Summer 8-22-2021

WAVELET PACKET POWER SPECTRUM OF THE SDSS LYMAN-ALPHA FOREST: A TOOL FOR LARGE-SCALE STRUCTURE DETECTION

Jason Pero
DePaul University, JPERO@depaul.edu

WAVELET PACKET POWER SPECTRUM OF THE SDSS LYMAN-ALPHA FOREST: A TOOL FOR LARGE-SCALE STRUCTURE DETECTION

A Thesis

Presented in

Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

August, 2 0 2 1

BY

Jason Pero

DEPARTMENT OF PHYSICS AND ASTROPHYSICS

College of Science and Health

DePaul University

Chicago, Illinois

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

LIST OF FIGURES

LIST OF FIGURES – *Continued*

ABSTRACT

One of the goals of astrophysics is to obtain a full understanding how the Universe is organized on large scales and how structure evolved. In this thesis we develop a method of detecting structure on Mpc scales by measuring the one-dimensional power spectrum of the transmitted flux in the Lyman-α forest. The method is based on the wavelet packet transform (WPT), which has several advantages over the Fourier transform. This includes reduced noise, resulting in less data manipulation and scrubbing in the early stages of analysis. Another advantage is localization of outliers in the data, which allows the general trend of the power spectrum to be revealed despite potentially problematic data. We apply the method to the set of 54,468 quasar spectra from the third collaboration of the Sloan Digital Sky Survey (SDSS-III) Baryonic Oscillation Spectroscopic Survey (BOSS) data release 9 (DR9) catalog. This is intended to be a proof of concept to determine if the wavelet packet power spectrum is a valid technique to extract the power spectrum in order to detect matter density fluctuations. Results are in good agreement with previous studies that used conventional Fourier techniques. The power spectrum vs velocity space plots show increasing power at smaller scales for both our results and earlier studies by [21] and [6]. We conclude that the wavelet packet power spectrum is a tool for detecting structure from transmitted flux in the Lyman-α forest. The advantages the wavelet packet power spectrum over the Fourier transform method are it requires less data manipulation and minimizes noise and propagation of errors and outliers in the data. As a next step we propose applying the tool to the larger more recent SDSS IV eBOSS dataset.

0.1 Acknowledgements

First I would like to thank my thesis advisor Professor Jesús Pando. Thank you for understanding what it is like being an older student returning to a full-time academic program, and providing tailored guidance and advice. Thank you for providing me research experience that has prepared me for doctoral level research. Finally, thank you for the never ending jokes centered around my Windows machine being inferior to a Mac!

I would like to thank Professor Bernhard Beck-Winchatz and Professor Anuj Sarma for their time and energy in being part of my thesis committee.

My deepest gratitude to all of the department physics professors, including those on my thesis committee and Professors Mary-Bridget Kustusch and Chris Goedde. Every physics professor at DePaul University is dedicated to ensuring all students are provided with every opportunity to succeed.

Thank you to all of my fellow graduate students for countless hours working with me on in-class activities and homework problem sets.

Thank you to my family for always believing in me.

0.2 Introduction

The questions of how the Universe is organized on large scales and how structure evolved from the early Universe to the present time are two of the most profound questions in modern astrophysics. Answering these questions provides details on the origin of the Universe, its evolution and fate. Cosmology is the study of the universe as a whole. The focus of cosmology is on structures in the universe which are larger than individual galaxies such as galaxy clusters and superclusters. The term large-scale structure (LSS) is used to refer to these massive structures which are $\sim 150-200$ Mpc across. Examining LSS gives us a clear picture of the evolution of structure in the Universe.

The reason we see structure today in the form of clusters and superclusters is because of the existence of small density fluctuations in the early universe. These density fluctuations got amplified due to gravitational attraction to form LSS. One way to detect fluctuations is by computing the variance as a function of scale (size), otherwise known as the power spectrum.

To study LSS we need large and accurate sky surveys capable of collecting data in various wavelengths of the electromagnetic spectrum. The existence of all-sky surveys has greatly improved the number of high-redshift spectra necessary to study LSS. This study uses data from the SDSS BOSS D9 catalog, which consists of 54,468 quasar spectra in redshift range $2.0 < z < 5.7$. From this dataset we extract the Lyman-α flux and apply the wavelet packet power spectrum. The transmitted flux is proportional to the amount of mass at that location. The more transmitted flux from quasars detected from any particular line of sight means more hydrogen is present, hence more dark matter. This means that the Lyman-α forest can be used as a proxy for the amount of dark matter along the line of sight.

This thesis is a proof of concept to determine if the wavelet packet power spectrum is a valid technique to extract the power spectrum from the Lyman-α forest to detect matter density fluctuations. The section on Large-Scale Structure and Cosmology (Chapter 1) explains the basics of LSS including the influence dynamics of the early Universe had on the formation of structure. It goes on to explain in more detail density fluctuations along with how they relate to the evolution of LSS. An introduction to the classical way of analyzing density fluctuations using Fourier methods is introduced, which includes explanations of the necessary concepts of variance and energy conservation in signal processing. This chapter concludes with a detailed explanation of the Lyman-α forest. Chapter 2 describes the Sloan Digital Sky Survey (SDSS) and the various collaborations leading up to and including the collaboration from which we selected our data, BOSS. Chapter 3 describes methods by which we bin the data and how the wavelet packet power spectrum is calculated. Some necessary background information is included on wavelet theory and the discrete wavelet packet transform (DWPT). This chapter also provides a brief discussion of the advantages of the wavelet transform method of calculating the power spectrum over the Fourier transform method. In Chapter 4 we present our results followed comparison with previous power spectrum studies. Finally, we state our conclusions and discuss future research.

CHAPTER 1

Large-Scale Structure and Cosmology

1.1 Large-Scale Structure

This thesis will develop a tool to probe large-scale structure (LSS). Before we do this, a clear definition of what we mean by large-scale structure and some basic background on how that structure arose is in order. The smallest scale density fluctuations in the universe are on the subatomic quantum scale. Planetary systems, star clusters and galaxies are on a larger scale, however they are relatively small compared to clusters, superclusters and voids which are on the largest scales. The largest superclusters and voids are $\sim$ 150 Mpc across. In cosmology the description *large-scale structure* of the universe indicates all structures larger than individual galaxies.

