed to the quasar itself. If it is considerably smaller, galaxies or their halos are believed to cause these lines.

- Lyman-α forests. Neutral hydrogen is, by far, the most abundant element in the universe. As light travels between the quasar and us, clouds of neutral hydrogen in the IGM (intergalactic medium) produce absorption lines in the Lyman-α line wavelength at their redshift. As the density of the universe is not uniform, a trough is not formed, but rather a series of absorption lines bluewards of the redshifted Lyman-α. That is the reason why this configuration is a 'forest': its rather messy shape.

Lyman-α forests are precisely one of the reasons why it is interesting to reconstruct the continuum. If precise information of this continuum is know and precise measurements are available, one can calculate how much was absorbed at every redshift (that is, at every distance from the Earth), and therefore obtain estimates of the density of the universe in the line between the quasar and us. This data is very valuable as it can be used to constrain the cosmological model we live in, constituting one of the main reasons why quasars are interesting to both theoreticians and observational astronomers.

In this work, I will focus exclusively on the UV part of a QSO spectrum. For an explanation on the shape of the rest of the spectrum of QSO, please refer to Leipski et al. (2014) or Harrison (2014).

3. Mock spectral data for WEAVE-QSO

In this section, firstly, I will introduce the WEAVE-QSO Survey, and secondly, I will introduce mock data. In particular, I will introduce why and how mock data, such as the one I work with in this work, is generated. The sketched procedure (subsections 3.2 and 3.3) will explain some of the features of the OpR3b data that will be treated in sections 4 and 5, which will be first presented at the end of this section (subsection 3.4).

3.1. WEAVE-QSO

Let me review and briefly summarize the objectives of the WEAVE-QSO survey. For further details, please refer to Pieri et al. (2016). The aim is to observe around 400.000 high-redshift ($z > 2$) quasars with magnitudes $m < 24$ as part of the broader WEAVE survey, using the William Herschel Telescope (WHT, in Roque de los Muchachos Observatory, La Palma, Canary Islands). The WEAVE Survey consists of 3 galactic Surveys and 5 extra-galactic Surveys, one of them being WEAVE-QSO. It will use 70 % of WHT time and will have its first light presumably in the last months of 2022 (it is planned for September 1^{st}).

The goal of the WEAVE-QSO Survey is to shed some light to big questions about the nature of dark matter and energy, the reionization of the universe after the Big Bang and the formation of galaxies.

3.2. Usages of mock data

There are two main usages of the mock data for new astronomical instruments and, in particular, new telescopes. The first one is, mainly, to evaluate what science could be done with that specific instrument. For instance, in this case, the data generated for the WEAVE Survey focuses on quasars on a specific region of the sky. How much information one can get from these data serves as an estimation of how much information one will actually be able to gather using the telescope, and therefore helps deciding whether the observational project is worth investing in or not.

The second one is to already have the analysis pipeline ready when actual data from the real instrument is available. For example, this work will show whether adjusting the slope α will be plausible and useful or not, and will study the best set of parameters for certain spline interpolations (see section 5).

3.3. Generation of mock data

3.3.1. Survey generation

I am going to explain the process by which mock data is generated for any instrument (and therefore, in particular, how the data I study in this work was generated) as it may be illustrative to understand certain characteristics of the data before its treatment. Please, refer to Bautista et al. (2015) for more detailed information about this topic.

The first step is to create a survey of the objects which are going to be studied in particular. In this case, those objects are quasars, located in the center of distant galaxies. In order to do so, the following procedure is carried out:

1. After choosing a cosmological model, using the cosmological equations, an initial density field is evolved (either analytically or through numerical situation).

2. Using the density field and a model for the location of QSO, a number of quasars is obtained, along with their positions and their magnitudes.

3. These QSO are filtered by the region of the sky and the magnitudes the instrument will study.

This way, one may obtain as many surveys as one wants for the chosen cosmological model.

3.3.2. Observation modelling

To each and every QSO generated in this survey, a QSO template[9] is asigned. In the particular case of OpR3b (the data set I treat), exactly 452 templates were used. I will refer to this spectra with the sub-index 'temp' (from 'template'). Note that these templates give the shape of the spectrum, proportional to the real spectrum actually emitted. Specifically, they are stored in arbitrary units.

Each QSO is associated to a certain redshift and apparent magnitude in r-band. Using the template associated to the QSO, the appropriate cosmological absorption lines (which are dependent on redshift) are added either analytically or through simulation. The resulting spectrum will be referred to with the sub-index 'model' (as it embodies the effects of the chosen cosmological model in the previous template).

[9] Based on theoretical models or previous observations. Further details are out of the scope of this work, and can be found in papers like Lusso et al. (2015) and Vanden Berk et al. (2001).

In order to generate the final 'mock' data resembling actual measurements with the new instrument, first of all, the quasar's flux renormalized in order to match the required magnitude and redshift. Now, to all of these mock spectra, it is simulated how the spectra would be observed through the instrument, recreating atmospheric effects, night-dependent effects, the different nuances of the instrument itself...

This is some raw data, similar to the data one would get directly using the instrument when it is available. At this step, one may forget that these are artificial data, and act as if they were real measurements. Therefore, to interpret these data, one must correct all of these undesired or 'artificial' features: the instrument's effect through its efficiency function, the atmospheric noises... This process is called 'reduction', it cleans most of the contamination effects and quantifies the uncertainties of the data. These final data will be referred to with the sub-index 'mock' (these are the data one would actually work with in real-life observations).

It is interesting to point out how the atmospheric corrections are carried out. Basically, the empty night sky spectrum, conveniently normalised, is subtracted from the measured spectra. This night sky spectrum is season-dependent, vary between the different regions of the sky and, of course, is also dependent on the location in the Earth, but its main features are the ones represented in figure 4.

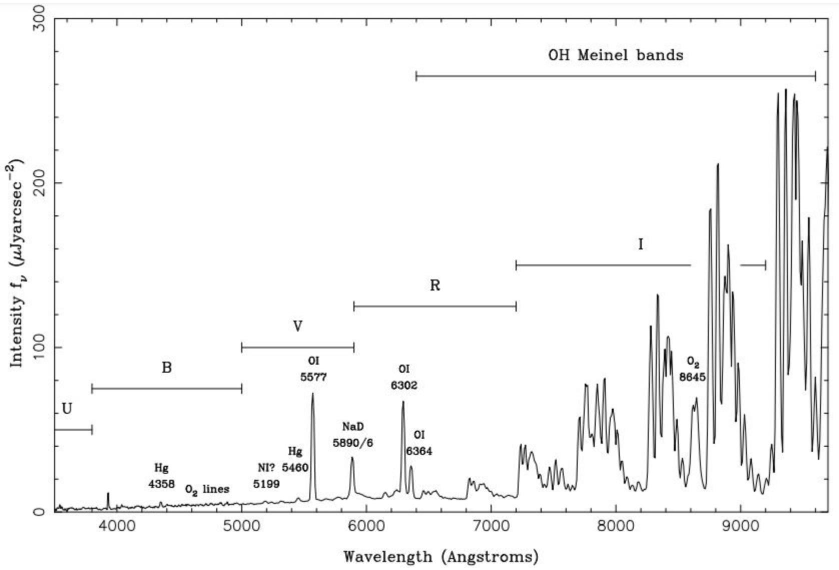
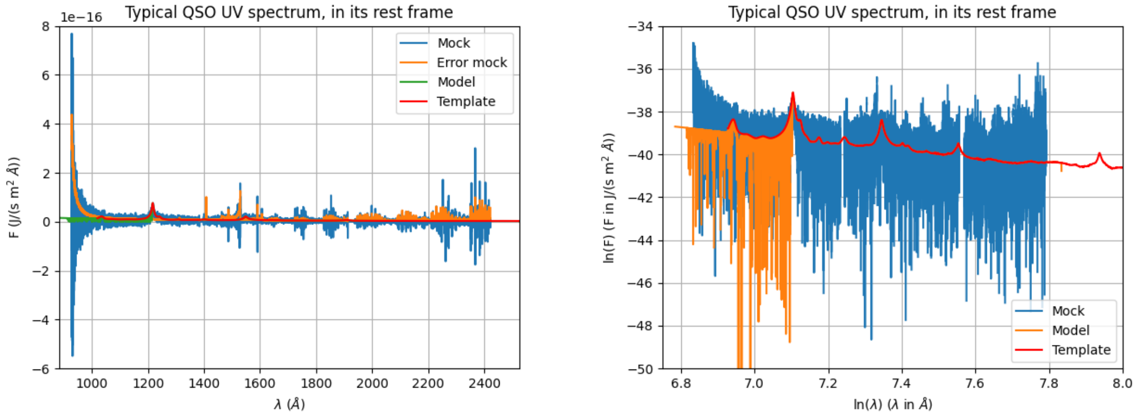


Figure 4: Spectrum from a moonless night sky in La Palma. Figure from Benn and Ellison (1998). The intensity I_ν is represented against wavelength. The flux $F = F_\lambda$ can be obtained as the integral of I_ν over the solid angle used and taking into account $|F_\lambda d\lambda| = |F_\nu d\nu|$. The horizontal lines such as 'U' and 'B' represent filter bands.

Different features can be noted, such as oxygen lines and mercury lines, the NaD doublet and the OH Meinel Bands (transition between rotational-vibrational levels within the OH ground state). These features are mostly due to airglow (emissions from atoms and molecules excited by solar ultraviolet radiation during the day), streetlight contamination and zodiacal light (sunlight scattered by interplanetary dust). A small component of scattered light from faint stars is also relevant.

As these features have some variation and inherent uncertainty, correcting the atmospheric effects leads to considerably higher uncertainties and noisier data in certain regions (for example, those corresponding to the OI lines and the OH Meinel Bands). In the panels in figure 5, the direct correlation between the errors in the typical spectra from the data I will treat and these regions from figure 4 becomes apparent.



(a) Typical shape of a QSO spectrum. Flux is represented against wavelength. Note the lack of physical meaning of the lower wavelengths.

(b) Typical shape of a QSO UV spectrum, using base e logarithms for both flux and wavelength. Note that it has a linear behaviour.

Figure 5: Typical spectra from the OpR3b mock data. The 'mock' data represents observations, the 'template' data represents the theoretical spectrum and the 'model' data represents this theoretical spectrum after the cosmological absorptions. The errors of the 'observed' (that is, 'mock') are included as well in panel 5a.

In figure 5, one of the quasar spectra from the files is represented in both linear and logarithmic scale. Its redshift is $z = 2.96$ and its magnitude $m = 22.92$. In both figures, the Ly-α peak, as well as the noise properties I mentioned before can be observed. Around 1400 Å and 1900 Å a 'flux drop' can be seen (the 'mock' flux is 0 in those regions). These are the regions where the noise is so high that the data is not reliable, and is therefore disposed-of. It can be seen as in the logarithmic representation as well: the parts without data (as the logarithm of 0 is not defined, those points are not represented). These flux drops will be cumbersome, as they do not occur in the same ranges of wavelengths for every QSO, thus eliminating them for my analyses is one of the challenges to overcome.

The renormalization used for this and all other representation follows the steps from subsection 4.1. Note how the true continuum (from either the 'template' or the 'model') follows a power law (in figure 5b, this is very clear). However, also note how such a dependence in the 'mock' spectrum is not that clear due to the noise, and how obtaining the red line from the blue data, which is my objective in sections 4 and 5, is not exactly trivial.

3.4. The OpR3b WEAVE-QSO mock data catalog

As may be inferred from the previous subsections, data may be presented in different frames of reference. In the so-called 'QSO rest frame', the values of wavelength and flux (along with its uncertainties, if applicable) are those one would measure if one were where the QSO is. Therefore, in this frame the Lyman-α peak is located at a wavelength of 1215.67 Å. On the other hand, data may be presented as they are (or would be) measured here on Earth. In this frame of reference, the Lyman-α peak is located at a wavelength of $1215.67 \cdot (1+z)$ Å, which is obviously dependent on the quasar. In order to clarify the frame of reference I am working on, I will adopt the following convention: the super-index 'rf' will refer to the 'QSO rest frame' (λ^{rf}, F^{rf}), while its absence will denote that the data is in our rest frame.

The data I worked with is the OpR3b WEAVE-QSO mock data (Operational Rehearsals, version 3b), and it consists on 3196 QSO unevenly divided in 19 files, each of them corresponding to an 'observational block' (certain amount of time, one hour in the case of WEAVE-QSO, dedicated to obtaining the relevant data from a small portion of the sky, which is different for every block, and corresponds to 2° diameter sections for WEAVE-QSO). For each of the quasars, there is the following information in the file:

- The r-band apparent AB magnitude of the QSO.

- The redshift of the QSO.

- The template spectra: A vector of uniformly-spaced wavelengths, which I will denote as λ^{rf}_{temp}, and another vector of the values of the F^{rf}_{temp} corresponding to those wavelengths. These are given in the QSO rest frame, and λ^{rf}_{temp} ranges from 1020 Å (approximately at Lyman-β) to 5100 Å, increasing approximately 0.1 Å every pixel (a total of 40800 points). These and all other values of the flux are given in J/(s m^{2}Å)

- The cosmological model spectra: A vector of uniformly-spaced wavelengths, which will be denoted as λ^{rf}_{model} (not necessarily coinciding with any wavelength from λ^{rf}_{temp}), and another vector of the values of the F^{rf}_{model} corresponding to those wavelengths. These values are given in the QSO rest frame, and covering different ranges (although always containing those from the observed or 'mock' data) in each QSO with a total of 5397 points.

- The observed spectra: Three vectors this time. A vector of equally-spaced wavelengths (not coinciding with the wavelenghts from the other two wavelenght sets), denoted as λ_{mock}, a vector of the values of the F_{mock} corresponding to those wavelengths, and a vector of uncertainties ΔF_{mock} associated to those values of the 'measured' flux. The λ_{mock} vector is always {3676 Å, 3676.25 Å, 3676.50 Å,..., 9593.75 Å}, it corresponds to the wavelengths that will be measured by the WEAVE spectrograph. This way, the observed spectra covers a total of 23672 data points for each QSO. It must be noted that due to the noise and the subsequent corrections, in regions where the flux is small, F_{mock} may become non-negative, which holds no physical meaning (although it is consistent with positive, close to zero flux taking errors into account). Furthermore, due to the shape of the telescope's efficiency, the values of the flux corresponding to the wavelengths closer to the lower bound have great uncertainty, more so with high magnitude (less luminous) quasars.

An important variable that is not included in the data but is easily computable is the 'signal-to-noise ratio' S/N is defined at any wavelength λ_{mock} as:

$$S/N = \frac{F_{mock}(\lambda_{mock})}{\Delta F_{mock}(\lambda_{mock})} \tag{23}$$

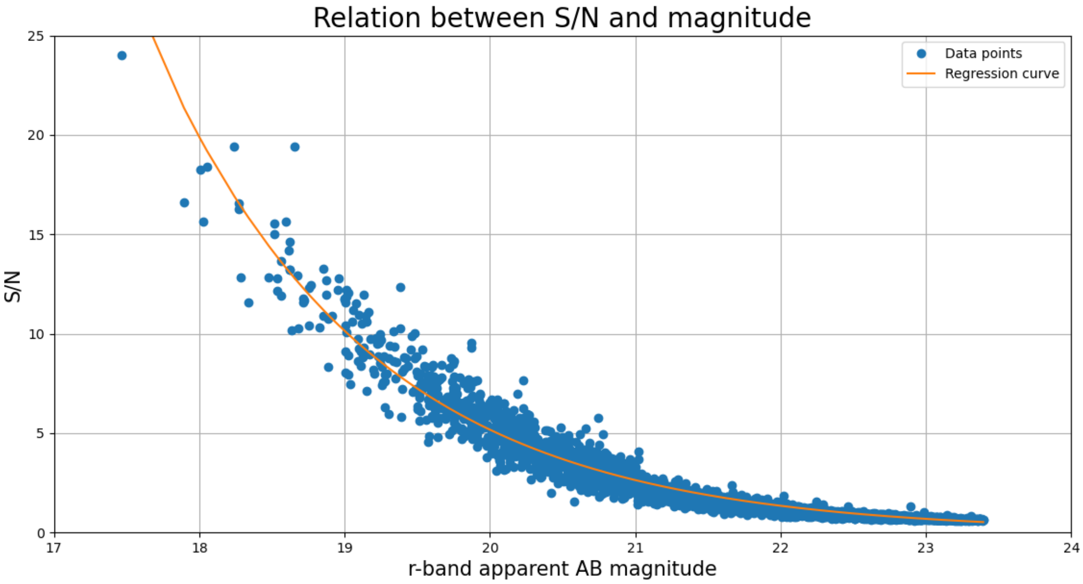


Figure 6: S/N is plotted against r-band apparent AB magnitude for the 3196 quasars of the WEAVE-QSO OpR3b dataset (blue points). The regression curve is plotted as well (orange line).

It is basically the inverse of the relative error, and gives an idea of the reliability of that data point: the higher, the more reliable. It is obviously dimensionless (as so are magnitude and redshift). For convenience, I will also define the average signal-to-noise ratio in a region as the average of the S/N values for that region. If the region is the whole measured range of wavelength, this new quantity (which, by convention, will be denoted as S/N too), shows the quality of a QSO spectrum as a whole. It is expectable that S/N will be directly correlated to the observed flux (therefore, higher for smaller apparent magnitudes). In fact, figure 6 may be obtained with the 3196 quasars from the data set.

The relation between the two variables of figure 6 seems to be an exponential (implying that S/N is proportional to a power of the integral flux, at least in the r-band). After a regression is done in the form of $y = A + Be^{-cx}$, with $(x,y) = (m,S/N)$, I present the following results:

A	**B**	**c**
0.00 ± 0.03	$(37 \pm 4) \cdot 10^5$	0.674 ± 0.005

Table 1: Results of the curve fitting to an exponential function between S/N and r-band apparent AB magnitude. The result is consistent with the shape showed on figure 6.

In particular, $A = 0$, as it should, because as the QSO becomes fainter, $m \to \infty$, and the spectrum must tend uniformly to 0, implying $S/N \to 0$.

An important matter is that, when analyzing the data, I will have to take into account the fact that the samples are not uniformly distributed, neither in redshift nor in magnitude or S/N, as is shown in histograms 7 to 10. I am left, then, with a fairly heterogeneous data set (corresponding, however, to the distribution expected for the WEAVE-QSO catalog).

In figures 7 and 8, histograms for the redshift and the r-band AB apparent magnitude of the QSO are represented. They are not uniformly distributed, nor follow a normal distribution. They reflect what the telescope will be measuring in the specific section of the sky it will point at.

On the one hand, figure 9 represents the QSO density, while its x and y coordinates represent its redshift and r-band apparent AB magnitude, respectively. One can see that the quasars of the data files have redshift $z > 2$, and that there is a region ($2 < z < 2.45$, $21 < m$) where there are no quasars. This is due to the scope of the survey: In order to sample the Lyman-α forest, higher redshift is necessary, as the wavelengths of the forest must be redshifted enough to be covered by the instrument's range of measurable wavelengths (basically λ_{mock}). However, absorption studies can be done with the rest of absorption lines (redward of Lyman-α), which means some quasars with $2 < z < 2.45$ will also be included in the survey, but for them more reliable data (higher S/N, that is, lower magnitude) is preferred, so in particular only $m < 21$ are included.

On the other hand, the redshift and magnitude distribution of the data (figures 7 and 8) are due to the actual densities already known from previous (real) surveys. In particular, figure 7 has the same shape as the QSO density vs redshift distribution I mentioned in subsection 2.3 (I remind it can be checked in Geller et al., 2019), save for the pit at $z \approx 2.25$ which can be explained by the absence of QSO in the $2 < z < 2.45$, $21 < m$ region seen in figure 9.

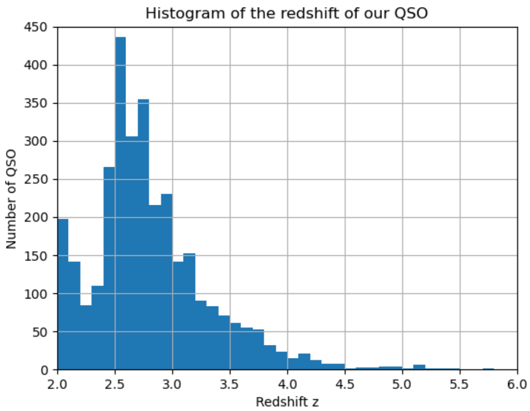


Figure 7: Redshift distribution (histogram) of the WEAVE-QSO OpR3b dataset. Note the discontinuity around $z \approx 2.45$.

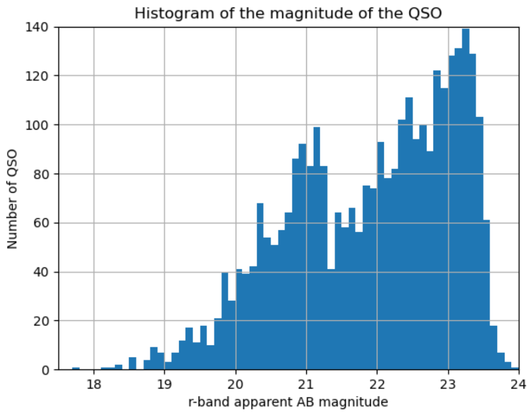


Figure 8: R-band apparent AB magnitude distribution (histogram) of the WEAVE-QSO OpR3b dataset.

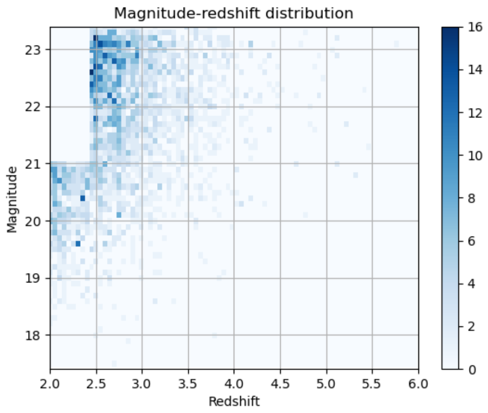


Figure 9: Joint distribution (2D histogram) of redshift and magnitude within the WEAVE-QSO OpR3b dataset.

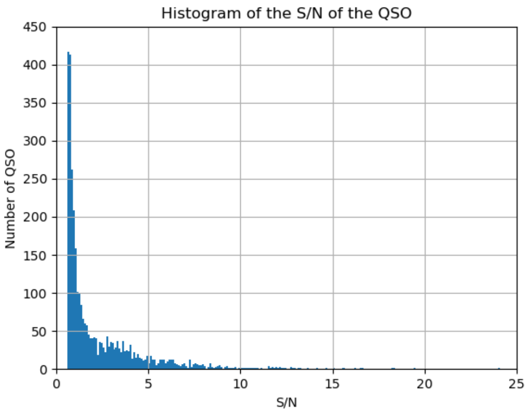


Figure 10: S/N distribution of the WEAVE-QSO OpR3b dataset. Most quasars have low S/N (74.4 % have $S/N \leq 3$).

4. Methodology

In this section, I will focus on introducing two continuum reconstruction techniques, namely computation of the UV slope (also referred to as power-law extrapolation) and cubic spline interpolation. The first one is a classical one, although nowadays, more sophisticated methods are becoming increasingly preferred, such as the spline interpolation or principal component analysis (PCA). The latter will not be discussed in this work (although some characteristics will be briefly mentioned in section 6). However, before explaining the two methods (slope of the UV spectrum and spline interpolation), I will present the method I followed for the renormalization of the 'template' spectra.

4.1. Renormalization of the template spectra

The aim of this section is to describe how to obtain the renormalization constant for the 'template' flux. Let me begin by claiming that I cannot simply make its magnitude coincide with that of the 'mock' spectrum, as absorption lines make the magnitude bigger and are present in the 'mock' but not the 'template'. However, the 'model' spectra do include these absorption lines, so I can begin by normalizing it. First, to transform F_{model} to F^{rf}_{model} in the QSO rest frame, I use equations (6) and (20). The data F^{rf}_{model} is now renormalized multiplying by a constant C_{model} such that its magnitude equals the magnitude required, m_r, which is included as part of the data in the files. Such constant may be easily determined:

$$0 = m^{new}_{model} - m_r = -2.5 \log \frac{C_{model} \int F_{model}(\lambda) T_r(\lambda) d\lambda}{\int F_{mock}(\lambda) T_r(\lambda) d\lambda} = -2.5 \log C_{model} + m^{old}_{model} - m_r$$

Therefore, the constant is simply calculated as:

$$C_{model} = 10^{\frac{m^{old}_{model} - m_r}{2.5}} \tag{24}$$

With the now normalized 'model' spectra, I can obtain the normalization constant for the 'template'. As I mentioned above, it is better to not use the same idea to obtain the C_{cont}, as absorption lines make this calculation imprecise in most QSO. A more correct approach is to impose that F_{cont} is as close as possible to F_{mock} in certain absorption-free regions, using the following lemma, which I include as will come in handy a few more times.

Lemma 1. *Let $\vec{v}, \vec{w} \in \mathbb{R}^n$, and consider the real variable function $f(C) = |C\vec{v} - \vec{w}|^2 = \sum_{k=1}^{n} (Cv_k - w_k)^2$. Then $f(C)$ attains its absolute minimum at $C_0 = \frac{\vec{v}\cdot\vec{w}}{\vec{v}\cdot\vec{v}}$. Also, if both $\vec{v}$ and $\vec{w}$ are non-negative, then $C \geq 0$.*

Proof. $f(C)$ is a convex second degree polynomial, so the minimum is the C_0 such that the derivative is 0.

$$f'(C) = 2\sum_{k=1}^{n} v_k (Cv_k - w_k) = 2(C\vec{v} - \vec{w}) \cdot \vec{v} = 0 \iff C = \frac{\vec{v} \cdot \vec{w}}{\vec{v} \cdot \vec{v}}$$

The non-negativity is trivial. □

Note that I am basically assuming a dependence of the form $y = Cx$ to the data $\{(v_k,w_k)\}_{k=1,\ldots,n}$ and applying least squares just like I would on a standard linear regression, but with interception 0.

The vectors of the lemma are the values of F^{rf}_{model} (already normalized) and those of F^{rf}_{temp}, as $\vec{w}$ and $\vec{v}$, respectively. Linear interpolation is performed in the latest in order to have both fluxes evaluated at the same wavelengths, precisely those of equation (25).

4.2. Slope of the UV spectra of the quasars

As I previously explained in subsection 2.3, the continuum part of the spectrum of a QSO can be fitted by a rather simply power law, $F(\lambda) = A\lambda^{\alpha}$ in the UV region. Through this part of the work, I will explore the slope distribution of the data and the correlation between these slopes and the rest of the properties of the QSO. Afterwards, I will check to which extent this power law is applicable, and whether it is actually possible to obtain consistent results with only the actual measurements or not, depending mainly on the S/N of the QSO. In order to do so, I will apply the fitting to both the 'template' and the 'mock' data. The fitting to the 'template' will show that this approximation is justified, and a comparison between both fittings will show under which circumstances this method of continuum reconstruction is advisable in real life.

Firstly, I may recall here that my main goal is to estimate the continuum of the quasars, especially in the Lyman-α forest region. That way, as this region is mostly emission-free, the continuum corresponds to the spectrum minus the absorption, and thus absorption may be estimated at every distance (or redshift) from the QSO to the Earth, as one may associate it with only Lyman-α absorption. In order to estimate the continuum in this region, an extrapolation from a different region shall be done.

Therefore, a subset of the λ_{temp} and λ_{mock} must be chosen where the regression will be performed. It cannot be the whole set in both cases for a number of reasons. Firstly, the slope is dependent on the region. In fact, some authors propose two or more power laws (at different regions) for each QSO (for example, in Bosman et al., 2021 or Tytler et al., 2004)[10]. However, I will fit only one slope (following the line of works like Dall'Aglio et al., 2008 or Meyer et al., 2019). Secondly what I am reconstructing is the continuum, and therefore I must eliminate the absorption (dominant in the Ly-α forest: the selected region will be redwards of Ly$-\alpha$) and emission lines, whose inclusion would distort my analysis.

Different regions in the spectra were tried to fit the slope, always taking into account the spectral shapes, represented in figure 5 for one of the QSO, and different articles where

[10] I checked this in the OpR3b data: By using a union of intervals between 1340 Å and 4800 Å (in the QSO rest frame), the slopes were higher than expected, the reason being an increment in the slope some quasars presented between 3000 Å and 4000 Å. Therefore, the region selected had to be narrowed to avoid this.

similar fittings were performed to a number of quasars (for example, Bosman et al., 2021 for a rather advanced but very complete article on these kinds of continuum reconstruction techniques). The one that returned the best results is the following (expressed in the QSO rest frame):

$$(1350\ \text{Å}, 1370\ \text{Å}) \cup (1440\ \text{Å}, 1470\ \text{Å}) \tag{25}$$

This region was selected taking into consideration three factors:

- The dispersion of the slopes for both the 'template' and the 'mock' (the lowest possible that is physically acceptable is desirable). Usually, the dependence of the slopes with the intervals used were mostly seen in the 'mock' spectra rather than in the 'template', given the smoothness of the latest. Nevertheless, some dependence was observed in the 'template' depending on the criteria followed to choose where the emission 'lines' begins and ends. A balance between the safer choice of eliminating more points near the line (physically, I am including the points which are closer to the assumed dependence, but mathematically the regression result is less reliable as less data points are used) and the more robust choice of having more points for the regression (quite the opposite, as some points are farther from the assumed dependence but the result is mathematically more reliable) was sought.

- The accordance between the 'template' and 'mock' slopes (that is, how accurate the predictions could be with each choice of region. The method I am proposing is meant to be the optimal regarding usefulness.

- The physical meaning of the fittings obtained. Are they really 'good fits' (usually, but not always, conveying higher r^2 values), or, despite complying with the first two points, the results are not what they are expected to be (most of the slopes must lay between -2 and -1, as stated in subsection 2.3)?

Additionally, in order to obtain a slope that fits better near the Lyman-α forest, I decided it was best to only keep intervals that were closest to it (diminishing extrapolation error and excluding regions with different slopes).

In any case, the condition $S/N > 0.1$ was added to the fitting of the 'mock' data. It excludes some of the points, but is mainly added to avoid using meaningless points and flux drops (as null flux obviously implies null S/N), should they happen to intersect that region for some QSO.

The simple functional dependence $F(\lambda) = A\lambda^{\alpha}$ guarantees that it is not important whether the 'template' data is normalized (a multiplicative constant can be grouped with A), or whether both are expressed in the same frame of reference or not (the difference is only a multiplicative factor $(1+z)^{\pm(1-\alpha)}$, which again can be grouped with A). However, to plot the comparison of the results, consistency in both is required. In this section, for the sake of convenience, I will work in the QSO rest frame. In order to obtain the normalization constant of the 'template' flux, I use the procedure from 4.1.

One may apply logarithms[11] to the assumed dependence $F(\lambda) = A\lambda^{\alpha}$:

$$\ln F(\lambda) = \ln A + \alpha \ln \lambda \tag{26}$$

I will use linear regression (via least squares), of the form $y = mx+n$, where $y = \ln F(\lambda)$ and $x = \ln \lambda$ first to the 'template' and later to the 'mock' observations. Therefore, the slope will give us the value of α. I will make use of the uncertainties (27) as weighs 'w' in the regression, using the convention $w(\lambda) = \frac{1}{\Delta(\ln F(\lambda))}$, where, by the usual propagation of uncertainties,

$$\Delta(\ln(F(\lambda)) = \frac{\Delta F(\lambda)}{F(\lambda)} \tag{27}$$

The weights are, therefore, the S/N values at each wavelength. Another possibility, however, is directly applying least squares to $F(\lambda) = A\lambda^{\alpha}$, which, contrary to popular belief, is not completely equivalent to the previous, simpler[12] way. In fact, as can be read in works like Xiao et al. (2011), the one that is more appropriate for positive-valued vectors depends on the weighs, that is, the distribution of uncertainties of the data. Moreover, in order to compute the linear regression with the logarithms, data points where the flux is smaller or equal to 0 (as I mentioned when explaining the data set, in regions where the 'real' flux is small, the measured flux may become non-positive) must be removed, and therefore information is being lost using this choice. However, computation is easier and faster with the linear choice, so to perform the almost half million analyses of the real quasars WEAVE will measure, if results are acceptable, it may be preferable. I checked both approaches with the WEAVE-QSO OpR3b mock dataset and, although with little nuances, in the end they present identical conclusions.

4.3. Cubic spline interpolation

In this part of the work, cubic spline interpolation will be performed through selected points obtained from each spectrum. A method for the optimal selection of points is sought making use of the 'true continuum' stored in the 'template' variables, so that the interpolation is as close as possible to it, in a sense that I will define. If such a method

[11] Any basis works for this procedure, as the slope 'α' is the same for any of them. In the literature, it is frequent to use basis 10, but in this work, natural logarithms will be used.

[12] As there is an analytical solution, while there is not for $\min_{\alpha,A} \sum_i (F(\lambda_i) - A\lambda_i^{\alpha})^2$, which is solved numerically.

is found, it can be used for real-life applications, as I have tested that the continuum reconstructed with it is close enough to the real continuum, and the error may be estimated (assuming it is similar to those using my data set).

In this case, my aim is not simply to reconstruct the naked continuum, but rather the continuum with the emission lines, basically 'fixing' all the absorption. Let me recall that, in emission-free regions, these two coincide, but maintaining the emission lines ensures richer information can be obtained.

First of all, let me define cubic splines. Given a set of n points in $\mathbb{R}$ $\{(x_k,y_k)\}_{k=1,...,n}$, $x_1 < x_2 < ... < x_n$, we are looking for a piecewise third grade polynomial function $f(x) = p_k(x)$, $x \in [x_k,x_{k+1}]$ such that $f(x_k) = y_k$ $\forall k = 1,...,n$ and f and its first and second derivatives are continuous. Now, as there are $n-1$ intervals $[x_k,x_{k+1}]$, $k = 1,...,n-1$, there are also $n-1$ polynomials, and therefore $4 \cdot (n-1)$ constants to determine (as each polynomial, referred to as 'spline', has third degree, has four constants to determine). On the other hand, $f(x_k) = y_k$ $\forall k = 1,...,n$ are $2 \cdot (n-1)$ conditions ($p_k(x_k) = y_k =$ and $p_k(x_{k+1}) = y_{k+1}$ $\forall k = 1,...,n-1$), and the conditions on the derivatives are $p'_k(x_{k+1}) = p'_{k+1}(x_{k+1})$ and $p''_k(x_{k+1}) = p''_{k+1}(x_{k+1})$ $\forall k = 1,...,n-2$, a total of $2 \cdot (n-2)$ constraints. The system of linear equations that must be solved in order to obtain the coefficients of the polynomials, then, lacks two more equations. A frequent choice is the *natural* cubic splines, where the second derivatives at the extreme points are set to 0: $p''_1(x_1) = 0 = p''_n(x_n)$. In this case, the matrix is tridiagonal and symmetric, and it can be proven that there exists only one solution to the linear system, and therefore only one natural cubic spline.

Nevertheless, as stated before, I am not going to simply apply interpolation between all the points of the 'mock' data, which would not solve the problems with its noisy shape and its absorption. By applying it to a different, smaller set of points carefully created to not represent any absorption line or region, some smooth estimation of the continuum component of the spectrum can be attained. Of course, some methods for smoothing noisy data already exist, such as the moving average, part of the more general Savitzsky-Golay filters, but such smoothing does not produce the best results near emission lines, and most importantly, does not exclude absorption, yielding a fake continuum. The procedure used with splines, thus, aims to solve these two issues.

The algorithm followed is based on Young et al. (1979) and Carswell et al. (1982), and has been implemented before in works such as Dall'Aglio et al. (2008). It divides the spectra in wavelength bins of two different sizes (before and after the Lyman-α peak) and either selects the 'best points' (closer to the average) from it or divides it in two bins if it intersects an emission line, finally using the resulting bins to compute their average wavelength and flux to use them as the points for the spline computation. Let us go over the algorithm in more detail. The following variables and parameters are needed: the vector of wavelengths, not necessarily equally-spaced, in Angstrom; the vector of the observed ('mock') fluxes corresponding to each wavelength; the vector of the the errors (uncertainties) corresponding to each of these flux values; the redshift 'z' of the QSO; 'Δpix_1', representing the width of the bins in pixels (or number of points) blueward of Lyman-α (that is, for wavelengths smaller than $Ly - \alpha$, 1215.67 Å); 'Δpix_2', representing the width of the bins in pixels (or number of points) redward of Lyman-α (that is, for wavelengths bigger than $Ly - \alpha$); 'minpix', or minimum number of pixels in a bin for the computation of the mean flux, which does not vary if there are less than minpix pixels left; 'slopethresh' or threshold for smaller bin sizes in the Lyman-α forest, so that if the slope between two spline points is larger than slopethresh, the bin is divided in two; 'fluxthresh', another threshold for the flux, used to detect and eliminate flux drops; and 'fluxscale', scale factor to be applied to the flux density and errors in order to avoid numerical problems with large or small numbers, which is corrected at the end of the computation, but is not really needed for the data I treat.

The algorithm itself consists on the following steps:

1. The spectrum is divided in bins of sizes Δpix_1 blueward and Δpix_2 redward of Lyman-α.

2. The average noise is computed in each bin, which will generally be smaller than the standard deviation of the data in that bin.

3. On each of the bins redward of Lyman-α, deselect the pixel with the largest deviation from the mean of the bin. On the bins blueward of Lyman-α, only deselect the pixels with negative deviation, as they correspond to absorption lines (Lyman-α forest), which are obviously not part of the desired continuum. Repeat this procedure with the rest of the points of each bin, until either the flux standard deviation is less than the average noise or the number of pixels left is lower than the parameter minpix.

4. Divide the bin size (number of points) by two if the absolute value of any of the slopes between the mean points in consecutive bins is greater than slopethresh.

5. Repeat the third step on the refined bins, and this time directly clip spikes that are more than two times the standard deviation away from the mean of the bin.