On scales of $\sim$100 Mpc in diameter, the universe can be approximated as being homogeneous and isotropic [13]. To say the universe is homogeneous means the universe looks the same no matter where you are or "same in all locations". To say the universe is isotropic means it looks the same no matter where you look or "same in all directions". Gravitational instability is the primary instrument for building large structures, voids, clusters, and superclusters. In the early universe, the cosmic microwave background provides clues that quantum fluctuations and dark matter provided an initial foundation for structure formation.

1.1.1 Cosmic Microwave Background

Observations of the cosmic microwave background (CMB) show that the early universe was isotropic and emitted an almost perfect black-body spectrum. This means it is almost homogeneous as well. Density fluctuations have imprinted their signature on the CMB, therefore we know they were present during the time of the last scattering. This was the last time CMB photons scattered from an electron and the CMB photons became free to traverse the universe uninhibited. Photon decoupling is the time when the expansion rate of the universe caused a decrease in interactions between particles. In other words, the rate photons scatter from electrons becomes smaller than the rate at which the universe expands. At the time of photon decoupling, the pressure was reduced which caused the baryonic Jeans mass to decrease suddenly. This caused the baryonic density perturbations to grow after the epoch of photon decoupling.

On scales larger than superclusters, the interactions of baryons and photons has had a profound effect on the evolution of density perturbations through baryon acoustic oscillations (BAO). BAO originated during the epoch of photon decoupling $z \sim 1090$. The gravitational collapse of high-density sections of the photon-baryon plasma was stopped by their own pressure, resulting in an acoustic rebound and an outward expansion of the fluid. Expansion caused the pressure to drop and the fluid to collapse again. This repetitive oscillations of the photon-baryon fluid produce sound waves and are called *acoustic oscillations*. The size of these high-density regions was close to the sound horizon distance, which is the maximum proper distance a sound wave in the photon-baryon fluid of the early universe.

After decoupling, the density fluctuations were low in amplitude, but grew via gravitational attraction to the amplitudes we see today. It was not until

photon decoupling that baryon density perturbations on galaxy and supercluster scales could grow in amplitude. During this time, baryons fell into preexisting gravitational wells of dark matter. This was the beginning of structure formation.

1.1.2 Density Fluctuations

To quantitatively describe density fluctuations, we start by defining the matter density at an epoch in the universe in terms of the average density $\overline{\rho}$ and a local fluctuation $\delta(\mathbf{x})$ known as the density contrast

$$\rho(\mathbf{x}) = \overline{\rho}(1 + \delta(\mathbf{x})) \tag{1.1}$$

Solving the above equation for the dimensionless density contrast yields

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \overline{\rho}}{\overline{\rho}} \tag{1.2}$$

Some important insights can be gleaned from equation 1.2 regarding the behavior of matter. For example, the value of $\delta(x)$ will be negative in underdense regions. However, when the value of $\delta(x)$ is positive, that indicates an overdense region. Overdense regions are where structure (clusters and superclusters) will form, while underdense regions will produce voids. When $\delta(x) = 0$, this corresponds to a critical point. At critical points we can't determine what will happen until later time. As time goes on if a little more matter gets added to the area it will become overdense and create structure. Conversely, if a small amount matter moves out of the region it will become underdense and possibly create a void.

Figure 1.1 plots a toy model of density contrast vs position which further illustrates the concepts of density fluctuations. The x-axis is a simplified position scale in a one-dimensional toy universe. One can clearly see that the overdensities are of different sizes. For example, in the range $0.91 < x < 0.92$ there is a large overdensity, while

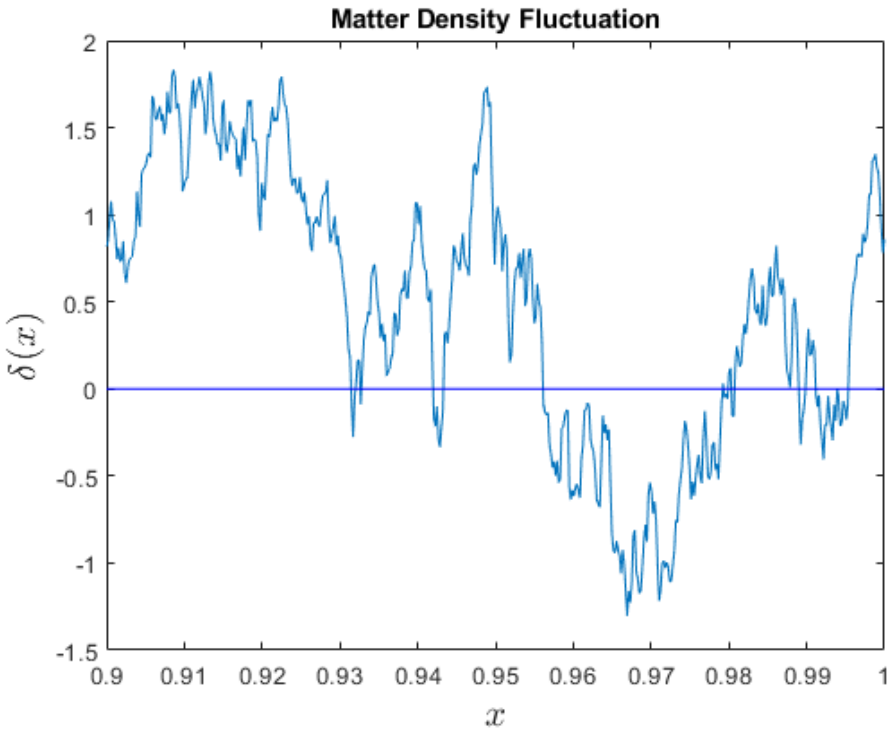

Figure 1.1: Plot of the density contrast $\delta(x)$ vs the position x in a one-dimensional toy model universe. $\delta > 0$ indicate overdense regions, this is where structure will form. Areas where $\delta < 0$ indicate underdense regions which will end up becoming voids. Critical points are when $\delta = 0$. More matter moving into or out of the region in a future time will determine if the region will become overdense or underdense.

$x \sim 0.985$ there is a smaller overdensity. This leads to the production of a variety of different size structures. This means that the size of clusters and superclusters follow a particular theoretical form. We are therefore looking at existing LSS to try and determine what the cosmology is. The pattern of how the structures form tells us about the cosmology that produced them. In other words, different plots like in figure 1.1 will produce different large-scale structure and a different cosmology to describe that structure. Different cosmologies will have different values for their cosmological parameters such as the Hubble constant or the dark energy equation of state.

1.2 Calculating Density Fluctuations using Fourier Methods

The traditional way of calculating density fluctuations is by computing the power spectrum using Fourier analysis methods. One can separate a density fluctuation field into individual Fourier components, and use those components to compute the power spectrum. The power spectrum indicates how much matter is present at each frequency. The following section will provide more details on the process computing the power spectrum by using the Fourier method.

1.2.1 Fourier Analysis

Analogous to equation (1.2), start by defining the density fluctuation field, also known as the density contrast $\delta(\mathbf{x})$, of a large expanding cube with comoving volume V

$$\delta(\mathbf{x}) = \frac{\rho(\mathbf{x}) - \overline{\rho}}{\overline{\rho}} \tag{1.3}$$

which has the following Fourier expansion

$$\delta(\mathbf{x}) = \frac{V}{(2\pi)^3} \int \delta(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3k \tag{1.4}$$

The density contrast can be represented in Fourier components by splitting it into an infinite number of sine waves. Each wave has comoving wavenumber $\mathbf{k}$ and comoving wavelength $\lambda = 2\pi/k$.

$$\delta(\mathbf{k}) = \frac{1}{V}\int \delta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^3x \tag{1.5}$$

1.2.2 Parseval's Theorem

The reason that we can use the power spectrum to determine density fluctuations is because in the act of performing the Fourier transform, energy is conserved. To demonstrate, we expand a signal s in an orthogonal basis function ϕ such that

$$|s(x)|^2 \equiv s(x)s^*(x) = \sum_{-\infty}^{\infty} c_n \phi_n(x) \sum_{-\infty}^{\infty} c_m^* \phi_m^*(x)$$

Where $c_{n,m}$ are the expansion coefficients and asterisk means take the complex conjugate. Next take the norm square and integrate over a Fourier basis interval.

$$\begin{aligned}
\int_0^L |s(x)|^2 \, dx &= \sum_{m,n=-\infty}^{\infty} c_n c_m^* \int_0^L \phi_n(x)\phi_m^*(x) dx \\
&= \sum_{m,n=-\infty}^{\infty} c_n c_m^* L \delta_{m,n} \\
&= L \sum_{n=-\infty}^{\infty} |c_n|^2
\end{aligned} \tag{1.6}$$

where the constant L will change depending on the choice of basis functions. The $|c_n|^2$ terms constitute the power spectrum. The result in equation (1.6) is called *Parseval's Theorem*. After computing the norm square, the basis function ϕ ends up cancelling out due to orthogonality, leaving only the constant L and the coefficent c_n terms. This shows that the *energy* of a signal is distributed to the coefficients of expansion.

1.2.3 Variance

Additional insight can be gained from the power spectrum by inspection of the *variance* because the variance captures local fluctuations. The variance (σ^2) of a function or signal s is defined as

$$\sigma^2[s(x)] = \left\langle s(x)^2 \right\rangle - \left\langle s(x) \right\rangle^2 \tag{1.7}$$

The first term represents the square of the coefficients, which is simply the energy. The second term is the zeroth coefficient, c_0. We can write the variance as

$$\sigma^2[s(x)] = \sum_{n=-\infty}^{\infty} |c_n|^2, \qquad n \neq 0 \tag{1.8}$$

It is common practice to plot n vs $|c_n|^2$ to get information on which components contribute most to the energy (or variance).

Now that we have outlined a method to calculate density fluctuations, the next question is what data can be used to perform these calculations? SDSS provides the flux from quasars. Dips in the transmitted quasar flux are caused by the Lyman-α line. The less flux that we see means there is more matter in that line-of-site. If matter is present, in particular Hydrogen clouds, the Lyman-α wavelength gets absorbed. The *Lyman-α forest* will serve as the data to perform density fluctuation calculations. The next section provides more details on what the Lyman-α forest is and how we will use it to measure large-scale structure.

1.3 Lyman-α forest

Although Lyman-α forest was discovered 5 decades ago, it is one of the essential tracers of the large-scale structure in the Universe. Hydrogen is the most abundant element and therefore can be used as an excellent tool to explore the formation of

structure. The Lyman-α forest contains information about the density of neutral hydrogen gas in the early Universe. Knowledge of the gas density reveals information about matter density fluctuations, which means the Lyman-α forest can be used as a proxy for the amount of dark matter along the line-of-sight.

1.3.1 Lyman series

The majority of our knowledge about the universe comes from the detection and analysis light. The wave-particle duality nature of light allows conceptualization as electromagnetic waves, or a flow of massless particles called photons. Electromagnetic waves are described by their wavelength λ or frequency $\nu = c/\lambda$, where c is the speed of light. Photons are described by their energy $E = h\nu$, where h is the Planck constant, $h = 6.626 \times 10^{-34}$ J s.

The model for a neutral hydrogen atom is a nucleus containing one proton and no neutrons, with a single electron orbiting the nucleus. The only stable state for the electron is in the ground state. Electrons in excited states will decay to lower energy states, emitting photons equal to the difference in energy between the two states in the process according to equation

$$\Delta E = E_n - E_{n,} = \frac{m_e c^2}{2} \alpha^2 Z^2 \left[\frac{1}{(n')^2} - \frac{1}{n^2} \right] \tag{1.9}$$

where m_e is the mass of an electron (9.109×10^{-31} kg) and Z is the number of protons in the atom [23], so in the case for neutral hydrogen $Z = 1$. The starting energy level of the electron is n and the destination energy level is n'. The term α is the fine-structure constant defined as follows

$$\alpha \equiv \frac{1}{4\pi\epsilon_0} \frac{e^2}{\hbar c}$$

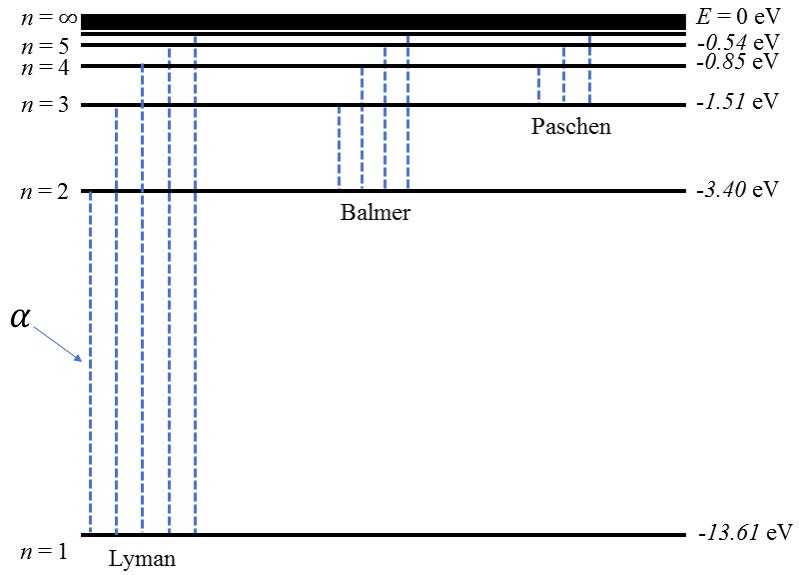

Figure 1.2: The energy level diagram for hydrogen, showing the lower-order lines of the Lyman, Balmer, and Paschen series. The transitions for the Lyman series are labeled on the left. The leftmost line is the Lyman-*alpha* line of the series.