6. By comparing each bin with its neighbours, flux drops (see 5) may be detected (if the neighbours have roughly the same mean flux and it is significantly higher than the central bin's) and therefore corrected (replacing the central bin's flux by the

mean flux of the neighbours). This works better for greater bin sizes[13].

7. The mean value of the wavelengths and the fluxes at each bin is calculated, and spline interpolation is performed with these points. In the introduced vector of wavelengths, the new flux value is obtained by merely substituting in the polynomial of the interval it is in. For the formulas to come, this spline-interpolated continuum will be referred to as $F_{spl}(\lambda)$ ('spl' for 'spline').

This spline interpolation was visually and statistically tested for many of the quasars of the data, hardly showing any dependence with 'minpix', 'slopethresh', 'fluxthresh' and 'fluxscale', and thus I kept them fixed at the following default values: minpix=4, slopethresh= 0.033, fluxthresh= 0.99 and fluxscale= 1.0. However, strong dependence of the resulting splines with the values of Δpix_1 and Δpix_2 could be noted, which roughly represent how smooth the reconstructed continuum is in each zone (bluewards and redwards of Lyman-α): the higher these values are, the smoother the splines result. This is due to the fact that bigger bins imply less points I am computing the interpolation with.

Therefore, the only parameters I am going to change are Δpix_1 and Δpix_2. After trying a fine set of values up to 150, the results indicated that the best values seemed to surpass the 150 limit (because most of the selected optimal $(\Delta pix_1, \Delta pix_2)$ were 150). Therefore, and taking into account the very long running times of my code, I opted to obtain a crude estimate but over a large range of values. In particular, I will show results where both independently run over the eleven values:

$$\{10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}. \tag{28}$$

This set was selected to cover the usable range for these parameters with a sufficient amount of values, without adding too many and simply getting way higher running times for the code[14]. I coded in Python a program that runs this analysis and, for every couple of values, computes certain residuals that quantify how close the interpolation was to the true continuum. Three different residuals were computed:

$$resid_1 = \sum_{\lambda_{mock} \in R} |F_{temp}(\lambda_{mock}) - F_{spl}(\lambda_{mock})| \tag{29}$$

[13] In this work I propose additional steps to reduce the influence of these flux drop: I exclude them from the data before calculating the splines, filtering only those points with $S/N > 0.1$, and I also exclude them from the definitions of residuals (29), (30) and (31).

[14] The following fact endorses my choice: the splines' shape vary slowly and continuously moving the Δpix values, and thus, most QSO where, say the optimal value for one of the Δpix is in the (250,350) interval will return the result '300'. Therefore, to see the global tendency of the optimal paramenters, it suffices to restrict oneself to these possible values.

$$resid_2 = \sum_{\lambda_{mock} \in R} (F_{temp}(\lambda_{mock}) - F_{spl}(\lambda_{mock}))^2 \tag{30}$$

$$resid_3 = \sum_{\lambda_{mock} \in R} \left| \frac{F_{temp}(\lambda_{mock}) - F_{spl}(\lambda_{mock}))}{F_{temp}(\lambda_{mock})} \right| \tag{31}$$

Note that I do not have $F_{cont}(\lambda_{mock})$, but can be estimated through linear interpolation, which may be performed as long as the restriction $\min(\lambda_{temp}) \leq \lambda_{mock} \leq \max(\lambda_{temp})$ is imposed. Due to the high sampling in λ_{temp}, no relevant effect (error) from this interpolation is expected, and it was checked to be the case. From the 11 × 11 matrices of residuals, I pick the minimum values, and select the pairs (Δpix_1,Δpix_2) that generated it as the optimal for the QSO and residual I am working with in particular.

The wavelength region R is, still, unspecified. It has to be the region of the spectrum of interest. In this work, I focused on three specific regions:

$$R_{total} = \{\lambda_{mock} | \ \min(\lambda_{temp}) \leq \lambda_{mock} \leq \max(\lambda_{temp})\}, \tag{32}$$

which is basically the whole usable spectrum, in case the information needed is global,

$$R_1 = \{\lambda_{mock} | \ 1040 \text{ Å} < \lambda^{rf}_{mock} < 1190 \text{ Å}, \min(\lambda_{temp}) \leq \lambda_{mock} \leq \max(\lambda_{temp})\} \tag{33}$$

which corresponds to the Lyman-α forest, without the influence of the Lyman-α and Lyman-β lines, and

$$R_2 = \{\lambda_{mock} | \ 1275 \text{ Å} < \lambda^{rf}_{mock}, \min(\lambda_{temp}) \leq \lambda_{mock} \leq \max(\lambda_{temp})\} \tag{34}$$

The region redwards of the Lyman-α line and avoiding its influence. The reason the Lyman-α line was removed from these last two regions is that its flux is usually comparatively higher and its sides are steeper than the rest of the spectrum, producing greater residuals for fittings that were otherwise good enough in the rest of the spectrum. In any case, I remind here that both in the spline computation and in the residuals values, the wavelength filter $S/N > 0.1$ was also applied.

The first residual is the sum of absolute errors, unweighted as the others. The second residual is the sum of absolute errors squared, which penalizes bigger errors more than smaller errors. The third residual is the sum of relative errors, which returns better results if what is sought is the bare continuum, rather than the emission lines (gives less importance to bigger F_{temp}).

Let me remind here that F_{temp} was renormalized as explained in the previous section (see equation (24) and lemma 1). However, this method may not necessarily return the best results possible for every QSO (see figure 26). Besides, the splines oscillate (sometimes being above and below the 'template' flux in consecutive regions). Furthermore, some global absorption may be present in the 'mock' spectra (mainly in the Ly-α forest), causing the calculated splines to be systematically below the real 'template' spectra. Therefore, it seems natural to adjust new normalization constants to both these fluxes, F_{temp} and F_{spl}. Nevertheless, in residuals (29), (30) and (31) one of the constants may become a common factor, and thus only one constant is needed. That is why I tried replacing F_{temp} with $C \cdot F_{temp}$ with a conveniently computed C[15] (for every spline computation, that is, every couple $(\Delta pix_1, \Delta pix_2)$ of the 10×11 matrix), which is expected to have a value close to one.

Several methods for computing this constant were tried as well. The first idea is: the constant shall be the one that minimizes the corresponding residual. It is easy to compute the constant for $resid_2$, as applying lemma 1 yields the desired result (which will be referred to as C_2). Determining such constant for the other residuals, however, is not an easy task, as computing the minimum of the L^1 distance (or taxicab metric) is a non-differentiable problem with worse properties than least squares, and solving it for the 121 couples of $(\Delta pix_1, \Delta pix_2)$ for the 3196 quasars would yield higher running times for my code, which already takes days. Furthermore, the constant that minimizes $resid_2$ already minimizes the (euclidean) distance between the two vectors, and thus using it for $resid_1$ and $resid_3$ is justified as well.

In any case, for $resid_1$ and $resid_3$ I also tried two more different constants: For $resid_1$, I tried imposing that the absorptions in the Lyman-α forest were similar in F_{spl} and F_{temp}, in a sense that:

$$\sum_{\lambda_{mock} \in R_1} (C_1 \cdot F_{temp}(\lambda_{mock}) - F_{mock}(\lambda_{mock}))^2 = \sum_{\lambda_{mock} \in R_1} (F_{spl}(\lambda_{mock}) - F_{mock}(\lambda_{mock}))^2$$

However, as is apparent from the fact that the equation is quadratic in C_1, for some QSO this had no solution, and for others it had two solutions, although only one could be the optimal. In any case, the results with this C_1 (on the QSO where it works) will be briefly commented as well in the results (section 5.2.1). On the other hand, for $resid_3$ although minimizing this residual itself was not a good idea, minimizing the same summatory but with its terms squared becomes an easily-solvable problem, again by lemma 1, and the minimum constitutes the constant C_3.

One last comment regarding these renormalization constants. As splines are oscillating functions, and the bin sizes used are the same irrespective of the QSO in question, for

[15] In real life application, a constant multiplying F_{spl} is sometimes used (of course, in these application F_{temp} is not known), which is computed differently as explained here (more details are out of the scope of this work), and thus adjusting a constant as I do in this work is completely justified.

a small portion of QSO some fits may cause the return of a constant with no physical meaning (for example, a negative constant). Those cases are directly discarded by filtering the values of these constants to be between 0.5 and 1.5.

Another important remark. If Lyman-α lays on 1215.67 Å, when the emitter is at redshift z, the perceived wavelength is $1215.67 \cdot (1+z)$ Å. If the lower limit the telescope can perceive is 3676 Å, in order to detect this peak,

$$z \geq \frac{3676}{1215.67} - 1 \approx 2.024 \tag{35}$$

This is a necessary condition for the program to run properly (it requires a certain range both bluewards and redwards of Ly-α). However, this just is not enough for an analysis with Δpix_1 as a parameter. Depending on the scope, of course, different lower bounds for z could be used. In my case, I opted for:

$$z \geq 2.5 \tag{36}$$

The reasons are mainly the following two. As I stated when explaining the redshift distribution of the data (see figure 7), Lyman-α forest is best sampled when the quasars have $z \gtrsim 2.45$, and the procedure I come up with must be useful for the Lyman-α forest analyses. The other reason is that, as there are 4 pixels per Ångstrom, a total of $4 \cdot [1215.67 \cdot (1+2.5) - 3676] \approx 2315$ pixels is enough to have more than two bins for any $\Delta pix_1 \in \{10, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$. This way, from the 3196 quasars, only 2397 will be considered due to this redshift cut.

Running the previously explained analyses on these 2397 $z > 2.5$ quasars from the data, I will study the dependence of the optimal parameters and the residuals with, say, the redshift, the S/N and the magnitude. In fact, I will suggest an algorithm that maps 'certain QSO properties' $\rightarrow$ (Δpix_1,Δpix_2), where these are the optimal parameters and the properties given are none but the redshift, magnitude and S/N.

In any case, a side result will also be shown in subsection 5.2: before running the analyses with the residuals, I represented and visualized hundreds of combinations of Δpix_1 and Δpix_2 for quasars of different S/N (therefore different magnitudes), selecting the best couple of parameters for each myself (visually). After running the analyses, this selection fell, as expected, in the global tendency that the residual minimization gives.

As the last point to this section, I would like to properly define what a 'good' spline interpolation is in the context of QSO spectra. After representing a number of spline interpolations along with the associated 'template' spectra, I am able to confidently assert that minimization of the residuals defined do produce very good results, but sometimes one of the following cases may unexpectedly occur:

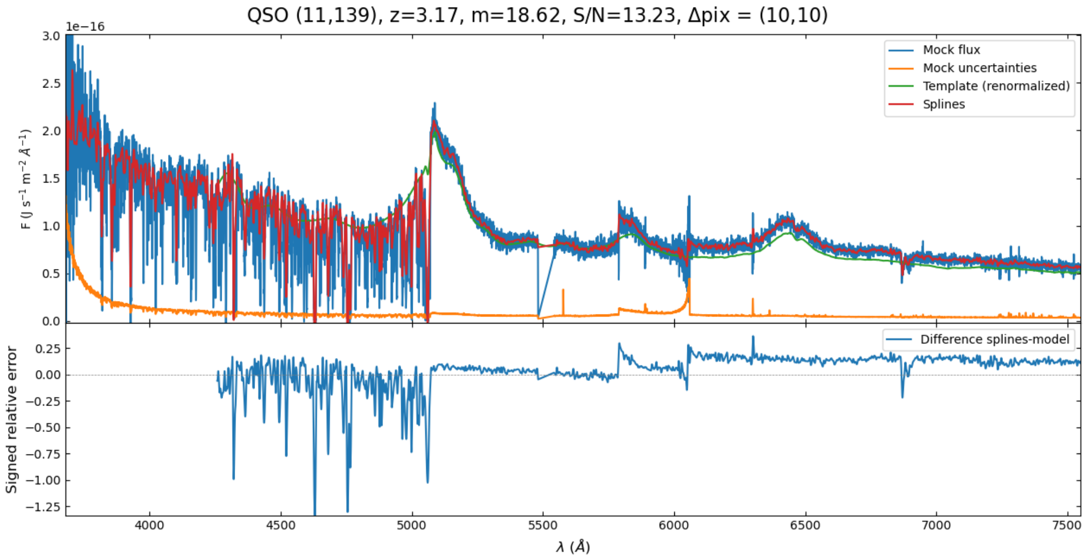


Figure 11: I the upper panel, mock flux, its uncertainties, the spline interpolation and the 'template' flux, renormalized using the constant C_2 over the whole usable spectrum. Note that these $\Delta pix = (\Delta pix_1, \Delta pix_2) = (10{,}10)$ are not the optimal. In the lower panel, the signed relative error ($\frac{F_{spl}-F_{temp}}{F_{temp}}$ with F_{temp} normalized) is represented in parts per unit. The numbers between brackets in the title are simply the file (of the 19 in total) and position within that file. Cut at a wavelength of 7500 Å for display purposes.

In this first figure, although the residual is very low, the little oscillations are not natural, that is, the real spectrum ('template') is globally smooth, while the spline interpolation is not as smooth (this is even more evident bluewards of the Ly-α region, where the splines undesirably pass through many emission lines). Therefore, this fit is not physical. It could be assumable, depending on the usage one wants from the constructed splines, but in this case, this does not even constitute the optimal fit for any residual. Note that it is a luminous, high S/N quasar, therefore these 'wiggles' affect the spline interpolation for any quasar. Also note that the renormalization constant makes the splines be below and above the 'template' bluewards and redwards of the Lyman-α line, respectively (it could be the other way around in some QSO). This suggests that better results can be obtained restricting the wavelength interval where the constant is computed to one of those regions. The results would be better, however, only in that region, but worse in the other, and therefore this restriction must be used only if the information needed is strictly contained in the region in question (for global information, the whole spectrum must be used to compute the constant, as in this figure 11). In any case, and back to the oscillations, in order to avoid them, one must set the parameter in question (depending on the region that presents them) to be ≥ 150 (depending on the QSO, for some it could be enough with ≥ 100, while for others even ≥ 200 could be necessary for more certainty that they are gone, if needed; 150 is some kind of mean value to establish a clear threshold).

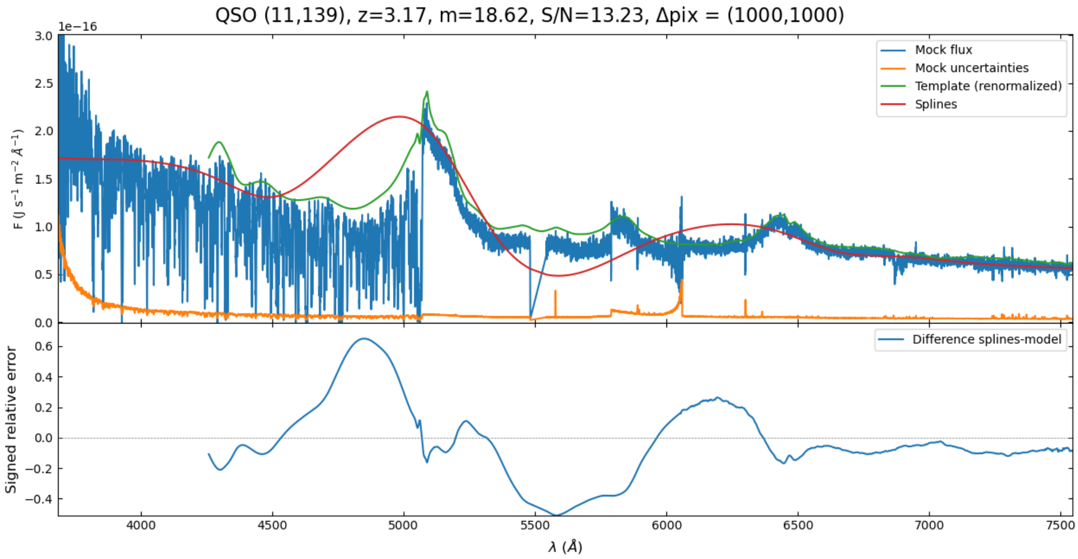


Figure 12: Same as figure 11 but with different Δpix = (1000,1000).

In this figure, however, a different kind of error is shown. For bin sizes too big, the corresponding spline will basically ignore most of the data points, sometimes returning an 'unphysical' continuum (look at the 60 % relative error in the Lyman-α forest). For quasars with lower S/N (this one has high S/N), however, the great fluctuations may not be present, as the spectrum is 'mostly plain' (see figure 5a), so the spline interpolation, computed with these points, will also be plain, and therefore big bin sizes could return the optimal results (in fact, bigger bin sizes are the expected in low S/N quasars, although not necessarily as high as 1000).

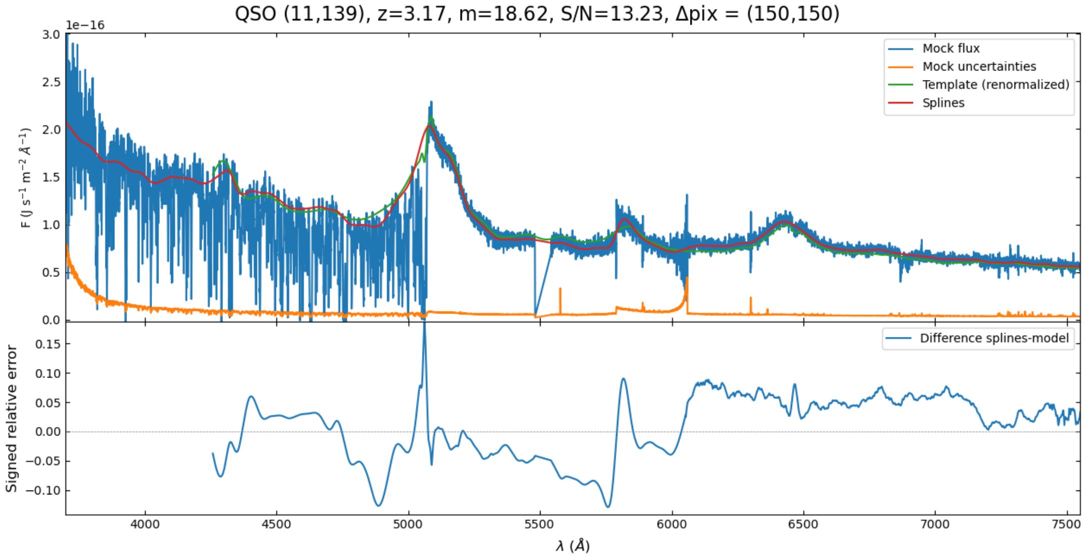


Figure 13: Same as figure 11 but with different Δpix = (150,150).

A good fit, then, may be defined as one having the same shape as the 'template', no visible unphysical oscillations and small relative errors, if conveniently renormalized. An example of this is, for the same QSO as figures 11 and 12, is figure 13:

As this is one of the highest S/N quasars, a first result is that relative errors of around 10 % are the minimum expected. Note that the 'wiggles' seen in the signed relative error plot are due to wiggles present in the template, not the splines in this case (contrary to figure 11). In any case, in appendix C more examples of handpicked optimal fits can be found.