Using the relation $\Delta E = h\nu = hc/\lambda$, the wavelength of the photon emitted is given by

$$\lambda = \left(\frac{2h}{m_e c\alpha^2 Z^2}\right)\left[\frac{1}{(n')^2} - \frac{1}{n^2}\right]^{-1} \tag{1.10}$$

The changes in energy levels in hydrogen are unique and easily recognizable. One example are downward transitions that end at the ground state energy level $n = 1$ produce wavelengths in the ultraviolet portion of the spectrum. These transitions are called the **Lyman series** and were discovered by American physicist Theodore Lyman [16]. Figure 1.2 shows the energy level diagram for hydrogen based on equation (1.9) which indicates the transitions of the Lyman series.

The lowest energy is the first line in the Lyman series corresponding starting energy level $n = 2$ and destination energy level $n' = 1$. This is known as Lyman-α,

or Lyα λ1216. The number 1216 in units of angstroms is the wavelength of the light emitted by this transition, which can be computed using equation (1.10). Emitted photons due to electrons moving to a less excited state are quantized at the energy level differences of the atoms from which they come from. Therefore, the emitted photons produce an *emission spectrum* which is the propagation of photons at a particular wavelength.

An analogous process is when a continuous spectrum containing light from all wavelengths, passes through a gas producing an *absorption spectrum*. This will produce absorption lines at wavelengths or energies corresponding to the atomic transitions of the elements that the gas is composed of. The energy transitions that produce the emission lines are the same as the transitions that produce the absorption lines. This is what allows the unique identification of the particular ions the gas contains. In this case the bright continuous spectrum is the light from distant quasars which shines through gas clouds in the interstellar medium and produces absorption lines. When corrected for redshift, the absorption lines are at wavelength 1216 Å, which identifies them as Lyman-α lines. Both absorption and emission from Lyman-α transitions occur, however our focus is on the transmitted quasar flux which is what remains after the Lyman-α absorption decreases the amount of quasar flux that reaches Earth.

1.3.2 Quasars and Absorption Spectra

Before introducing the Lyman-α forest, which will be our matter tracer in this thesis, it is necessary to introduce the objects that will serve as a background light: the quasars. Quasar stands for "Quasi-stellar radio source". Although they are not transient, quasars are similar to transient events such as supernovae or title disruption events because they are highly luminous and can be seen at large

redshifts ($z > 6$). This makes quasars beneficial for probing the status and content of the intergalactic medium during the early stages of the universe [3].

Quasars contain active galactic nuclei with strong emission line spectra, which is responsible for their high luminosity. Accretion of gas onto the galaxy's central supermassive black hole heats up the gas and produces enormous amounts of radiation, making the quasar much more luminous than all of its stars combined. The most distant quasar known, ULAS J1120+0641, is energized by a $2 \times 10^9\, M_\odot$ black hole and is over 63 trillion times brighter than the Sun. At redshift $z = 7.085$ the spectral lines from this quasar allow visibility into the early Universe only 0.77 billion years after the Big Bang [20]. Like all other galaxies, quasars are dispersed throughout the universe and as their light travels to us it gets intercepted by interstellar gas.

Neutral hydrogen atoms in their lowest state will interact with any source of light that has been redshifted to ultraviolet wavelength of 1216 angstroms when it reaches them. Therefore, as quasar light passes through interstellar gas it kicks electrons in neutral hydrogen into the first excited state. The electron quickly decays back to the ground state and a photon is emitted at the rest wavelength of 1216 angstroms, which is again the Lyman-α line. There are clouds of hydrogen gas positioned between quasars and Earth. Since the gas clouds are closer to Earth than the quasar, their absorption and emission lines will will have smaller redshifts resulting in bluer or shorter wavelengths than the quasar emission line. Multiple gas clouds along the line-of-sight will produce a grouping of Lyman-α absorption or emission lines called the "Lyman-alpha forest". Roger Lynds of Kitt Peak National Observatory is credited with discovering the Lyman-alpha forest in 1971 [4]. Figure 1.3 is an artist's depiction of Lyman-α forest lines created by absorbed light from a distant quasar.

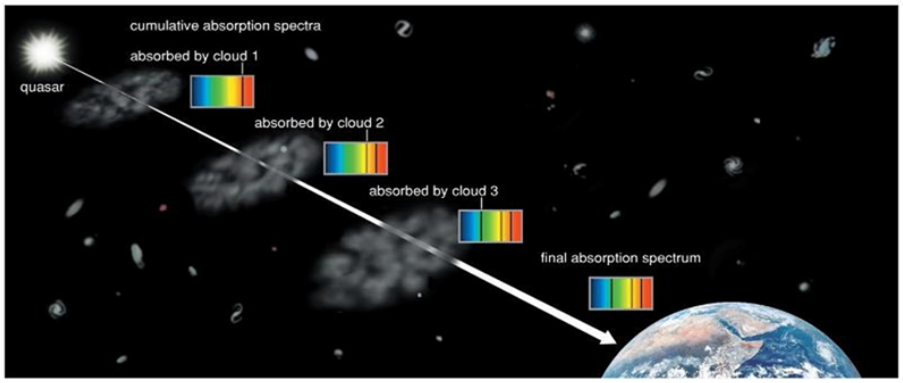

Figure 1.3: Illustration of Lyman-α forest consisting of absorption lines produced by multiple hydrogen clouds along the line-of-sight between the us and the quasar. The absorption lines get redshifted by a value corresponding to the distance the absorbing cloud. The transmitted flux from the quasar is proportional to the amount of mass in the location of the emitting cloud, which is a tracer of dark matter.