5. Results

In this section I will present the results for the two methods that have just been described in section 4: the slope of the UV spectra of the quasars (subsection 5.1) and the cubic spline interpolation (subsection 5.2).

5.1. Slope of the UV spectra of the quasars

I will analyze the data thoroughly in this section. After I obtain the estimations of the slopes of the 'template' and the 'mock', I will try to check whether classifying QSO by slopes is plausible with the data, I will study some properties of the data (in particular, the dependence of the slopes with the rest of the variables, such as magnitudes and redshift), and most importantly, I will check whether the regressions are good, and whether the continuum component from 'template' can really be reconstructed from the 'mock' data. Unless specified otherwise, the linear regression (26) will be used. To visualize some of the casuistry, check the appendix B.

For the uncertainties and, in general, the presentation of the results, I follow the conventions that are taught in the Degree of Physics, University of Seville.

5.1.1. Classification by slopes

Once I ran the linear fitting for the 3196 QSO of the WEAVE-QSO OpR3b dataset for the couples $(\lambda, F(\lambda))$ in the wavelength region (25), for the 'template', 'model' and 'mock', the first thing I did was check the distribution of slopes of the quasars (admitting that the real slope is the obtained from 'template'), to see if it followed the expected tendency (laying around -1 and -2) and if more than one different population (that is, group of QSO with similar slopes that are somewhat different from those of other groups) could be clearly distinguished or not. In order to do so, I represented a histogram of the vector of 'template' slopes, α_{temp}, only to check that it roughly follows a normal distribution. I tried, then, to adjust a Gaussian,

$$y = Ae^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{37}$$

to it, which is represented in figure 14.

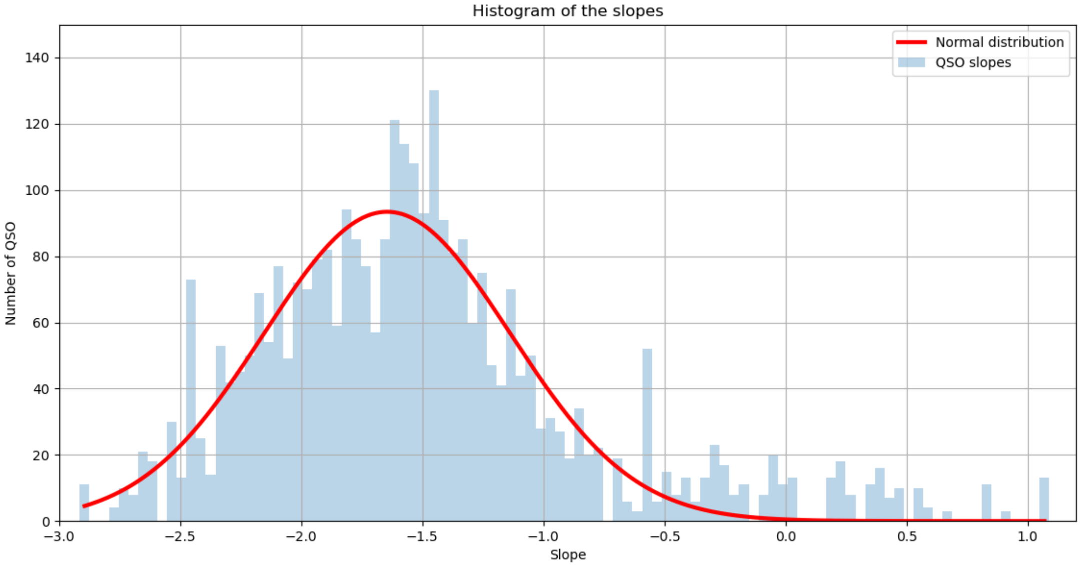


Figure 14: In blue, the histogram (in 100 bins) of the QSO ‘template’ slopes obtained from the data as explained in the previous subsection 4.2. In red, Gaussian fit to this histogram.

Mean slope	**Standard deviation**	**A**	μ	σ	r^2
-1.501	0.689	93 ± 3	-1.645 ± 0.022	0.508 ± 0.022	0.8

Table 2: This table includes the mean slope and the standard deviation of the 3196 QSO slopes (directly calculated from the data), as well as the results of the Gaussian fit.

Obviously, this distribution will be dependent on the region chosen for the line regression. Even if 3196 is a big number of quasars, it is not enough to define a smooth Gaussian (or to show that the slopes do not follow such a law). In any case, the median is $-1{,}576$, which is smaller than the average. Such asymmetry can already be perceived in figure (14): some QSO present positive slope, disrupting the symmetry, and constituting a dilemma themselves: slopes are expected to be negative. It can be checked that the reason for these positive slopes is that the QSO spectrum strongly differ from the power-law approximation that others do follow to some extent. See figures from appendix B for an example of this.

On the other hand, should one be able to recognise different populations of QSO by their slope, a sum of more than one normal distribution would produce a better fit. I tried from two to five, and only the bimodal (as the sum of two normal distributions) produced a physically plausible result that in fact adjusted the curve better.

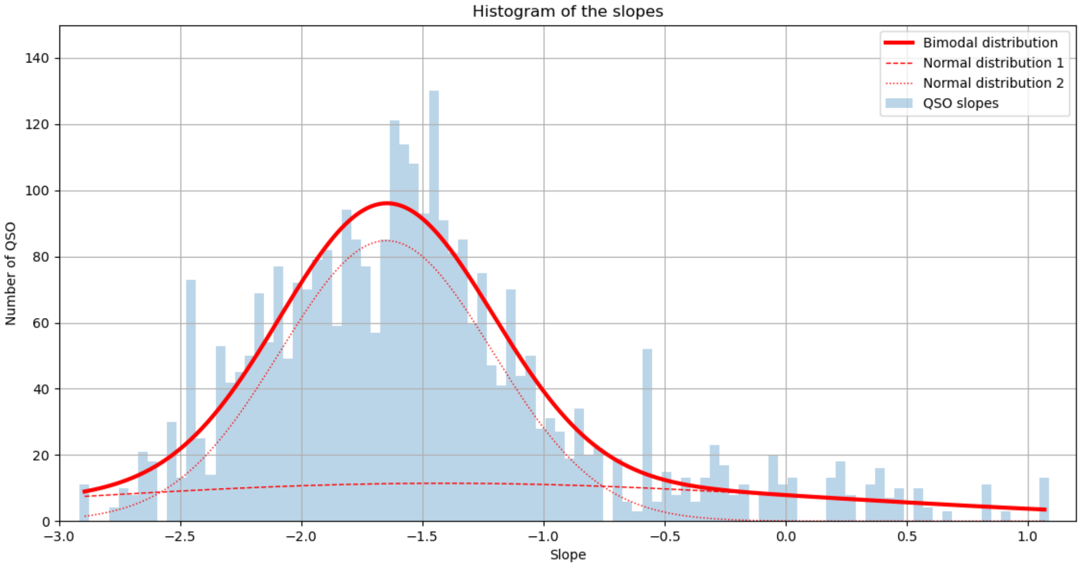


Figure 15: In blue, the histogram of the QSO slopes. In red, the bimodal fit to this histogram. The two Gaussian are represented in dotted lines.

A_1	μ_1	σ_1	A_2	μ_2	σ_2	r^2
85 ± 7	-1.65 ± 0.03	0.44 ± 0.04	11 ± 7	-1.4 ± 0.7	1.6 ± 1.0	0.9

Table 3: Fitting the sum of two normal distributions, I get the results showed. The r^2 is slightly bigger, conveying this may be closer to the ideal distribution of slopes.

In any case, the overlapping between the two normal curves suggest that they do not correspond to two different populations of QSO, but rather to the same population that may not be exactly normally distributed (either that or that the secondary populations do not have enough representation to be clearly distinguished from the obvious −1.5-centered one; the errors of the parameters of the second Gaussian are substantially higher than those of the first Gaussian, supporting this idea). Furthermore, it should be noted that this distribution doesn't take into account the regression errors, which could change slightly the distribution.

5.1.2. Dependence of the slope with magnitude and redshift

In this subsection, I show the results of the statistical study of the correlation between r-band apparent magnitude and slope, and between redshift and slope. I refer to appendix A at the end of the work for the complete study. The results simply indicate that there is no correlation whatsoever between slope and redshift or magnitude. This is the expected result, however, for the following reasons:

- Redshift represent both distance from us and time since the light was emitted or, equivalently, time at which the light left the QSO measured since the Big Bang. As the QSO lifespans are very small, redshift is not related to the QSO age, which may have an impact on the slope. This way, assuming space-temporal homogeneity, there is no reason why redshift should affect the QSO's physical properties and therefore its slope.

- Apparent magnitude represents the QSO's integral flux. This depends on their distance from us and on the amount of luminosity actually emitted. As I argued, no dependence with distance is expected. The only dependence the slope could have with magnitude is through luminosity, but at least in the data such dependence is not apparent.

This is an important result because it indicates that magnitude or redshift are not needed to be taken into account when stacking quasars, and that no method for the continuum reconstruction based on the slope of quasars could use solely the magnitude and redshift of the quasars.

5.1.3. Continuum reconstruction

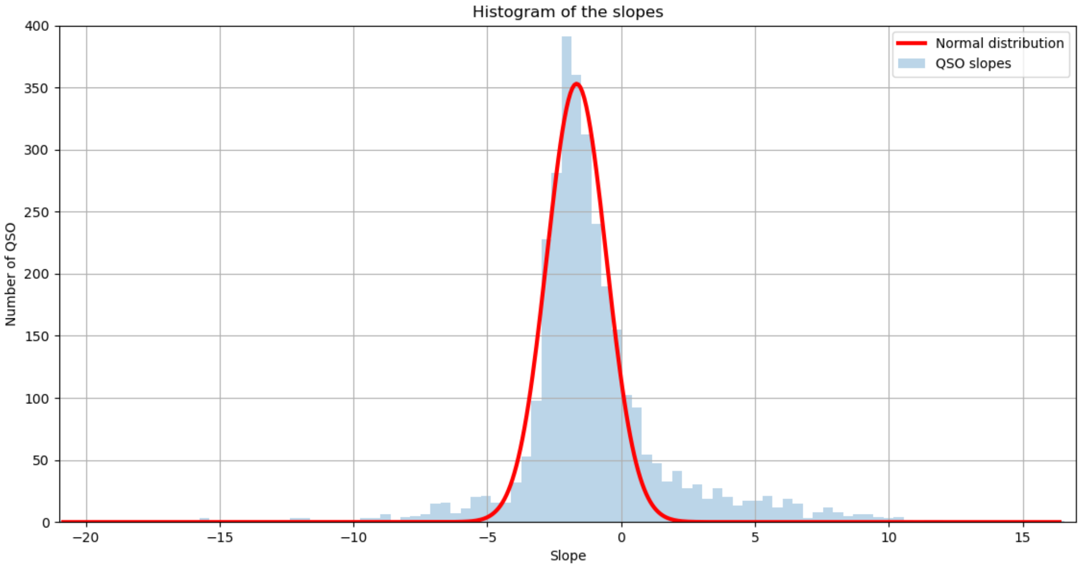


Figure 16: Again, the slopes roughly follow a normal distribution. The Gaussian fit is included along with the histogram of the 'mock' slopes.

In this section, the 'mock' slope distribution will be studied, and these slopes will be correlated to the 'template' slopes. Plotting the histogram of 'mock' slopes yields the result of figure 16. Again, a Gaussian of the form $y = Ae^{-\frac{(x-\mu)^2}{2\sigma^2}}$ may be adjusted, with the following results.

Mean slope	**Standard deviation**	**A**	μ	σ	r^2
-1.077	2,968	353 ± 8	-1.67 ± 0.03	1.10 ± 0.03	0.96

Table 4: This table includes the mean slope and the standard deviation of the 3196 QSO "mock' slopes, as well as the results of the Gaussian fit.

The parameters do differ: mean slope and μ parameter of the Gaussian do not coincide, in fact the distribution seems more skewed to the right than before (the mean value is as high as -1.077). That may be due to the information lost by cutting the non-positive fluxes when taking logarithms. In fact, I checked the distribution of the slopes using directly the power law for the fitting, and the mean was -1.336, closer to the -1.501 from the 'template'.

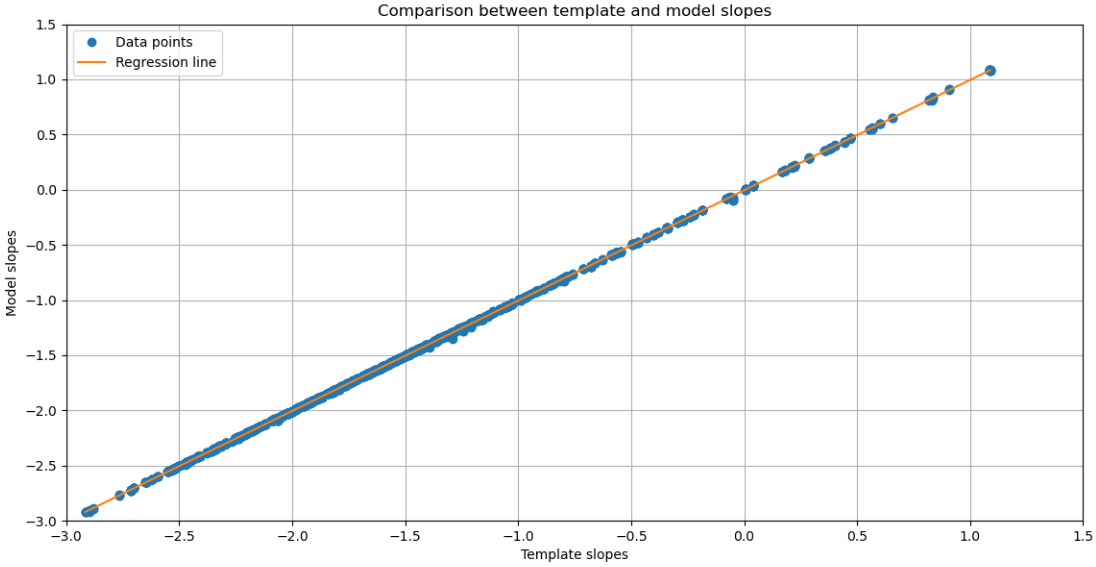
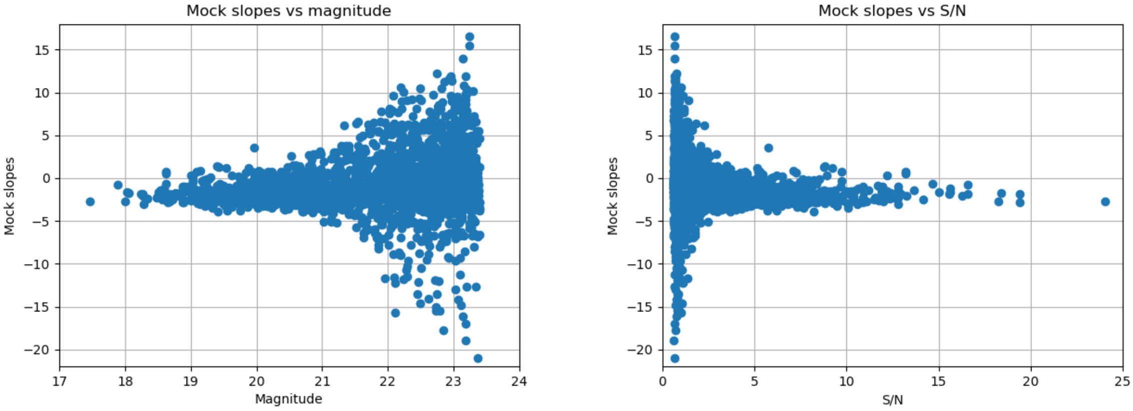


Figure 17: Comparison between the slopes of the 'template' and the 'model'. The equation of the regression line is $y = (0.99975 \pm 0.00012) \cdot x + (-0.00503 \pm 0.00020)$, with $r = 0.99998$,

It also is observed in figure 16 that values extend from -21 to 17, which is obviously has no physical sense. By visualizing examples, one notices that the enormous oscillations of some QSO 'mock' spectra cause some slopes to differ greatly from the real value, which in some cases may not be clear at all in them. Let me remind that these oscillations are mainly due to the atmospheric and instrumental effects and their subsequent corrections, which are inevitable for the current Earthly instruments. This hypothesis can be proven by computing the slopes to the ‘model’ variable, which is the same as the ‘mock’ save these corrections (that is, it roughly coincides with the ‘template’, save for the cosmological absorptions). By doing so, slopes from ‘template’ and ‘model’ may be compared, obtaining figure 17.

Therefore, the method for obtaining the slopes is correct, as it does avoid the absorption regions as was intended, and the strong variation has been proven to be due to the noisy, messy spectra of the QSO (see figures 5). In fact, a different check ultimately supports this:



(a) Slope distribution against magnitude.

(b) Slope distribution against S/N.

Figure 18: Each point represents one QSO. Clearly, higher magnitude and lower signal to noise ratio S/N (which are correlated) imply more dispersion and considerable less reliable ‘mock’ slope values.

The less luminous (higher magnitude) produce less reliable results, which cannot be helped, as it is directly related to lower S/N and the regression, whatever the region chosen, yields worse results. In fact, the dependence of the dispersion with S/N is stronger, and therefore the dependence of the dispersion with the magnitude is simply a consequence of this and the fact that magnitude and S/N are correlated.

In what follows I state the main results of this subsection.

In figure 19, the ‘template’ slopes were divided in 20 bins of the same size (the horizontal error bars of the blue data points are as big as this bin size), and the median of the ‘mock’ slopes corresponding to the quasars laying in each bin is computed; this corresponds to the blue data. Its vertical error bars are the 68^{th} percentile of the vector of distances between this median and the rest of the data (one σ). In red, the line representing the theoretical dependence between these two variables, $y = x$. In purple, the regression line of the blue data, which slightly differs from this dependence (note that the red line is graphically compatible with the blue set of points). The points that laid outside of the error bars were included in two different colors: green for those with $S/N > 3$ (a total of 60 points) and orange for those with $S/N \leq 3$ (a total of 968). Note how most of the high S/N points lay either inside or very close to the vertical error bars. In fact, this is (another) visual proof that the lower S/N QSO are definitely the cause of the huge dispersion in the 'mock' slope distribution. The results of the regression with the blue data points is shown in table 5.

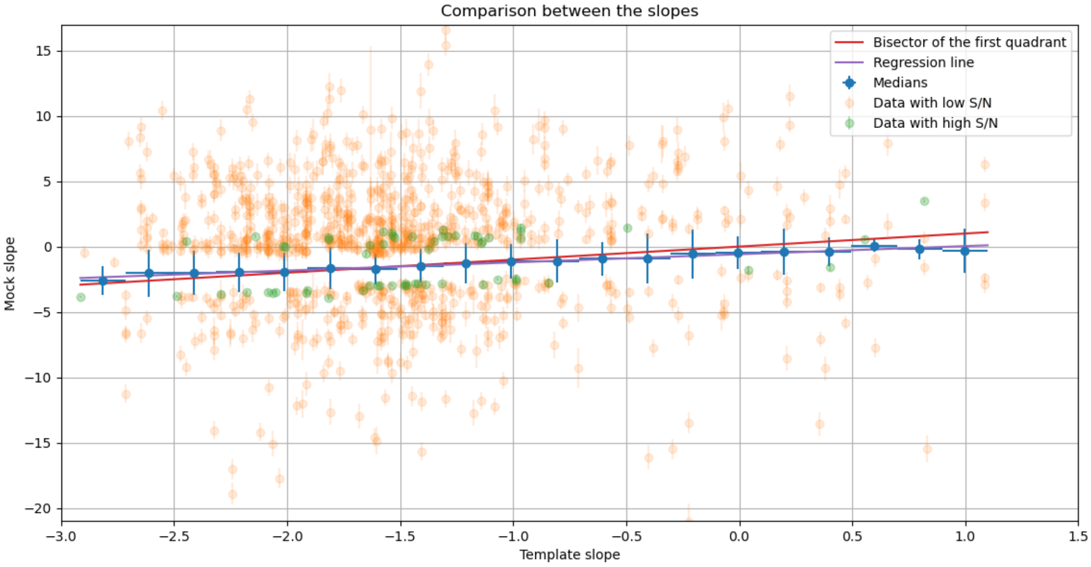


Figure 19: Results of the slope comparison. 'Mock' slopes are plotted against 'template' slopes, after binning the data points in 'template' slope. The data points that lay out of the error bars have been included to offer an overview of the great dispersion of the data: the orange represent quasars with $S/N \leq 3$ and the green represent quasars with $S/N > 3$. The line $y = x$ (bisector of the first quadrant) and the regression line of the binned data is also included.

Slope	**Intercept**	**r**
(0.62 ± 0.03)	(-0.59 ± 0.04)	0.98

Table 5: Slope, intercept and Pearson correlation coefficient of the linear regression of the binned pairs (α_{temp},α_{mock}).

The slope and the intercept do differ significantly from the expected values (1 and 0, respectively). The difference, however, is not very visually noticeable, and must be due to the effects of the atmosphere and the instrument and their correspondent corrections (the way they are carried out probably accounts for this difference), as such a difference in these two values is not present in the 'model' slope distribution (see figure 17). The most important fact is that the two variables are obviously correlated, and that, on average (or 'on medians', because, as commented before, both returned similar results), the method for obtaining the slope does work and give consistent results, although worse so in lower S/N quasars.

Nevertheless, in order to establish a proper procedure to be used when final data comes out (which will be the equivalent to the 'mock', exclusively), the opposite regression is more appropriate (the real slopes, equivalent to my 'template' slopes, will be sought, and the observed slopes, equivalent to the 'mock' slopes, will be the available data). However, as could be inferred from figure 19, the high S/N quasars could lead to values of slopes that have no physical meaning (those in orange), and obtaining the real slope of, say, a measured slope of 10 from a problematic QSO just is impossible.

If the number of QSO is small enough, of course, the slope can be determined in a one-by-one basis, obtaining optimal results. However, let me remind here that the objective is to obtain results for a number of quasars of the order of a million: one-by-one analysis is not an option. Therefore, in order to obtain a straightforward method for obtaining the slope of tons of quasars, I suggest utilizing the higher S/N quasars only, with a threshold depending on the reliability of the results sought. In particular, for a threshold of $S/N > 3$, the following results may be obtained:

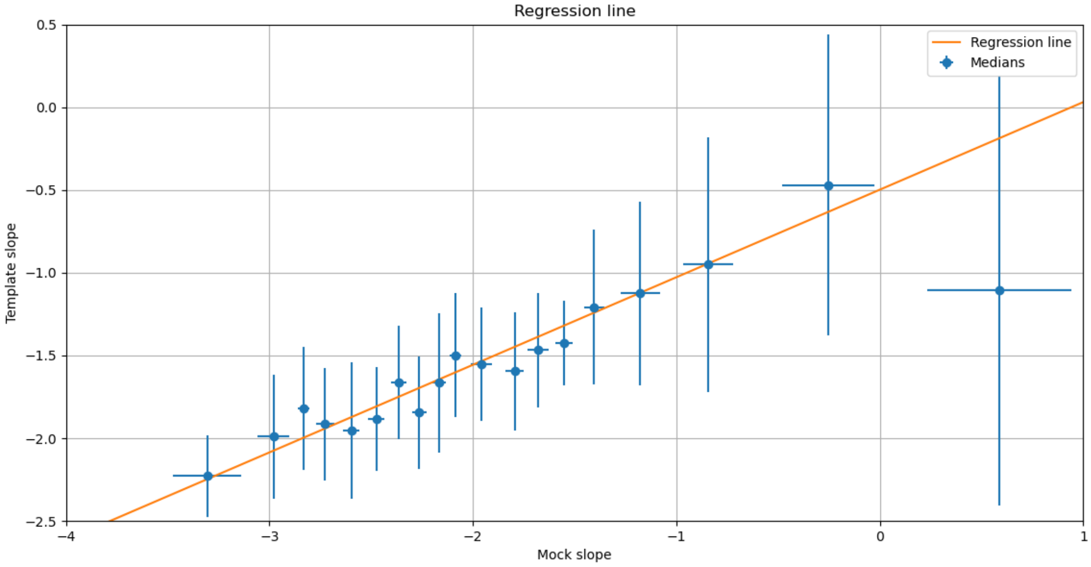


Figure 20: Regression line (orange) for the mock slope-binned $S/N > 3$ QSO couples (α_{mock},α_{temp}) (blue). The last point does not follow the general tendency, and therefore was excluded from the regression line.

Slope	Intercept	r
(0.53 ± 0.03)	(-0.50 ± 0.07)	0.98

Table 6: Slope, intercept and Pearson correlation coefficient of the linear regression of binned couples (α_{mock},α_{temp}).

In figure 20, the bins were selected, in this case, to have the same amount of data points

(19 bins of 43 data points each). The median and the 68^{th} percentile of the desviations from it were used as the (x,y) coordinates of the data points and their vertical and horizontal error bars. The regression of the table 6 was carried out excluding the last point (that is, the positive slopes, which do not follow the general tendency). Note that the slope and intercept are not 1 and 0, as a result of the atmospheric and instrumental noises and their subsequent corrections, but what is important is that the regression is good ($r = 0.98$) and predictions can be made with it. With these results I am able to conclude with the following.

Method: Once the real data comes out, in order to carry out a slope analysis of it, first, filter the QSO by a S/N threshold. The regression from figure 20 and table 6 used $S/N > 3$, being therefore only advisable for this threshold (if a different threshold needs to be used, I suggest running a similar analysis to mine with these mock data and the new threshold to calibrate the regression line again). Once this filter has been applied, run the linear regression of equation (26) on every QSO. Later, bin the QSO by slopes in the same fashion as the figure 20. Now, apply the map $\alpha_{mock} \rightarrow \alpha_{temp}$ from table 6 to get a more reliable value of the slope[16]. Although I suggest tossing the positive slopes before binning the data, in case a positive value is needed, simply extrapolate this line to cover it. Lastly, extrapolate the power-law obtained to the rest of the spectrum. The constant A from $F = A\lambda^{\alpha}$ may need an adjustment after changing the slope (simply minimize the distance between this extrapolated power law and the measured spectrum in the same points the regression took place using lemma 1, for example). In any case, this procedure will only bring good results if enough quasars are used and the results averaged to eliminate the statistical error that is being committed with it.

5.2. Cubic spline interpolation

This section is structured as follows. Firstly, I will present the results for all the combinations of residuals, regions and constants defined on section 4.3. Secondly, selecting the best choice of these three for the data based on certain criteria, I will propose a method to be used for the real WEAVE-QSO Survey. Lastly, the method will be checked stacking the punctual residuals of all $z > 2.7$ QSO and (separately) those with $S/N > 3$ within these.

5.2.1. General results

As I mentioned above, in the methodology section (subsection 4.3), I ran the code for 3 different residuals, both with and without a moving ‘renormalization’ constant,

[16] Of course, the α_{mock} values could be used directly too, but, as I checked, this slope is not the real slope for most cases due to the data treatment (noises and corrections) prior to the spectra analyses.

conveniently computed in several different ways as well. The amount of results to present is, then, very wide, and thus I will try to summarize them in table 7.

The optimal parameters did not show any dependence on the value of the redshift 'z', as expected: only magnitude or S/N may change the smoothness needed (the messier the data, the more smoothness is needed, implying larger bin sizes), redshift may only stretch the shape of the spectra, but does not convey higher noise. Therefore, the dependence was only studied between the optimal parameters and magnitude. Both magnitude and S/N are easily computable for any spectrum, and follow the relation given in table 1, therefore composing the line equations given in table 7 with the relation from table 1 yields exponential relations between the optimal parameter in question and the S/N value. Of course, a regression can be performed directly between the optimal parameters and the S/N. However, one thing must be kept in mind: the roughly linear behaviour between optimal parameters and magnitude does not apply to magnitudes below approximately $19 - 20$ (the exact limit depends on the case), where saturation can be observed when representing all data points. Nevertheless, only 43 quasars from the WEAVE-QSO OpR3b dataset have magnitude lower than 19: quasars so luminous are not abundant. In the case of S/N, the shape of the curves obtained representing the optimal parameters against it somewhat resembles a very steep negative exponential, or a negative-sloped line followed by a horizontal line of value oscillating between 200 and 100 (there are not many quasars with high S/N to define it clearly). The choice of using magnitudes instead of S/N is based on the facts that the linear tendency was clearer with them and that the quasars were distributed more sparsely over a broader range in magnitude than in S/N, which makes the regression more reliable.

Note that the intercepts are very prone to high uncertainties. This is due to the fact that the magnitudes of these QSO range between 17 and 24, therefore m=0 is far from the data points and the theoretical Δpix for a hypothetical m=0 (for the intercept) is not a reliable value. Obviously, the optimal parameters obtained from these regression lines will not be integer: the value must be rounded to be usable.

The regression lines were computed between with magnitude-binned data (bins of equal number of data points)[17]. The number of bins was chosen to be a number close to 10 which divides the value in the column length in each case (to use all quasars). It was checked that computing the regression lines without binning (raw data points) returned similar results (with, however, considerably lower - in absolute value - r values), but the binning was performed to obtain the statistical tendency, avoiding possible statistical deviations from it.

[17] Similarly to section 5.1, the medians were used.

Residual	**Region (parameter)**	**Constant**	**N**	**Slope**	**Intercept**	r
$resid_1$	R_{total} (Δpix_1)	None	2397	27 ± 17	$(-4 \pm 4) \cdot 10^2$	0.4
		C_1	321	79 ± 25	$(-16 \pm 6) \cdot 10^2$	0.8
		C_2	2394	78 ± 16	$(-15 \pm 4) \cdot 10^2$	0.8
	R_{total} (Δpix_2)	None	2397	15 ± 3	$(-32 \pm 6) \cdot 10^2$	0.8
		C_1	321	$(19 \pm 5) \cdot 10$	$(-41 \pm 10) \cdot 10^2$	0.9
		C_2	2394	145 ± 17	$(-29 \pm 4) \cdot 10^2$	0.93
	R_1 (Δpix_1)	None	2937	$(16 \pm 4) \cdot 10$	$(-32 \pm 9) \cdot 10^2$	0.7
		C_1	223	35 ± 25	$(-6 \pm 6) \cdot 10^2$	0.6
		C_2	2181	95 ± 14	$(-18 \pm 3) \cdot 10^2$	0.92
	R_2 (Δpix_2)	None	2397	$(20 \pm 3) \cdot 10$	$(-42 \pm 6) \cdot 10^2$	0.9
		C_1	912	$(21 \pm 4) \cdot 10$	$(-44 \pm 9) \cdot 10^2$	0.91
		C_2	2394	148 ± 11	$(-288 \pm 25) \cdot 10$	0.97
$resid_2$	R_{total} (Δpix_1)	None	2397	-3 ± 6	$(26 \pm 14) \cdot 10$	-0.1
		C_2	2392	73 ± 13	$(-14 \pm 3) \cdot 10^2$	0.9
	R_{total} (Δpix_2)	None	2397	122 ± 20	$(-25 \pm 4) \cdot 10^2$	0.8
		C_2	2392	137 ± 6	$(-28 \pm 5) \cdot 10^2$	0.9
	R_1 (Δpix_1)	None	2397	130 ± 23	$(-25 \pm 5) \cdot 10^2$	0.8
		C_2	2180	95 ± 14	$(-18 \pm 3) \cdot 10^3$	0.92
	R_2 (Δpix_2)	None	2397	164 ± 20	$(-34 \pm 4) \cdot 10^2$	0.91
		C_2	2393	156 ± 6	$(-31 \pm 3) \cdot 10^2$	0.96
$resid_3$	R_{total} (Δpix_1)	None	2397	$(136 \pm 18) \cdot 10$	$(-27 \pm 4) \cdot 10^2$	0.9
		C_3	2386	120 ± 17	$(-24 \pm 4) \cdot 10^2$	0.9
		C_2	2397	57 ± 8	$(-105 \pm 19) \cdot 10$	0.9
	R_{total} (Δpix_2)	None	2397	200 ± 25	$(-41 \pm 5) \cdot 10^2$	0.9
		C_3	2386	158 ± 11	$(-307 \pm 25) \cdot 10$	0.97
		C_2	2397	144 ± 10	$(-278 \pm 22) \cdot 10$	0.97
	R_1 (Δpix_1)	None	2397	$(18 \pm 4) \cdot 10$	$(-36 \pm 9) \cdot 10^2$	0.8
		C_3	2264	39 ± 13	$(-5 \pm 3) \cdot 10^2$	0.8
		C_2	2221	106 ± 16	$(-20 \pm 3) \cdot 10^2$	0.92
	R_2 (Δpix_2)	None	2397	$(19 \pm 3) \cdot 10$	$(-39 \pm 6) \cdot 10^2$	0.9
		C_3	2385	31 ± 9	$(-9 \pm 21) \cdot 10$	0.7
		C_2	2396	125 ± 11	$(-232 \pm 23) \cdot 10$	0.95

Table 7: The results of the regression lines between the relevant bin sizes and the magnitude are included (slope, intercept and Pearson correlation coefficient). The column 'N' indicates the number of quasars that satisfy both $z > 2.5$ and $0.5 < C < 1.5$, C being the renormalization constant, if applicable.

Note that the number of quasars (column 'N') for $R_{total}(\Delta pix_1)$ and $R_{total}(\Delta pix_2)$ coincide (it is the same analysis), that for the analyses without renormalizing this 'N' is precisely the total number of quasars with $z > 2.5$ (2397), and that the number of those whose renormalizaton constant was unphysical rarely surpasses 100 in any row. There are only two exceptions. The first one is the region R_1. It is a smaller and considerably more problematic region (noise and absorptions are both strong and can be hardly differentiable), thus worse results are expectable. The second is those cases where the constant C_1 is used. As it was computed solving the quadratic equation (4.3), in some cases it was not a real number, and in other it was, but also was unphysical. This shows that this way of ‘equalizing’ absorptions is not entirely correct. In fact, I may remind here that an absorption produced near the QSO will stretch along with the rest of the spectra as light travels towards us, and therefore will cover a bigger part of the spectrum than an absorption produced near us (which, then, has not stretched at all). Thus, giving all absorption features the same ‘weight’ is very imprecise. The definition of more accurate weights is not an easy task, and is beyond the scope of this work.

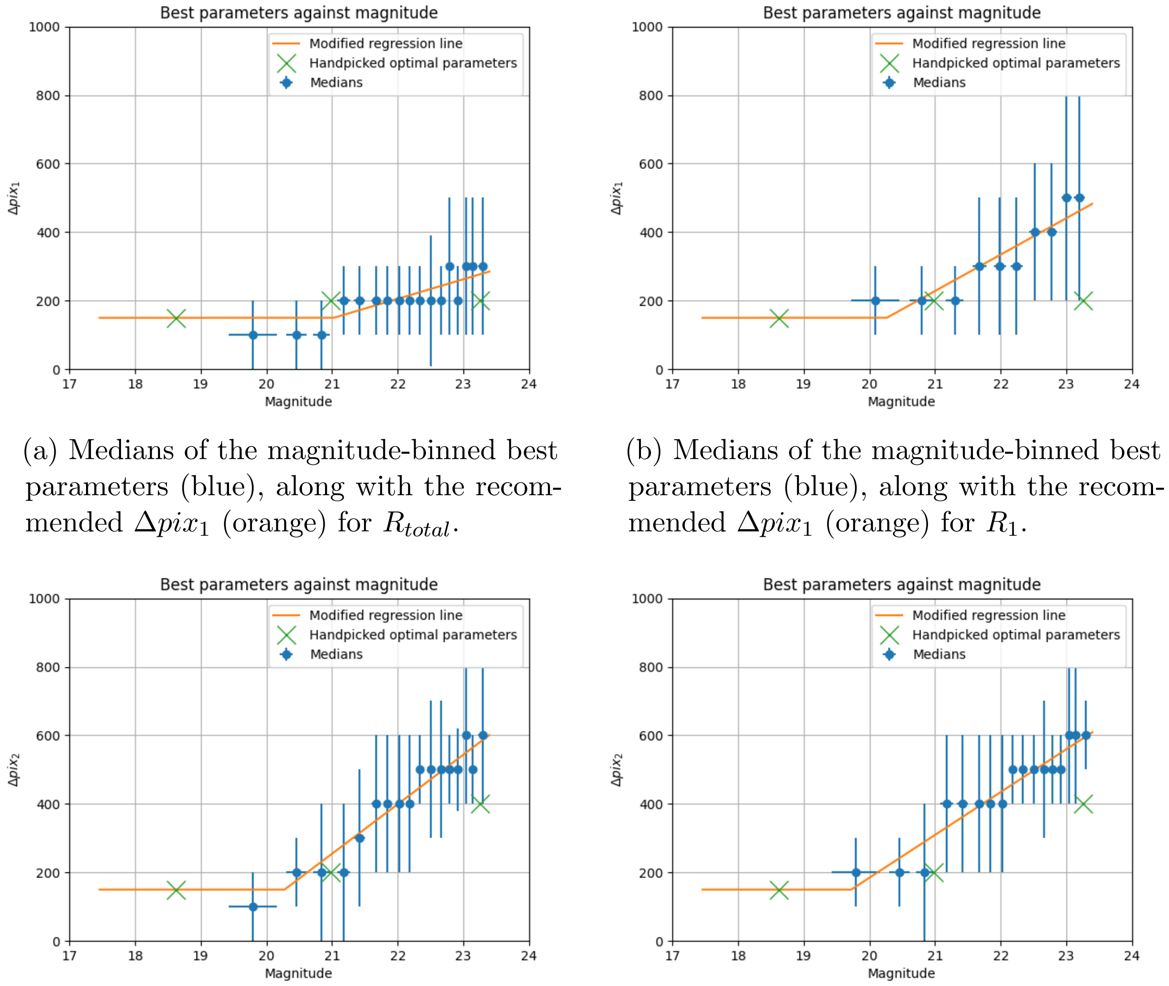
Another important remark that was visually checked for a sample of QSO is that, fixing one of the Δpix and changing the value of the other, the region corresponding to the fixed value did not show important changes. This is due to the fact that, mathematically, in that region, the same linear system is being solved to compute the splines save for the equations corresponding to the point of union, and therefore the change in the matrix of the linear system is small, implying that the change in the solution (computed splines in that region) is also small. This justifies the fact that, in R_{total}, the regressions for Δpix_1 and Δpix_2 are computed independently.