There are hydrogen clouds between distant quasars and Earth along our line-of-sight. The clouds absorb ultraviolet light equal to the wavelength of the hydrogen Lyman-α line at a wavelength of 1216 Å and kick electrons in Hydrogen atoms into the first excited state. As we image distant quasars, we are actually viewing light that is passing through dark matter and hydrogen gas. The more quasar flux that we detect means there is less matter along the line-of-sight. If neutral hydrogen is encountered by the quasar light, it will create a decrease in flux due to absorption at the Lyman-α line. Therefore, we can use the clustering models of neutral hydrogen to gain insight into the underlying distribution of matter.

The Lyman-α forest contains information about the density of neutral gas in the early Universe. The positions and clustering patterns of Lyman-α absorption lines is determined by the expansion of space and the matter power spectrum. While the shape of the absorption lines is influenced by the growth and evolution of structure through the column density produced by the absorbing atoms [8].

The Lyman-α line is good for studying high redshift objects because the Lyman-α absorption lines are broad and readily observable in the optical regime. For a Lyman-α absorption line, the observed wavelength λ_{obs} will be shifted due to the expansion of space-time by:

$$\lambda_{obs} = \lambda_{rest}(1 + z) \tag{1.11}$$

where λ_{rest} is the restframe wavelength. Another way to describe the restframe wavelength is the wavelength emitted when measured at the location of the source object. While the central wavelength of the Lyman-α line of the absorbing hydrogen resides in the ultraviolet portion of the spectrum (1216 Å), the cosmological expansion stretches the wavelength that is measured on Earth. For redshifts larger than 1.6, the line is shifted into the visible spectrum and can be detected through Earth's atmosphere[8]. Another feature of the Lyman-α lines which make them noticeable is their broadening (or widening). Lyman-α lines can be broadened by several methods including natural broadening, thermal broadening, and scattering at different densities.

1.3.3 Lyman-α line strength

The optical depth measures how much the intensity of light is diminished when moving through a gas. As the column density of absorbing atoms along the line-of-sight increases, the width of any absorption lines created also increases. The Lyman-α optical depth observed at frequency ν as a result of a homogeneous density distribution of hydrogen is

$$\tau(\nu) = \int_0^{z_s} n_{HI}(z)\, \sigma\left(\nu(1 + z)\right) \frac{c}{(1 + z)H(z)} dz, \tag{1.12}$$

[12], where $n_{HI}(z)$ is the number density of neutral hydrogen, $H(z)$ is the Hubble parameter at redshift z, z_s is the redshift of the source, and σ is the Lyman-α

cross section as a function of frequency. Using the frequency-wavelength relation $\nu = c/\lambda$, the optical depth can be expressed in terms of wavelength (λ)

$$\tau(\lambda) = \int_0^{z_s} n_{HI}(z)\,\sigma\left(\frac{c(1+z)}{\lambda}\right)\frac{c}{(1+z)H(z)}dz \tag{1.13}$$

Higher densities of hydrogen gas contain more atoms to absorb photons, producing a larger optical depth. Higher optical depths means more photons will be absorbed, creating stronger absorption lines in observed spectra. In addition, the large cross section for the Lyman-α transition makes this technique the most sensitive method for detecting baryons at any redshift [2].

Thermal broadening involves the random motion of individual atoms or ions produced line-of-sight velocities which are a function of temperature. Thermal broadening b_T is given by

$$b_T = \sqrt{\frac{2k_BT}{m_H}} \tag{1.14}$$

where m_H is the mass of the hydrogen atom, k_B is Boltzmann's constant, and temperature T is given by

$$T = T_0\left(\frac{\rho}{\langle\rho\rangle}\right)^{\gamma-1} \tag{1.15}$$

where T_0 is the temperature at the average density $\langle\rho\rangle$. During the re-ionization epoch when the gas temperature was practically constant, $\gamma \sim 1$. Long after re-ionization $\gamma \rightarrow (1 + 1/1.7)$ [11]. This shows that temperatures above zero will smear out emission or absorption lines produced by the gas. Doppler broadening also involves the motion of gas particles, but on a larger bulk scale. Doppler broadening does not depend on molecular mass of the particles and can produce larger line widths depending on the gas temperature.

Due to its finite lifetime, the excited state has a degree of uncertainty known as the

Heisenberg uncertainty principle. Since the energy is not specifically determined, the wavelengths of the photons that produce the excited transition are also uncertain. by an amount known as the *natural width* of the line. In the case of pure natural broadening the line profile is determined by the Lorentz distribution

$$\phi(\nu)d\nu = \frac{(\gamma_n/4\pi)}{(\nu-\nu_0)^2 + (\gamma_n/4\pi)^2}\frac{d\nu}{\pi} \tag{1.16}$$

where ν_0 is the frequency of the line center. The term γ_n is a damping which depends on the Einstein A coefficient for transitions from state n to state n'

$$\gamma_n = \sum_{n'<n} A_{nn'} \tag{1.17}$$

The broadened width of the Lyman-α line due to natural broadening can be calculated by [23]

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\nu}{\nu} \approx \frac{\gamma_n/4\pi}{\lambda/c} \approx 2\times 10^{-8}$$

Therefore, the line widening produced by natural broadening is significantly smaller than the widening produced by thermal and Doppler broadening.

The Lyman-α forest lines (typically several hundred per quasar) are far too numerous to be produced exclusively by normal galaxies. Also, the clouds have a homogeneous distribution, which differs from the clustering patterns we see of galaxies. Finally, the objects causing the Lyman-α forest seem to have become rarer as the universe evolved, meaning there are less of them in recent times [4].

Hydrogen clouds responsible for the Lyman-α are diffuse and have low self-gravity. Their gravity is too low to collapse into a star or galaxy, therefore they have not contributed to the formation of structure. However, they also have not dissipated yet. This is further evidence for dark matter, that there is something keeping the intergalactic hydrogen clouds together and in place.

High redshift quasars are useful tools to obtain information about the reionization of the intergalatic medium due to their high intrinsic luminosity and strong Lyman-α absorption lines. In particular, they provide information on the amount of neutral hydrogen in the intergalactic medium. Optical depth is proportional to the density of neutral hydrogen, and the density of hydrogen is a tracer of dark matter. Broadening (thermal, natural, and Doppler) increases the line width of Lyman-α lines making them easier to detect. On account of the prevalence of hydrogen in the intergalactic medium, each quasar produces copious amounts of Lyman-α lines which represent various distances leading to the edges of the early universe. We therefore are able to gain a deeper understanding of the distribution of dark matter and the evolution of large-scale structure.