Lastly, if no renormalization constant is used, the optimal parameters have greater tendency to be the extreme values (10 and 1000, which, as I just argued, are not desirable) and also have greater residuals. Therefore, renormalization is strongly recommended.

All of the results in 7 are included for the interested reader. I recommend only using the fits where a renormalization constant is used, in particular those using C_2 seem to produce both the best results (visually, and also note that this constant’s values are the most physical in most cases) and the most reliable fittings. I will also make this remark: The lines with worse r values were not clear lines, and I suggest that they should be replaced with horizontal lines of some mean value (for example, ($resid_1$, R_{total}, None) with $\Delta pix_1 \approx 200$ and ($resid_3$, R_2, C_3) with $\Delta pix_2 \approx 500$).

In any case, a detailed analysis of the best residual to use (which probably depends on the spectra and the scientific objectives of the interpolation) is not the main goal of this section, butit is rather to give an applicable method to real data of the WEAVE-QSO Survey, likely coming out on late 2022. For that purpose, in the next subsections, I will only use the third residual ($resid_3$), for the following reasons. It penalizes the emission lines in favor of the continuum component, which is the most interesting to study absorption, especially the Lyman-α forest. Furthermore, the residuals have more physical meaning. It can be checked that the residuals from $resid_1$ and $resid_2$ follow increasing tendencies with S/N, therefore decreasing as the magnitude increases. This is simply due to the fact that the most luminous quasars have bigger fluxes, and these residuals represent absolute errors, so even though the relative errors may decrease the more definite the QSO spectrum is (higher S/N), the absolute errors increase. However, the residuals from $resid_3$ show the opposite tendency, as will be presented in subsection D. It is more assumable that most luminous (less noisy) quasars produce less residuals.

5.2.2. Proposed methodology to analyze the WEAVE-QSO dataset



(a) Medians of the magnitude-binned best parameters (blue), along with the recommended Δpix_1 (orange) for R_{total}.

(b) Medians of the magnitude-binned best parameters (blue), along with the recommended Δpix_1 (orange) for R_1.

(c) Medians of the magnitude-binned best parameters (blue), along with the recommended Δpix_2 (orange) for R_{total}.

(d) Medians of the magnitude-binned best parameters (blue), along with the recommended Δpix_2 (orange) for R_2.

Figure 21: The results of the four relevant modified regression lines between optimal parameters and magnitude, along with the original binned best parameters, is included. The magnitude is, as in the rest of the work, the r-band apparent AB magnitude.

When time comes to use the results of these analyses to the real WEAVE-QSO data, I suggest using $resid_3$ (better for continuum, worse for emission lines, as discussed in 5.2.1; this decision is supported by works like Meyer et al., 2019). Furthermore, the renormalization constant is mostly used in real-life applications for the Lyman-α forest region (that is, R_1), because it is the splines what is renormalized (remember, no such thing as 'the template', a theoretical continuum, is known in these cases, in fact the purpose is estimating it) and is mostly needed where global absorption may be present (only in the Lyman-α forest). However, as I am doing a comparison with a theoretical 'template' that has a norm (subsection 4.1) prone to some error, even if small, the comparison is more exact renormalizing it irrespective of the region. The choice I recommend is to use C_2 (which, remember, was computed by minimizing the distance between F_{spl} and F_{temp} in the region being considered), as explained in subsection 5.2.1. Therefore, on the region R_{total} one regression for Δpix_1 and another for Δpix_2 is included below, while for R_1 (respectively R_2) only the regression for Δpix_1 (respectively Δpix_2) is included, as the other parameter is not important.

The mean value of the relative error by using this method is expected to be below 5 % irrespective of the region for magnitudes lower than 20, and up to $20 - 25$ % (with more dispersion) for the highest magnitudes of the data set. See appendix D for more details on this.

These four regressions were performed in the same way. The data is divided in magnitude bins of the same number of data points (the number of resulting data bins was chosen to be a value close to 10 that divides the number of usable quasars, from the column 'N' of 7). In each bin, the median in both magnitude and Δpix (1 or 2) was computed to form the blue dots of the figures, while the 68^{th} percentile of the deviations from this median were chosen as the error bars. The regression lines (orange) are forced to saturate at $\Delta pix = 150$ (they are 'modified' regression lines), to avoid errors as the one in figure 11 (see subsection 4.3 for the full explanation). As the highest values lay around 600, no concern for the error in figure 12 is needed. The green crosses are the three QSO for which I handpicked the optimal parameters, from figures 13, 30 and 31.

Note that the regressions in figures 21c and 21d are similar, but the one in figure 21a is not similar to the one in figure 21b. To understand what is happening, I recall that in figure 21a the regression is done with best fits after minimizing $resid_3$ and adjusting C_2 in the whole spectrum, R_{total}. As $R_{total} \setminus R_1$ is a considerably bigger set than R_1, or in other words, there are many more points redwards than bluewards of Lyman-α, it is no surprise that the global residual value is dominated by the residual in the R_2 region.

The slopes and intercepts of the regression lines are given in table 7, along with their uncertainties. Taking into account all of this, I may conclude the subsection with the following:

Method: Given a QSO whose spectrum has meen measured, compute its r-band AB magnitude. Use the line equations from 7 (corresponding to $resid_3$, C_2 and the region and parameter of interest) to obtain the optimal parameters. If the obtained value is less than

150, substitute it by 150. If the region is R_1 (respectively R_2), then set Δpix_2 (respectively Δpix_1) to 200. Use the algorithm explained in subsection 4.3, for instance through the spline interpolation code from Dall'Aglio et al. (2008), with these parameters (and the other four, whose value was specified in subsection 4.3, minpix=4, slopethresh= 0.033, fluxthresh= 0.99 and fluxscale= 1.0). Specially if the region is R_1, adjust a constant to the resulting splines, using the most convenient method depending on the situation (one possibility is minimizing the distance to the measured spectrum in absorption-free regions, if the region is not R_1). These renormalized splines should, in average, give the optimal (best possible) results.

As a check of the extent of the applicability of this method, I performed a stack of more than half of the total number of quasars of the OpR3b data set, presented in the following subsection 5.2.3.

5.2.3. Stacks for $z > 2.7$ quasars

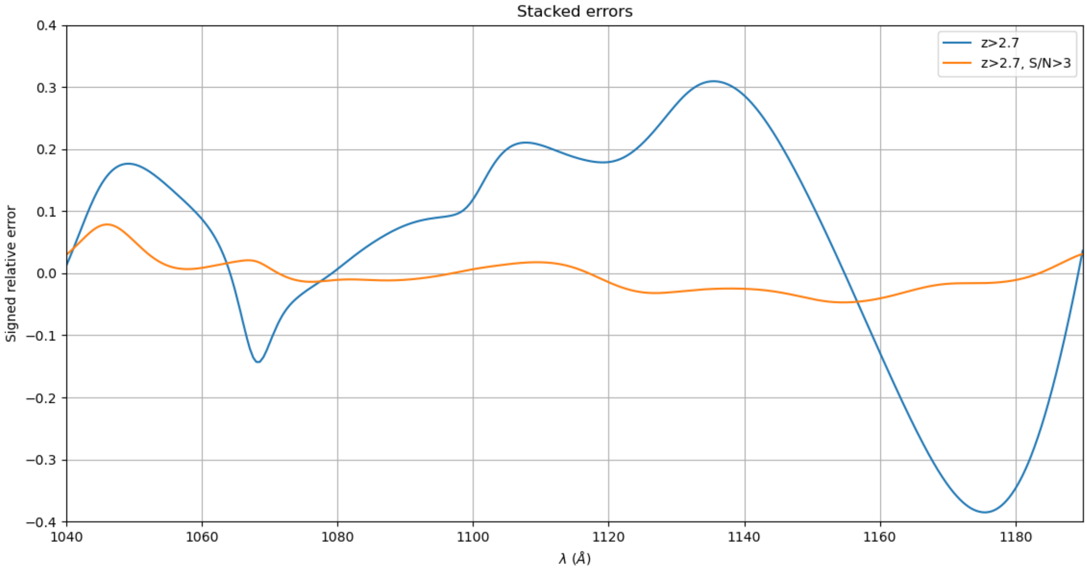


Figure 22: Stacks in the Lyman-α forest region (in the QSO rest frame, it corresponds to 1040 Å $< \lambda <$ 1190 Å). In blue, the stack for the 1665 $z > 2.7$ quasars. In orange, the stack for the 190 $z > 2.7$ and $S/N > 3$ quasars. The greater goodness of the fits for higher S/N quasars is apparent, as the stacked errors are considerably smaller in spite of having stacked approximately 10 times as less quasars. The mean values of the blue and orange curve are 4 % and -0.5 %, and the mean of the absolute values of the blue and orange curve are 16 % and 2.2 %.

In this subsection, I will use the regression line from table 7 corresponding to $resid_3$, R_1 and C_2 to compute the optimal spline interpolation for the spectra of the quasars with $z > 2.7$. The reason for this threshold is that, in order to be able to stack in the whole interval [1040 Å, 1190 Å] (in the QSO rest frame), the redshifted 1040 Å must be measured with the telescope, that is, $1040 \cdot (1+z) > 3676$[18], so $z \gtrsim 2.53$, but the closer wavelengths to 3676 Å are prone to higher errors due to the drop in the instrument's precision near its observable limits (see figure 5), and so to avoid them I chose $z > 2.7$.

Once all the splines are computed, for each 1040 Å $< \lambda^{rf}_{temp} <$ 1190 Å the following average is what is called the stack of the signed relative errors of the spline interpolation:

$$\frac{1}{N} \sum_{j \in QSO} \left(\frac{F_{spl} - F_{temp}}{F_{temp}} \right)_j \tag{38}$$

N is the number of QSO (i.e. terms in the summatory). Note that, if errors are purely statistical, by stacking enough quasars they should average to 0, or close, for any wavelength, that is, the stack is expected to be close to 0. The result is shown in figure 22.

The blue line represents the stack of all > 2.7 quasars, 88.6 % of which are low S/N quasars ($S/N \leq 3$). A certain bias can be seen in the right half of the graph, seemingly implying that splines tend to give high negative errors near the Lyman-α line. However, the orange line, that only stacks the $N = 190$ quasars with $S/N > 3$ (that is, the 'best' spectra in terms of S/N), although does not directly discard this hypothesis (the errors on the right are indeed negative), seem to blame the almost 40 % relative error value on the low S/N of the quasars rather than the method itself. In fact, as the relative errors of the orange line do not even reach 10 % after stacking only 190 quasars, the errors seem to be indeed statistical, and thus it is possible to eliminate them with stacking, expecting better results as the number of quasars increases. Therefore, if averaging 1665 quasars did not lower the stacked errors below 10 %, the complete WEAVE-QSO Survey, with a number of quasars with $z > 2.7$ of the order of 200,000, would probably accomplish it.

As a final conclusion, what is certain is that the $S/N > 3$ cut lowers the errors below 10 % on the Lyman-α forest region, which is the most problematic region observationally but is also the most scientifically interesting. Depending on the precision needed a different cut at perhaps higher S/N is recommended.

6. Conclusion

In this work, I defined magnitude (in particular, r-band AB apparent magnitude) as a measure of the luminosity of celestial objects, and redshift as a measure of the difference between emmited and received wavelengths. In particular, in quasars, black holes in the center of very distant galaxies surrounded by very hot material that fall spirally towards it emitting an enormous amount of light, I showed that this redshift is cosmological, i.e.,

[18] This was the lower limit for the telescope's range of wavelengths, check 3.4.

it is due to the cosmological hypothesis that the universe is expanding. I also defined the continuum component of a QSO spectrum, arguing that in the UV part it roughly follows a power law, $F = A\lambda^{\alpha}$. Rich information about the distribution of matter between the QSO and us can be extracted from absorption features if this continuum is known, and in this work I presented two different approaches to its reconstruction from observational data: computation of the slope (α) of the UV part of the spectra and cubic spline interpolation. To show the extent of the applicability of these methods, I used the OpR3b WEAVE-QSO mock data catalog (see subsection 3.4), consisting in 3196 mock quasar observed spectra along with their 'theoretical' continuum ('template'), which allowed me to find the best way to apply these two methods to the real WEAVE-QSO data expected to come out on September 2022.

To apply these two methods to the OpR3b mock dataset, I wrote the analysis pipeline as several thousand lines of Python code (in order to do that, the first part of the work consisted in learning Python in the first place). These codes presented the problem of high running times. In particular, the slope computation for all quasars lasted around 3 hours, while the splines analysis lasted up to 3 days. Therefore, a first conclusion is that fast computers will be required to treat the WEAVE-QSO Survey, when available, and to perform more sophisticated analyses than the ones I present.

Regarding the slopes, firstly in subsection 4.2, I argued that applying linear regression to ($ln(\lambda)$,$ln(F)$) in the region (1350 Å,1370 Å) $\cup$ (1440 Å, 1470 Å) (filtering $S/N > 0.1$ for the observed or 'mock' spectra, and after correcting the redshift, that is, in the QSO rest frame) was the best choice for the computation of the slope of a given spectrum, as it excludes the less reliable points with the condition $S/N > 0.1$ and the emission and absorption lines that would distort the result of the regression, and also minimizes the error of the extrapolation to the Lyman-α forest region, (1040 Å, 1190 Å), given its proximity. In order to test the slope approach, the slopes were computed, for every QSO, for both the observed ('mock') spectrum and the theoretical ('template') spectrum. The results indicate that the slope computation approach is not recommendable for low S/N quasars (in particular, I suggest $S/N > 3$).

Secondly, the slopes of the 'template' were studied. The power law proved to adequately fit most quasars, while others seemed to differ from such a behaviour (around 5 %). The cause is not in the method, but rather that these QSO showed exotic profiles. More research on how to detect the quasars that significantly differ from this expected behaviour could be useful for future analyses of this kind. The distribution of slopes showed a single population of quasars with normally distributed slopes centered in $\alpha = -1.5$. Furthermore, any correlation between the value of these slopes and the redshift or the magnitude of the quasars was statistically discarded with the discussion made in subsection 5.1.2. Therefore, no ‘check’ for the slope value can be based solely on these physical quantities, but no caution has to be taken about them when grouping quasars to study their slopes, either.

Thirdly, the slopes of the ‘mock’ (both ‘mock’ and ‘template’ slopes are computed with the same method and region, save the $S/N > 0.1$ condition for the ‘mock’) followed a similar distribution, merely with more dispersion, due to the noise and the oscillations of the data caused by atmospheric and instrumental traits and their subsequent corrections. By comparing the slopes of the ‘template’ and the ‘mock’ quasar by quasar, the usability of the method to make predictions was tested. In particular, I showed that the slopes of the ‘template’ spectra were in good agreement, on average, with the slope of the corresponding ‘mock’ spectra (more so with higher S/N quasars). This result, along with other checks that were presented in subsection 5.1.3, serve as proof that the region choice is satisfactory.

Finally, both slopes for each QSO did, in general, coincide, taking into account that certain bias was present mainly due to data treatment prior to these analyses (atmospheric and instrumental noises and their corresponding corrections). For real-life applications, I proposed a method to correct this bias, consisting on assigning, to the measured α_{mock}, a theoretical $\alpha_{temp} \neq \alpha_{mock}$, given by the regression line from 6 (being, by construction, closer to the real slope value, at least in the OpR3b data set). In order to do so, unphysical values of α_{mock} had to be discarded by filtering quasars by $S/N > 3$ (as I stated, I do not recommend using the slope computation method for low S/N quasars). Different cuts to quasars based on S/N can be applied, and in case another is needed instead of the presented $S/N > 3$, I suggest to perform the linear regression 6 again. In appendix B, along with other representations for the slope approach, there are two examples where this bias correction method produces better results than directly assigning $\alpha_{temp} = \alpha_{mock}$.

As further remarks for the slope computation analysis, some ideas for further development of the proposed method are the following. Similarly to in cubic spline interpolation, stacking and error estimation are future areas to continue the research I carried out throughout this work. In fact, works like Dall’Aglio et al. (2008) conclude that cubic spline interpolation produce more reliable results, and thus building a numerical test to mathematically tell how much better it is using the OpR3b data constitutes a matter of interest for future studies following this work.

Regarding the cubic splines, I used the algorithm described in Young et al. (1979) and Carswell et al. (1982). This algorithm uses certain parameters and variables, in particular the quasar spectrum, and basically bins it in wavelength, eliminates the points that are further away from the mean behaviour in the bin (dividing the bins near emission lines and removing null flux regions), computes the mean wavelength and flux of each resulting bin and returns the spline continuum reconstruction using these points (emission lines included), interpolating in the rest of the spectrum. The input variables are wavelength, flux, error of the flux and redshift, and the most important parameters are Δpix_1 and Δpix_2, which represent the size of the bins in which the spectrum is divided bluewards and redwards, respectively, of the Lyman-α line (1215.67 Å in the QSO rest frame), in order to filter the most adequate points from them, compute their mean wavelength and flux, and use these points to construct the cubic splines. My objective was to find a procedure to give the optimal Δpix_1 and Δpix_2 parameters based on the QSO characteristics.

To find this method, firstly, I selected some quasars with different S/N to determine what a good fit is using them. In order to do so, I visually analyzed hundreds of combinations of $(\Delta pix_1, \Delta pix_2)$ as they ran over a fine parameter space to select, by hand, the best fit. This approach, however, is not efficient for large number of quasars, and is definitely not feasible for the whole WEAVE-QSO dataset. In order to automatically extrapolate this to the rest of the quasars of the OpR3b data (only those with $z > 2.5$ to better sample the Lyman-α forest), I defined three residuals, with different objectives, that can be computed after interpolation using a couple $(\Delta pix_1, \Delta pix_2)$ on a spectrum has been carried out. The $(\Delta pix_1, \Delta pix_2)$ (both changing independently in a grid ranging from 10 to 1000) that gives the lowest value of the selected residual is the optimal couple. For the region on which the residual is computed, I also covered three possible choices: the whole spectrum, bluewards of Lyman-α and redwards of Lyman-α. Furthermore, the relative normalization between the 'template' and the computed splines was also adjusted (as a multiplicative constant close to 1 on 'template', in particular) to produce better fits. Depending on the usage, different choices for the residual, the region and the renormalization can be needed, and thus I included all the options I considered on table 5.2.1. No dependence between the optimal parameters and redshift was found, while a linear behaviour between the optimal parameters and the r-band apparent magnitude was observed, allowing me to construct the method based on a linear regression between the parameters and the magnitude.