CHAPTER 2

Sloan Digital Sky Survey

To study large-scale structure surveys are needed that are larger than the largest structures in the survey itself. Advances in telescopes, CCD sensors, other observational instruments allows scientists to view objects at higher redshifts than ever before. As a result, recent large-scale galaxy surveys are able to sample volumes larger than the largest structures ($\sim$150 Mpc in size). The Sloan Digital Sky Survey [26] and its collaborations have been producing data sets that have the sample volumes necessary for studying large-scale structure. The datasets produced by SDSS are available to the public and are what we used in this thesis.

SDSS is the most detailed three-dimensional astronomical survey of the Universe ever made. SDSS maps or images a region of the sky without regard to any one particular observational target. SDSS "first light" was in May 2000 and has been in operation for over two decades. Our focus will be on the third-generation survey SDSS-III. During the course of SDSS-III operation multiple collaborations were formed which studied supernovae to measure the expansion history of the universe, stellar spectra observations to determine the dynamics and chemical evolution of the Milky Way (APOGEE and SEGUE-2), populations of extra-solar giant planets (MARVELS), and mapping the clustering of galaxies and intergalactic gas in the distant universe (BOSS). Our focus will be on BOSS, however we give a brief overview of the four collaborations in the rest of this section. Then in the next section, 2.3 we provide a detailed account of the BOSS survey.

2.1 SDSS-I and SDSS-II

SDSS-III builds on the instrumentation and work done by the prior surveys SDSS-I and SDSS-II, so a brief introduction to these first two surveys is provided next. The first eight years of operation included the first two surveys, SDSS-I from 2000-2005 and SDSS-II from 2005-2008. Together these first two surveys covered over 25 percent of the total sky by capturing deep multi-color images to create three-dimensional maps of the universe. The conclusion of SDSS-II marked the successful image capture of 930,000 galaxies, 120,000 quasars, and 460,000 stars.

SDSS-II consisted of three collaborations: the Legacy Survey, the Supernova Survey (SN), and the Sloan Extension for Galactic Understanding and Exploration (SEGUE). The Legacy Survey was the original sky imaging plan which was in operation from 2000 to 2008. Sky coverage for the Legacy Survey provided a 7,500 square degree map of the North Galactic Cap and 740 square degrees in the South Galactic Cap. SN found and monitored hundreds of supernovae in a southern equatorial stripe with the goal of measuring the expansion of the universe. SDSS-II SN also measured lightcurves for more than 500 spectroscopically confirmed Type Ia supernovae in the redshift range $0.1 < z < 0.4$. SEGUE focused on the Galactic plane and charted the motions and composition of 240,000 stars in the Milky Way. At the conclusion of SDSS-II in July 2008, SDSS-III began in autumn of the same year. Both surveys were executed with the dedicated, wide-field of view, Sloan Foundation 2.5-m telescope at Apache Point Observatory (APO), a large mosaic CCD camera, and a pair of double spectrographs, each outfitted with 320 optical fibers plugged into custom-drilled aluminum plates. To process all of the data, an elaborate system of data reduction, calibration pipelines, and data archiving systems was developed [10].

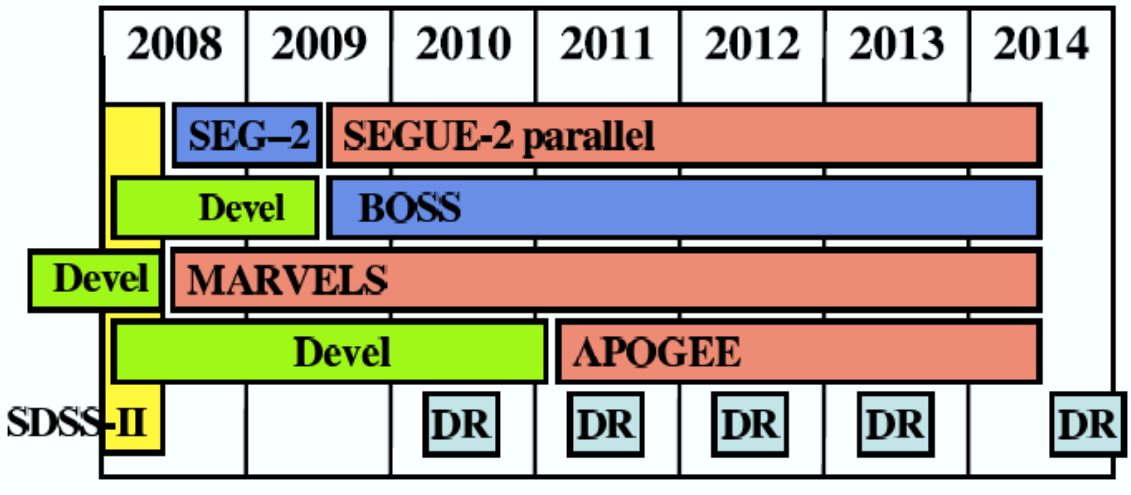

Figure 2.1: High-level SDSS-III observing schedule. Dark-time observing programs are marked in blue and bright-time observing programs are marked in red. Hardware development activities are marked in green, and light blue squares mark the five public SDSS-III data releases; DR8, DR9, DR10, DR11, and the final SDSS-III data release, DR12 in 2014. Image from `http://www.sdss3.org/collaboration/description.pdf`.

2.2 SDSS-III Overview

This third generation of SDSS consisted of four different collaborations: The APO Galactic Evolution Experiment (APOGEE), the Sloan Extension for Galactic Understanding and Exploration 2 (SEGUE-2), Multi-object APO Radial Velocity Exoplanet Large area Survey (MARVELS), and the Baryon Oscillation Spectroscopic Survey (BOSS). The details of BOSS will be provided in the next section, after a brief description here of the other three surveys. Figure 2.1 displays the high-level schedule that was followed for the four SDSS-III surveys and their development activities.

2.2.1 APOGEE

The APO Galactic Evolution Experiment used near-infrared spectroscopy to survey over 100,000 red giant stars in the Milky Way Galaxy. Measurements of chemical composition, stellar parameters, and high precision radial velocity allow APOGEE to provide clues to the dynamical and chemical history of our Galaxy. The APOGEE survey took advantage of "bright" time, when the Moon is more than 60% illuminated, a time which the BOSS survey cannot operate.