Finally, I focused on the residual $resid_3$ (given that it gives more importance to the continuum rather than the emission lines) and the renormalization constant C_2 (defined on subsection 4.3), on all the three regions, and the results were presented in detail on subsection 5.2.2. The proposed algorithm to obtain the optimal parameters for any quasar works on average by definition, and this can be checked in subsection 5.2.3. In that subsection, a stacking was performed: the average value of the relative errors of the cubic spline interpolation using $resid_3$, C_2 and R_1 was represented, performing this average on all 1655 $z > 2.7$ quasars and the 190 $S/N > 3$ (best quasars in terms of signal-to-noise ratio) among those. The errors of the first stack went up to 30-40 % (with average of the absolute value 16 %), while those of the second stack, in spite of being computed with less quasars, did not reach the 10 % in any specific wavelength (its absolute values averaging 2.2 %). In WEAVE-QSO, around 200,000 $z > 2.7$ quasar spectra will be measured, and around 24,000 $z > 2.7$, $S/N > 3$ (based on the distribution on the OpR3b, and assuming it is representative). With more than 100 times as many quasars, the stacked error is expected to be considerably lower. Therefore, my results indicate that statistical density, temperature and distribution of matter between the quasars and the Earth will be determined with high precision, on average (the stacking procedure returns statistically average results).

In addition to the aforementioned analyses, further ideas arose that remained untried due to lack of time, but I am willing to explore these possibilities in the near future. Some of them are presented below.

The bins length being fixed at each side of the Lyman-α peak (unless line emission is detected) may become a problem for some QSO where the oscillations of the splines become prominent for some values of (Δpix_1,Δpix_2). In order to solve this problem, I suggest more adaptability on the bins sizes. In fact, working directly on the QSO rest frame (transforming between this and the observed is easily done following equations (6) and (20)) could allow to localize most of the possible emission lines, which could help define restrictions for the bin sizes near them in order to produce a better fit in them. One (more advanced) method to obtain better results in the spline analyses could be to add weights on the residuals. For example, instead of simply running the analyses and compute the residuals on the Lyman-α forest, it could grant better results to run it on the entire spectra, but with different weights on the Lyman-α forest, the Lyman-α line and the rest of the spectrum. In fact, this could become as sophisticated as one wants, because redwards of the Lyman-α forest more sub-regions could be considered (for example divided by the stronger carbon lines). In any case, and using the unweighed regions I used in this work, more possible values for the Δpix parameters' grid would define better their distribution (against magnitude as presented, for example). For that, however, either more efficient codes or access to faster computers would be needed.

Another typical method with growing interest is PCA or Principal Component Analysis (see Suzuki, 2005 and Pâris et al., 2011). It consists on, using a number of quasars (considered representative of the quasar population), diagonalizing the matrix of covariances of their spectra to obtain the principal components and replacing each of the spectra by the mean spectrum plus a linear combination of these principal components. The principal components (PCS) are orthonormal by construction (therefore attaining negative values at some points), and the first PCS (ordered by eigenvalue) hold physical meaning (for example, in Suzuki, 2005, the first PCS take the Lyman-α line, the Lyman-β line and other emission lines). Therefore, the coefficients of these PCS in a particular spectrum (when expressed as sum of the mean spectrum plus a linear combination of the PCS) shed light into the importance of these particular features in that particular QSO. One key advantage of this method is that, similarly to the UV slope method, the continuum in the Lyman-α forest can be reconstructed from information at higher wavelengths than the Lyman-α emission line. In the article Bosman et al. (2021), a comparison between the two methods I applied in this work and PCA is carried out. A similar comparison using the OpR3b data is definitely of interest for my near-future studies.

7. References

Bautista, J. E. et al. (May 2015). «Mock Quasar-Lyman-α forest data-sets for the SDSS-III Baryon Oscillation Spectroscopic Survey». *J. Cosmology Astropart. Phys.* 2015.5, 060, p. 060.

Benn, C. R. and S. L. Ellison (Nov. 1998). «Brightness of the night sky over La Palma». *New Astronomy Reviews* 42.6-8, pp. 503–507.

Bosman, S. E. I., D. Ďurovčíková, F. B. Davies, and A.-C. Eilers (Feb. 2021). «A comparison of quasar emission reconstruction techniques for z $\geq$ 5.0 Lyman-α and Lyman-β transmission». *Monthly Notices of the Royal Astronomical Society* 503.2, pp. 2077–2096.

Burbidge, G. and M. Burbidge (1967). *Quasi-Stellar Objects*. W. H. Freeman.

Carrol, B. W. and D. A. Ostile (2017). *An introduction to modern astrophysics*. Cambridge University Press.

Carswell, R. F., J. A. J. Whelan, M. G. Smith, A. Boksenberg, and D. Tytler (Jan. 1982). «Observations of the spectra of Q0122-380 and Q1101-264.» *Mon. Not. Royal Astro. Soc.* 198, pp. 91–110.

Condon, J. J. and A. M. Matthews (July 2018). «ΛCDM Cosmology for Astronomers». *Pub. of the Ast. Soc. of the Pac.* 130.989, p. 073001.

Dall'Aglio, A., L. Wisotzki, and G. Worseck (Nov. 2008). «An unbiased measurement of the UV background and its evolution via the proximity effect in quasar spectra». *AAP* 491.2, pp. 465–481.

Francis, P. J. et al. (June 1991). «A High Signal-to-Noise Ratio Composite Quasar Spectrum». *ApJ* 373, p. 465.

Garcia-Bellido, J. (Feb. 2005). «Cosmology and Astrophysics». *Proceedings of the CERN-JINR European School of High Energy Physics.*

Geller, R. M., R. A. Freedman, and W. J. K. III (2019). *Universe*. New York: MacMillan.

Goswami, G. K., A. K. Yadav, and M. Mishra (Jan. 2015). «ΛCDM-type cosmological model and observational constraints». *International Journal of Theoretical Physics* 54.1, pp. 315–325.

Harrison, C. (Sept. 2014). «Observational constraints on the influence of active galactic nuclei on the evolution of galaxies». PhD thesis. Durham University, UK.

Kembhavi, A. J. and J. V. Narlikar (1999). *Quasars and Active Galactic Nuclei, An Introduction*. Cambridge University Press.

Leipski, C. et al. (Apr. 2014). «Spectral Energy Distributions of QSOs at z ¿ 5: Common Active Galactic Nucleus-heated Dust and Occasionally Strong Star-formation». *ApJ* 785.2, 154, p. 154.

Lusso, E. et al. (June 2015). «The first ultraviolet quasar-stacked spectrum at $z \approx 2.4$ from WFC3». *Mon. Not. Royal Astro. Soc.* 449.4, pp. 4204–4220.

Meyer, R. A., S. E. I. Bosman, K. Kakiichi, and R. S. Ellis (Feb. 2019). «The role of galaxies and AGNs in reionizing the IGM - II. Metal-tracing the faint sources of reionization at $5 \lesssim z \lesssim 6$». *Mon. Not. Royal Astro. Soc.* 483.1, pp. 19–37.

Pâris, I. et al. (June 2011). «A principal component analysis of quasar UV spectra at z ˜3». *AAP* 530, A50, A50.

Pieri, M. et al. (Nov. 2016). «WEAVE-QSO: A Massive Intergalactic Medium Survey for the William Herschel Telescope». *Proceedings of the SF2A conference, Lyon.*

Riess, A. G., S. Casertano, W. Yuan, L. M. Macri, and D. Scolnic (May 2019). «Large Magellanic Cloud Cepheid Standards Provide a 1% Foundation for the Determination of the Hubble Constant and Stronger Evidence for Physics beyond ΛCDM». *ApJ* 876.1, 85, p. 85.

Suzuki, N. (Mar. 2005). «Quasar Spectrum Classification with PCA - II: Introduction of Five Classes, Artificial Quasar Spectrum, the Mean Flux Correction Factor dF,and the Identification of Emission Lines in the Ly alpha Forest». *arXiv e-prints*, astro-ph/0503248, astro–ph/0503248.

Tytler, D. et al. (Dec. 2004). «Cosmological Parameters $\sigma 8$, the Baryon Density Ωb, the Vacuum Energy Density $\Omega\Lambda$, the Hubble Constant and the UV Background Intensity from a Calibrated Measurement of H I Lyα Absorption at $z = 1.9$». *The Astrophysical Journal* 617.1, pp. 1–28.

Urry, C. M. and P. Padovani (Sept. 1995). «Unified Schemes for Radio-Loud Active Galactic Nuclei». *Pub. of the Ast. Soc. of the Pac.* 107, p. 803.

Vanden Berk, D. E. et al. (Aug. 2001). «Composite Quasar Spectra from the Sloan Digital Sky Survey». *AJ* 122.2, pp. 549–564.

Xiao, X., E. White, M. Hooten, and S. Durham (Oct. 2011). «On the use of log-transformation vs. nonlinear regression for analyzing biological power laws». *Ecology* 92, pp. 1887–94.

York, D. G. et al. (Sept. 2000). «The Sloan Digital Sky Survey: Technical Summary». *AJ* 120.3, pp. 1579–1587.

Young, P. J., W. L. W. Sargent, A. Boksenberg, R. F. Carswell, and J. A. J. Whelan (May 1979). «A high-resolution study of the absorption spectrum of PKS 2126-158.» *ApJ* 229, pp. 891–908.

A. Dependence of the slope with r-band apparent AB magnitude and redshift

In this section, I aim to solve the following questions: does the distribution of slopes depend on redshift or magnitude? In order to answer them, the scatter plots do not offer a clear answer, and 2D histograms are not the best idea either, as the quasars are not evenly distributed in neither redshift nor magnitude. The choice I finally made was to divide the redshift (respectively, the magnitude) in 100 equally-spaced bins, and represent the medians (percentile 50) of the slopes at each bin. The error bars added are the percentile 68 of the distances between the median slope and the rest of the slopes in each bin (as a generalization of the σ, standard deviation, of normally distributed random variables). Mean values and standard deviations produced graphically similar results, but are not used for the sake of consistency, because as the distribution of slopes (in each bin) is not symmetrical (slopes bigger than the mean or median reach farther than the smaller ones), these median and associated error bars are more statistically representative. In blue, the slopes, along with their regression errors are included as well. Note that their errors are very small and can hardly be noticed at all; in fact the bigger slope error is 0,09.

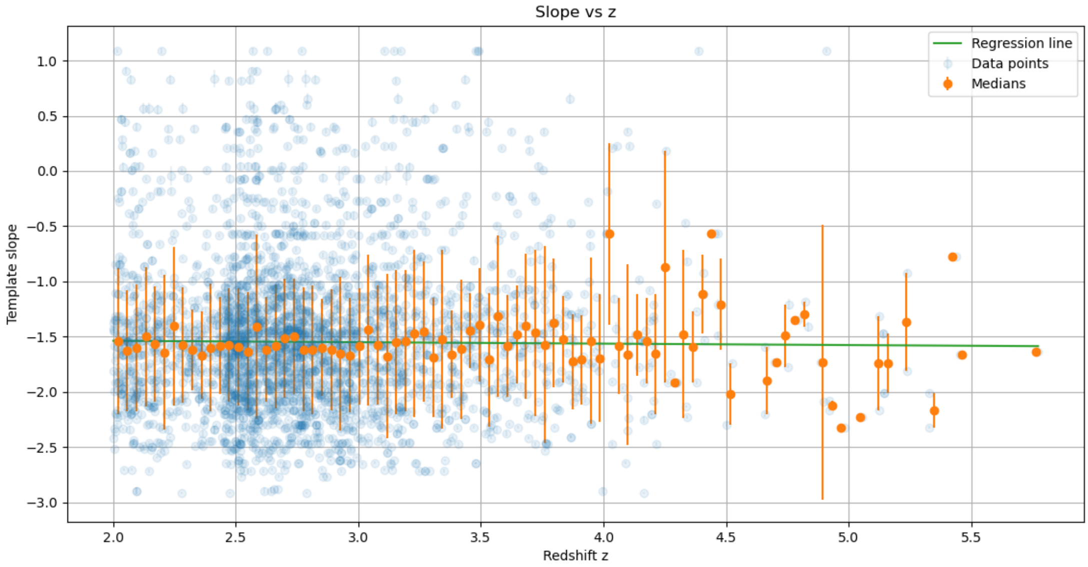


Figure 23: Slopes of the 'template' against redshift, z. The original slopes, in blue, their medians in uniformly distributed bins, in orange, and the regression line of these, in green. Note that some bins are empty (no median is shown in those) and some other, those at higher redshift, have very few (less than 5) points, so neither the median nor the mean really represent the real value that would be obtained with a higher number of cuasars in those bins.

In order to check whether there is some dependence between the slopes and the redshift, a linear regression is performed with all the data points (in blue in figure 23), having the following result:

$$\alpha = (0.024 \pm 0.024) \cdot z + (-1.57 \pm 0.07) \tag{39}$$

The usual notation α for the slope and z for the redshift was used in this line equation. The Pearson correlation coefficient[19] has the value $r = \frac{cov(z,\alpha)}{\sigma_z \sigma_\alpha} = 0.02$. Let us remind that this coefficient measures how strongly correlated[20] this values are: positive values imply a co-growing tendency (the greater one variable is, the greater the other is expected to be) while negative values imply just the opposite (the greater one variable is, the smaller the other becomes). Its absolute value ranges between 0 and 1, 0 being completely uncorrelated and 1 implying they follow an exact functional dependence (therefore r gives no information of the slope of the regression apart from its sign). Both are very unlikely to be obtained with any real data set, given the uncertainties intrinsic to the measuring instruments and procedures[21]. In any case, the value obtained $r = 0.02$ is very small, implying no correlation is apparent between these two variables. In fact, the slope of the line equation, along with its error, is compatible with 0, and the intercept is compatible with the mean value. However, one may argue that there is indeed a dependence, hidden behind the great dispersion of the slopes, and therefore the coefficient obtained accounts for this dispersion rather than this hidden dependence. Nevertheless, this asseveration may be quickly discarded, as such dependence is not observed eliminating the dispersion using the medians or the means in a bin division of the redshift (orange data). In fact, repeating the linear regression but with these orange data[22],

$$\alpha = (-0.01 \pm 0.03) \cdot z + (-1.51 \pm 0.12). \tag{40}$$

Again, this is compatible with the slope being null and the intercept being the mean value of α. Furthermore, the correlation coefficient is $r = -0.05$ (negative as the slope), sufficiently small (in absolute value), and it cannot be blamed on the dispersion in this case. Therefore, these two variables are completely uncorrelated.

A similar analysis can be carried out with the magnitude:

[19] The usual r^2 value is just this value squared.

[20] In a linear sense, but no dependence other than linear is recognisable from neither the scatter plot nor the medians.

[21] In the case of an exact functional dependence of $\alpha = \textit{constant}$, the coefficient would in fact be undefined, as $\sigma_\alpha = 0$. However, this is obviously not the case.

[22] No weighs were used (that is, all points were given the same weighs), because the data points with less error are, ironically, the less reliable ones, as they clearly do not follow the global tendency, given that they are the median of a considerably smaller data bin.

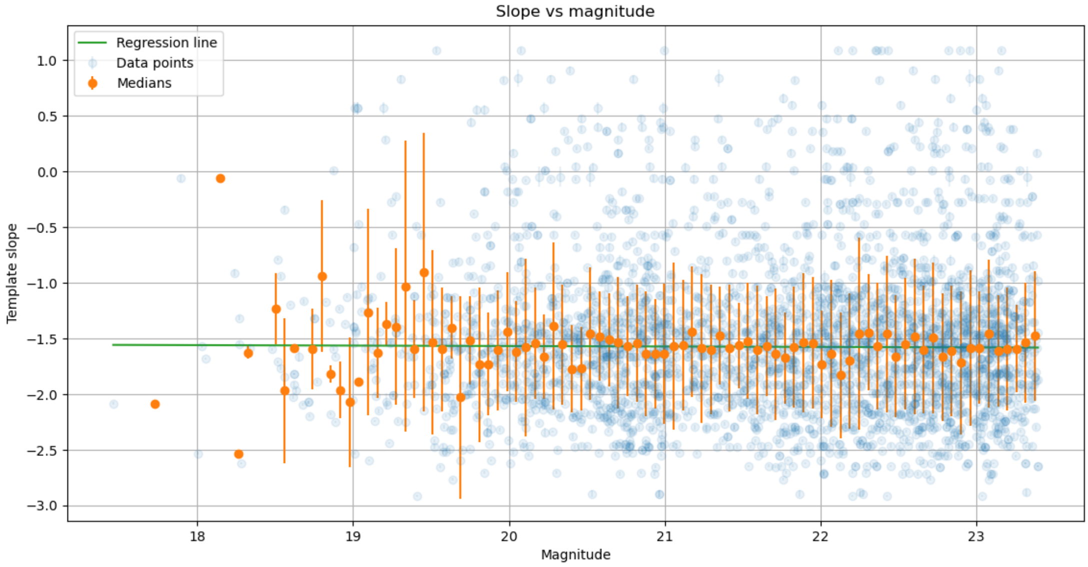


Figure 24: Slopes of the 'template' against r-band apparent AB magnitude. The original slopes, in blue, their medians in uniformly distributed bins, in orange, and the regression line of these, in green. Again, note that some bins are empty (no median is shown in those) and some other, those at smaller magnitude, have very few (10 or less) points, so neither the median nor the mean really represent the real value that would be obtained with a higher number of cuasars in those bins.

The result of the linear regression using all the points yields the following result:

$$\alpha = (-0.005 \pm 0.010) \cdot m + (-1.40 \pm 0.23). \tag{41}$$

As usual, 'm' stands for the r-band apparent AB magnitude. The correlation coefficient has the value $r = -0.008$. No correlation can be inferred by this regression. The same argument as with the redshift applies, although the intercept has bigger error and a value further away from the mean (but the value along with its error is compatible with the mean), the reason being the values in the x-axis are farther away from the origin than with the redshift, so the value at 0 can only be less reliable than in that case. Lastly, as before, I may do the regression with the medians,

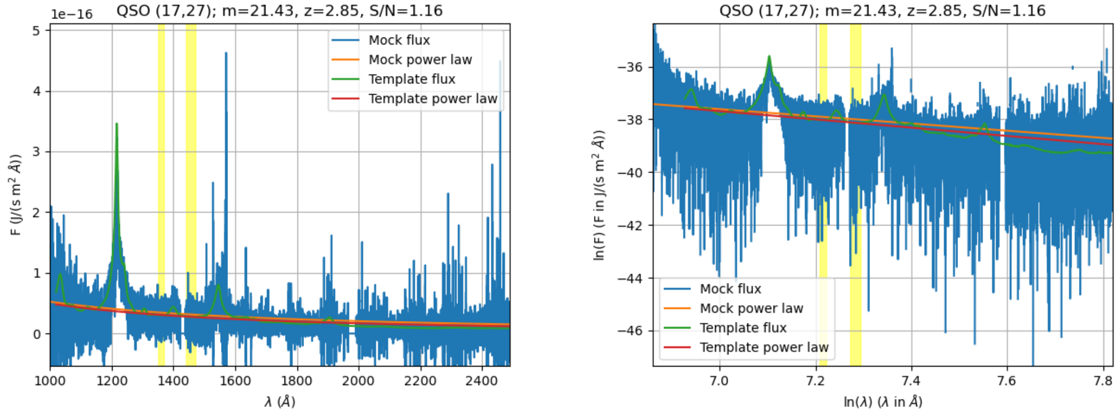
$$\alpha = (-0.004 \pm 0.020) \cdot m + (1.5 \pm 0.4) \tag{42}$$

Therefore, any dependence may be mathematically discarded (it could not be noted visually, either).