2.2.2 SEGUE-2

Sloan Extension for Galactic Understanding and Exploration 2 was an extension of SEGUE-1 and together they provided spectroscopic observations of approximately 350,000 stars with the goal of investigating the kinematics and chemical composition of the Milky Way. Specifically, these surveys focus on objects in the distant Galactic halo and how it evolves over time. The goal is to understand stellar dynamics in the halo and how they influence formation and metal enrichment throughout the Galaxy.

2.2.3 MARVELS

The goal of the Multi-object APO Radial Velocity Exoplanet Large-area Survey is to test theoretical models of the formation, migration, evolution of giant planet systems. To accomplish this goal MARVELS observed radial velocities of 11,000 bright stars in attempts to find giant planets with orbital periods as short as a few hours and as long as two years. Each star was observed approximately 20-40 times, with a typical exposure lasting 50-60 min. Data collection began in Fall 2008 and ended in Summer 2012.

2.2.4 BOSS at a glance

The Baryon Oscillation Spectroscopic Survey (BOSS) consists of three scientific goals: dark energy and cosmological parameters, the history and structure of the Milky Way, and extrasolar giant planets. Its goal was to carry out precision BAO measurements from the Lyman-α forest at $z \sim 2.5$ [14]. The BOSS survey measured redshifts of 1.5 million high mass galaxies and Lyman-α forest spectra of 150,000 quasars to obtain estimates of the distance scale and Hubble expansion rate at $z < 0.7$ and $z \approx 2.5$. The next section gives a more detailed account of BOSS including instrument upgrades and data processing.

2.3 BOSS

2.3.1 Telescopes, Cameras, and Spectrographs

SDSS-II and SDSS-III used the same two telescopes for imaging. The first telescope is a wide-field Sloan 2.5m telescope at APO in New Mexico and the other is a du Pont 2.5m telescope at Las Campanas Observatory (LCO) in Chile. Although the telescopes are different models, they both have identical spectrographs attached (see figure 2.2). The primary goal of the SDSS spectrographs is the creation of a three-dimensional wide-area map of the universe to reveal its large-scale structure. To accomplish this, spectroscopic observations use between six and nine plates each night. These are large aluminum plates with small holes drilled in them. Each hole has a fiber optic cable which must be manually plugged in each observing night. Each hole corresponds to an astronomical object (star, galaxy, quasar, or random empty area to subtract background noise), and each plate is custom drilled to correspond to a specific patch of sky. The 2.5m telescopes have a field of view of 7 deg^2, and each plate views a sky area of 5 deg^2. A high number of fibers and

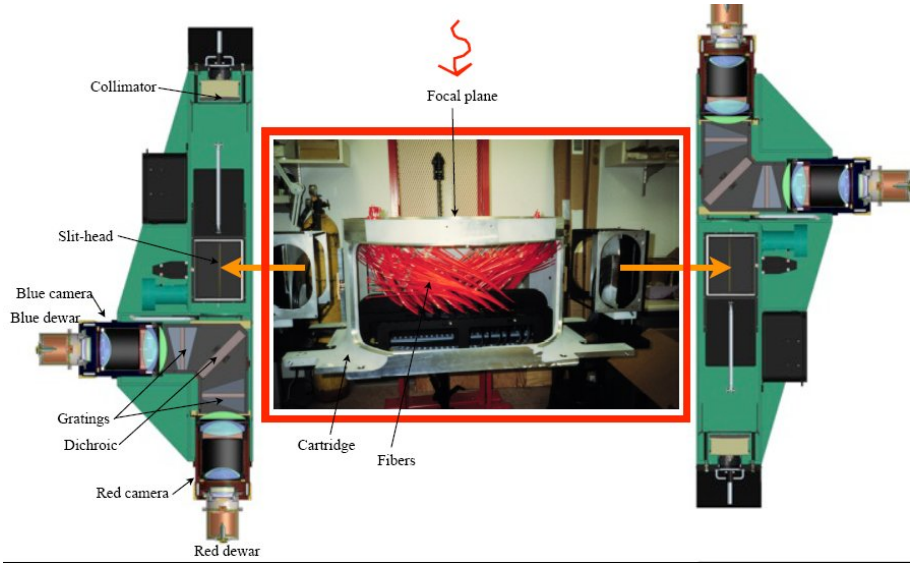

Figure 2.2: Illustration of the BOSS spectrograph setup. Image from `https://www.sdss.org/instruments/boss_spectrograph/`

higher density plates were used to ensure the survey could complete analysis of the entire sky. To ensure complete sky coverage, there is a need for overlap between fields of view. Also, sky-background measurements and cross-calibration of the full survey require multiple observations.

The telescopes are Cassegrain models, in which the focus is behind a parabolic primary mirror. The focus is accessible through a hole in the primary mirror through which reflected light from a secondary hyperbolic mirror passes. The advantage of the Cassegrain model is that telescopes with a long focal length can still have a compact design. Also, the focus is at the base of the telescope behind the mirror, allowing easy access to mount large cameras and spectrographs. Consequently, the telescope has two double fiber-fed spectrographs permanently mounted on the image rotator (see figure 2.3). There is also a photometric/astrometric mosaic camera, mounted at the Cassegrain focus. The camera images the sky with a scanning path that follows great circles at the sidereal rate. The telescope enclosure