To sum up, the slope of the quasars (the real one, as it was calculated using the 'template' data) has no dependence on neither redshift nor magnitude.

B. Some representations of the UV slopes

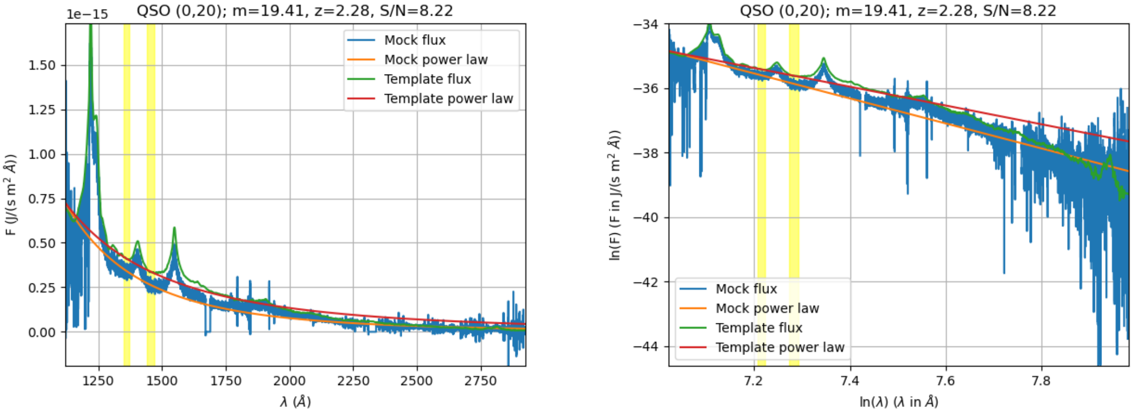
In figures 25 to 29, the title should read ‘QSO (file, number in the file), magnitude, redshift and signal-to-noise ratio. The yellow bands indicate the range of wavelengths used for the regression. The wavelengths are expressed at the QSO rest frame. Different cases are presented here, for the sake of completeness, in using both (λ, F) (‘linear spectrum’) and $(ln(\lambda), ln(F))$ (‘logarithmic spectrum’, the one used for the regression lines to compute the slopes). Note how the errors on the slopes of the ‘mock’ are always bigger.



(a) Linear spectrum.

(b) Logarithmic spectrum.

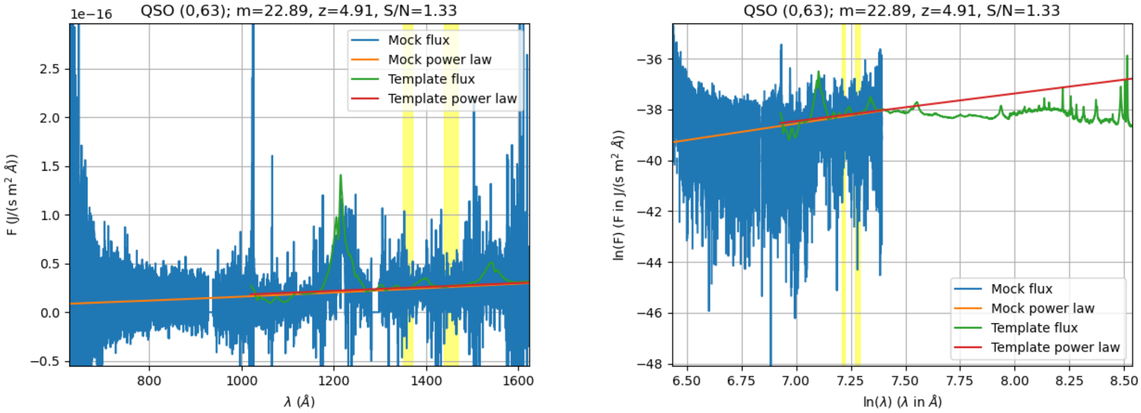
Figure 25: The values of the slopes are $\alpha_{temp} = -1.567 \pm 0.017$ and $\alpha_{mock} = -1.4 \pm 0.4$. This is an example of how a seemingly different slope produce very similar results visually.



(a) Linear spectrum.

(b) Logarithmic spectrum.

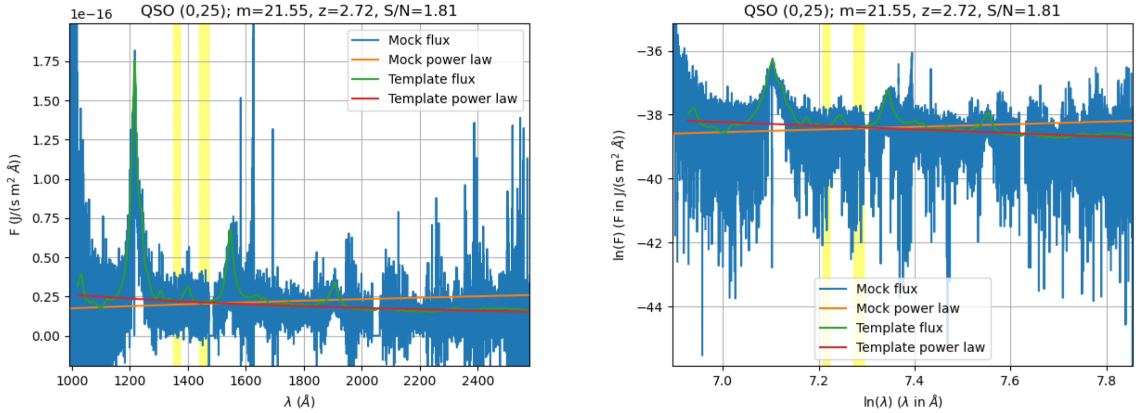
Figure 26: The values of the slopes are $\alpha_{temp} = -2.914 \pm 0.009$ and $\alpha_{mock} = -3.88 \pm 0.06$. This is an example of a very high (in absolute value) slope, which, even if far from the expected $[-2, -1]$ range, is correct. The difference between the curves is mainly due to the normalization constant: the method in 4.1 is good, but not perfect, and sometimes cases like this are produced. However, and especially in the Ly-α forest, the slope itself does not produce a great difference, only the normalization constant. A change of slope can be perceived, especially from 26b.



(a) Linear spectrum.

(b) Logarithmic spectrum.

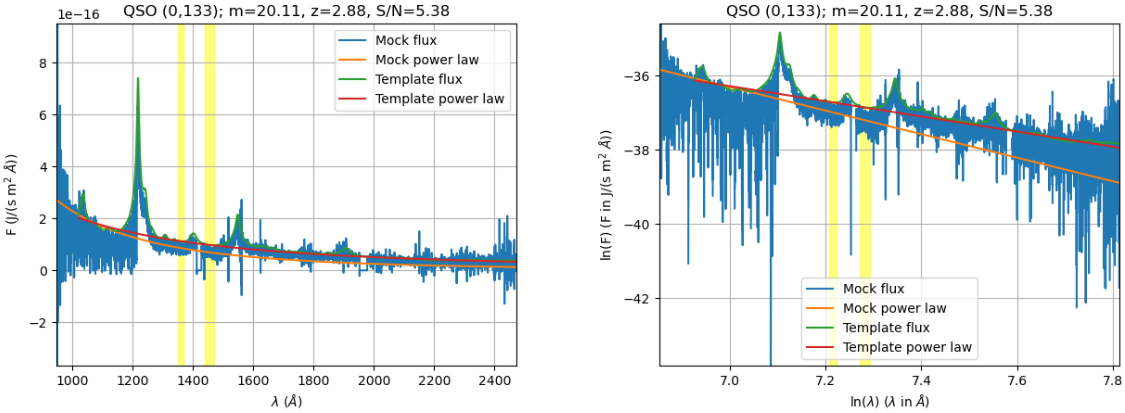
Figure 27: The values of the slopes are $\alpha_{temp} = 1.087 \pm 0.020$ and $\alpha_{mock} = 1.3 \pm 0.3$. This is an example of positive slope. In the linear spectrum, the fit seems appropriate, because I cut the rest of the model, that extends up to 9000 Å. However, the whole spectrum is represented in the logarithmic spectrum, and one can see that the fit was only good locally, the QSO does not follow a power law.



(a) Linear spectrum.

(b) Logarithmic spectrum.

Figure 28: The values of the slopes are $\alpha_{temp} = -0.569 \pm 0.025$ and $\alpha_{mock} = 0.4 \pm 0.4$. This is an example of a slope not well predicted. The reason is purely the low S/N of the QSO in question. Using the method from section 5.1.3 yields the estimation $\alpha \approx -0.288$, considerably better.



(a) Linear spectrum.

(b) Logarithmic spectrum.

Figure 29: The values of the slopes are $\alpha_{temp} = -2.046 \pm 0.008$ and $\alpha_{mock} = -3.18 \pm 0.13$. This is another example where the method from section 5.1.3 evokes better results, as the estimated 'new' slope is $\alpha = -2.18$, pretty close to the 'true' value (α_{temp}).

C. Other visual inspections of optimal spline fits

In figures 30 and 31, two additional examples of handpicked optimal parameters for medium and low (respectively) S/N are included. These two, along with the one from figure 13 are mentioned in the figures of subsection 5.2.2.

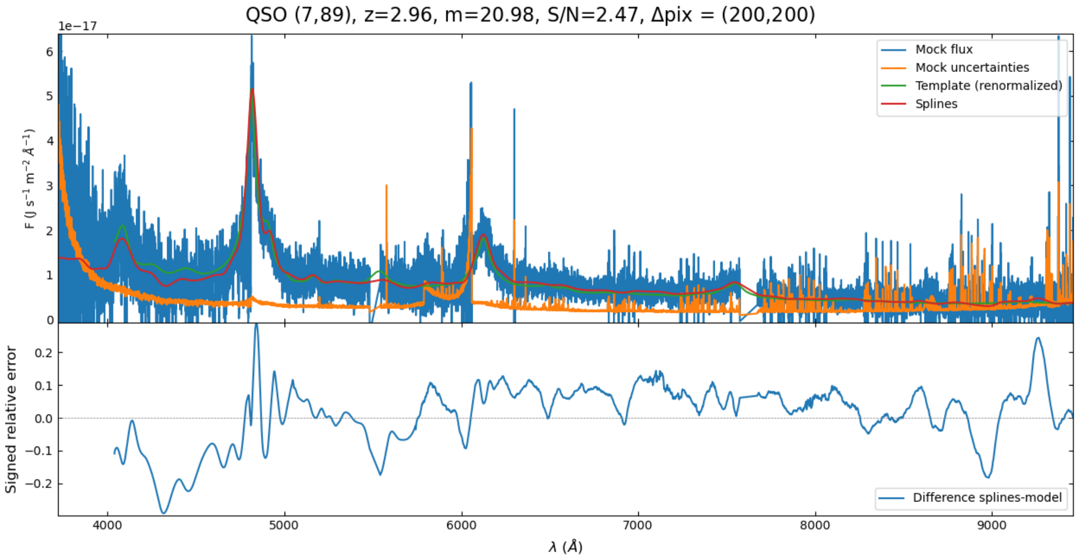


Figure 30: I the upper panel, mock flux, its uncertainties, the spline interpolation and the 'template' flux, renormalized using the constant C_2 over the whole usable spectrum. These $\Delta pix = (\Delta pix_1, \Delta pix_2) = (200{,}200)$ were selected 'by hand' to be the optimal. In the lower graph, the signed relative error ($\frac{F_{spl}-F_{temp}}{F_{temp}}$ with F_{temp} normalized), in parts per unit. The numbers between brackets in the title are simply the file (of the 19 in total) and position within that file.

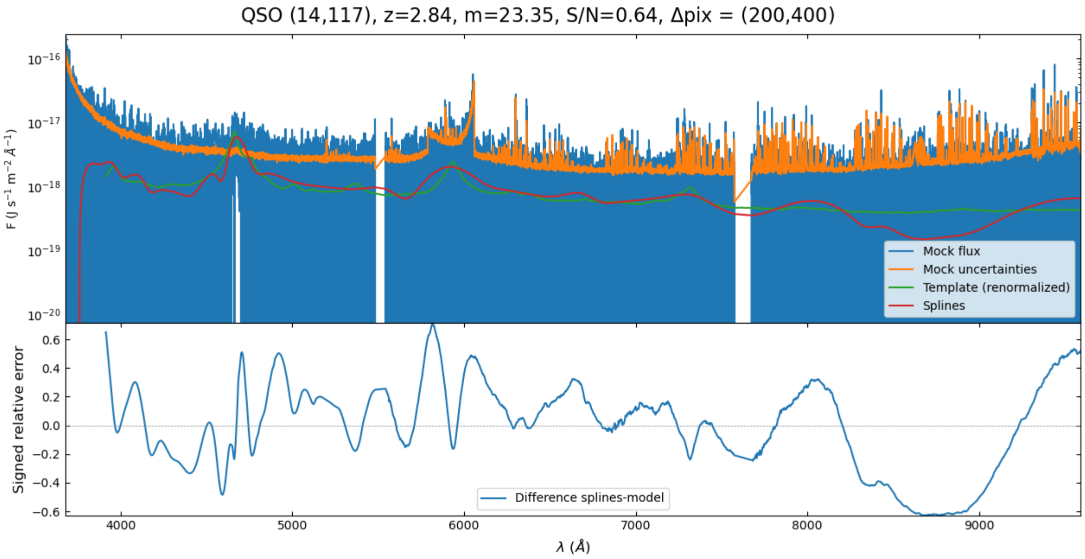


Figure 31: Same as figure 30, but in semilogarithmic scale to appreciate the shape of the template and the splines, that could not otherwise be noticed due to their very small values.

D. Mean value of the relative error using splines

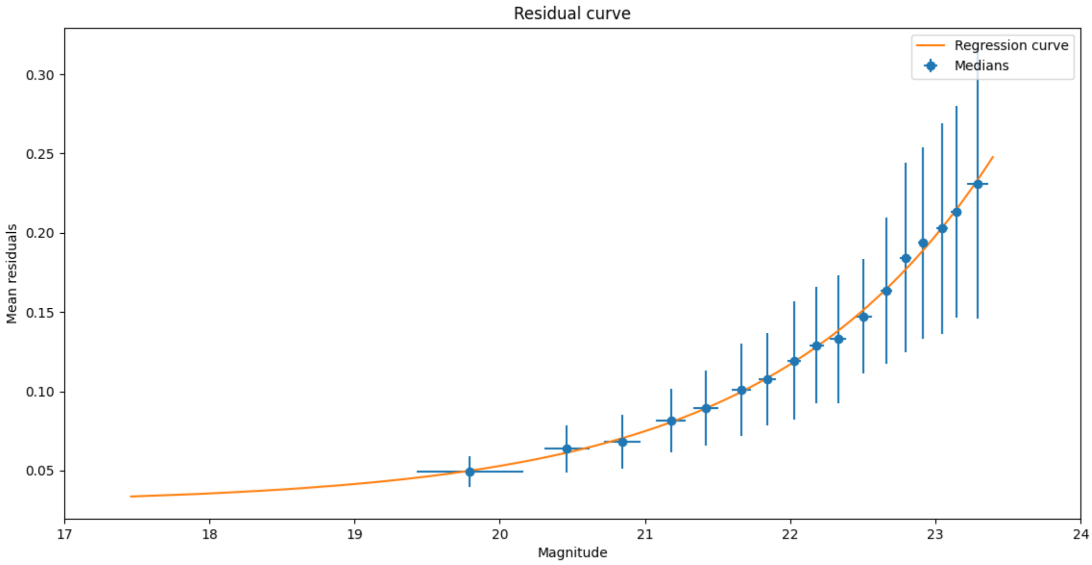
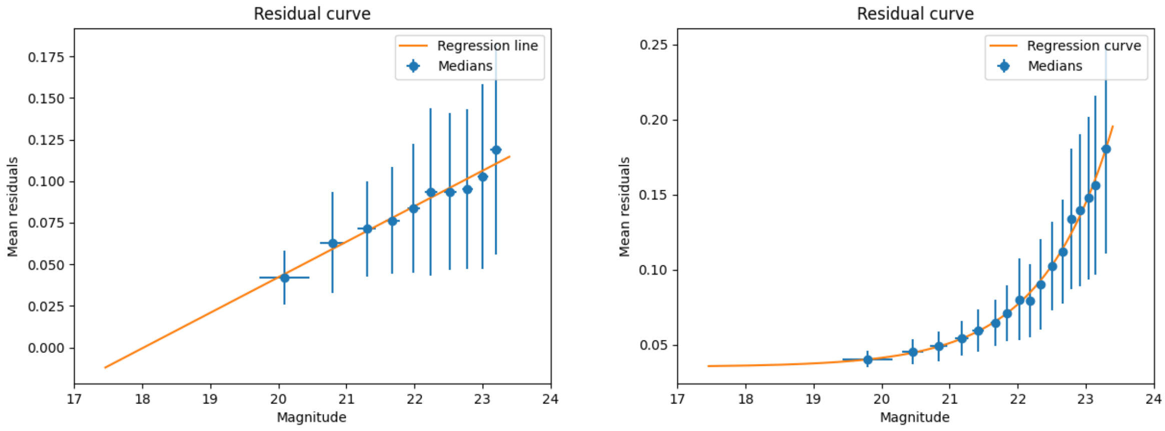


Figure 32: Mean relative error against r-band apparent magnitude for the region R_{total}. By adjusting $err = A + Be^{Cm}$, I obtain $A = (0{,}029 \pm 0{,}005)$, $B = (5 \pm 4) \cdot 10^{-8}$ and $C = 0.65 \pm 0.04$.

As for real data the 'template' variable is not available, estimating the uncertainties of the results derived from continuum reconstruction is not an easy task. In this appendix, I suggest a method for its estimation in the case the continuum is reconstructed using spline interpolation with the optimal of Δpix_1 and Δpix_2 minimizing $resid_3$ and using C_2. By dividing equations (29), (30) and (31) by the number of terms in the summatory and computing it for the optimal Δpix_1 and Δpix_2, an estimation of the corresponding mean uncertainty is obtained. In particular, with $resid_3$, the mean relative error is obtained. Binning in magnitude in the same fashion as in figures 21a to 21d, the following figures 32 and 33 may be obtained.



(a) Mean relative error against r-band apparent magnitude for the region R_1. By adjusting $err = A + Bm$, I obtain $A = (-0.39 \pm 0.03)$, $B = (0.0214 \pm 0.0015) \cdot 10^{-8}$ and $r = 0.98$.

(b) Mean relative error against r-band apparent magnitude for the region R_2. By adjusting $err = A + Be^{Cm}$, I obtain $A = (0{,}035 \pm 0{,}003)$, $B = (2 \pm 3) \cdot 10^{-11}$ and $C = 0.97 \pm 0.06$.

Figure 33: Fittings of the relative errors bluewards (left) and redwards (right) of Lyman-α. In R_1 (left), a line equation produced a better fit.