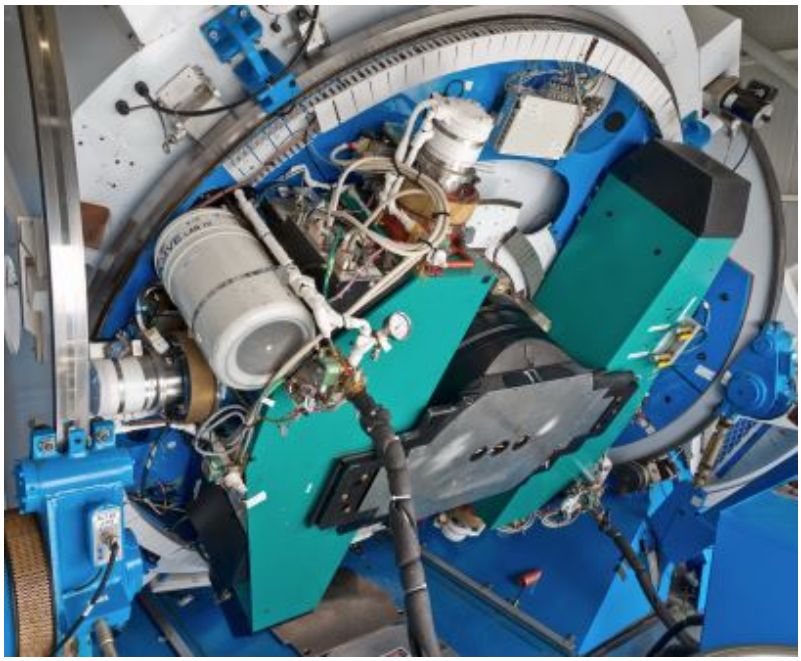

Figure 2.3: The actual SDSS spectrographs mounted to the Cassegrain image rotator. Spectrograph 1 is the green device on the right. Spectrograph 2 is the green device on the left. In this photograph, the telescope is at a zenith angle of 60°. Figure from [24]

and mounting were constructed for ease of access and to facilitate fast swapping between fiber plug plates and between imaging and spectroscopic modes. This allows for an ideal observation strategy where sky imaging is done during optimal weather conditions, and spectroscopy is performed during less pristine observing conditions.

Imaging data was captured during clear, dark nights then reduced and calibrated. The processed imaging data was then used to select spectroscopic targets. The imaging procedure begins with scanning the sky on the first night. After scanning the sky, one thousand holes are drilled in an aluminum plug plate. One thousand optical fibers are then plugged into the holes to allow light from the focal plane to flow to the pseudoslit of the spectrographs. Flux captured from quasar

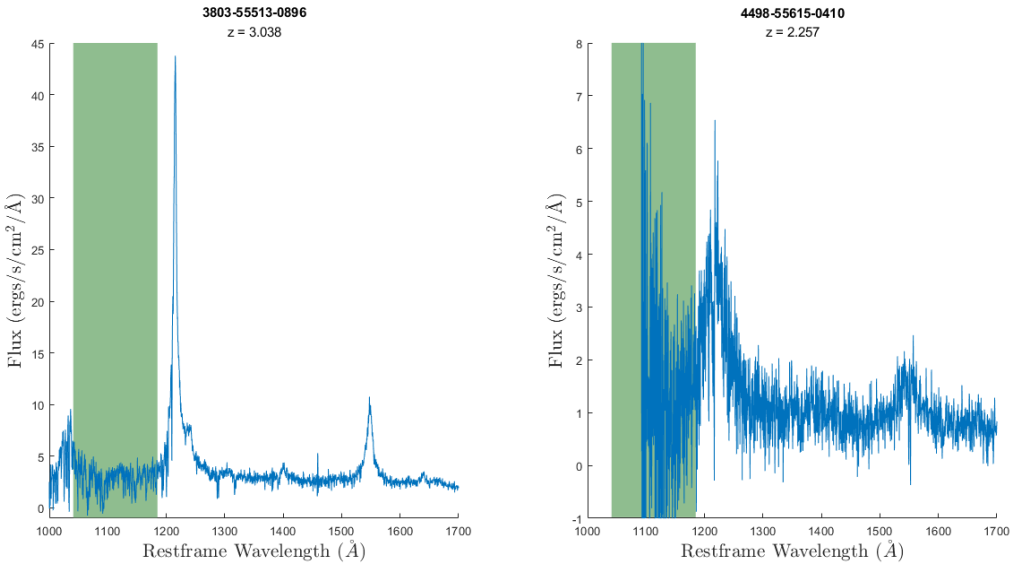

Figure 2.4: Spectra of randomly selected quasars from our sample, BOSS Lyman-α data release 9. The green region highlighted in each graph represents the selected range of the Lyman-α forest region $1041 < \lambda_{rest} < 1185$ Å.

light can be seen in figure 2.4. On the second night, the sky is re-scanned and the spectra for the one thousand objects is collected. Of this detail collected we'll be using the Lyman-α forest spectra produced by quasars. For the SDSS-III survey smaller diameter fibers that better matched the angular scale of BOSS targets were installed. The original 640 fibers used for SDSS-I and SDSS-II where increased to a total of 1000 fibers for SDSS-III BOSS. This increased the total number of simultaneous spectra readings from the two spectrographs. The SDSS imaging camera measures wavelengths on both the blue and red ends of the spectrum. On the blue side, the short wavelength limit was set to 3900 Å. For longer wavelengths, the red channel was added to enable observations of H-α to a redshift of $z = 0.2$ or more, and the observation of quasars out to redshifts beyond $z = 5$.[24] The upper wavelength limit was therefore set to 9100 Å. For the BOSS survey the limits were extended to 3560 Å$< \lambda <$ 10,400 Å. Also, upgraded CCD cameras were installed with higher quantum efficiency and smaller pixels. Overall the peak instrument

efficiency was increased from 45% to 70%.

The spectroscopic resolution of 1500 at 3800 Å was set to measure spectroscopic redshifts of galaxies accurate up to the limit imposed by Doppler broadening caused by velocity dispersions in the range 100 to 200 km/s. The BOSS upgrades increased the resolving power to 2500 at 9000 Å.

2.3.2 Data Selection

Our selected data set is from the BOSS Lyman-α Forest Sample from SDSS Data Release 9. This release is comprised of 54,468 quasar spectra with $z_{qso} > 2.15$. This data set examines absorption redshifts in the range $2.0 < z_\alpha < 5.7$ with sky area coverage over 3275 square degrees, enclosing an estimated comoving volume of $20\,h^{-3}\,\mathrm{Gpc}^3$ [14]. Figure 2.5 shows the redshift distribution of the quasars, and Figure 2.6 shows their distribution across the sky. There are many steps to perform against a set of quasar spectra before it is ready for cosmological analysis of the Lyman-*alpha* forest. These steps include removing regions affected by damped Lyman-α absorbers (DLAs), flagging inaccurate data, and accurate evaluation of noise.

2.3.3 Data Reduction Pipeline

Lee et al. (2013) [14] describes several steps that must be taken to prepare a set of quasar spectra for cosmological analysis of the Lyman-α forest. These steps involve removing unreliable pixels, corrections to account for contamination by damped Lyman-α absorbers (DLAs) or broad absorption lines (BALs), noise corrections, and conclusively defining the full continuum baseline, before it is altered by absorption. All of this difficult work was already completed by the SDSS consortium and we

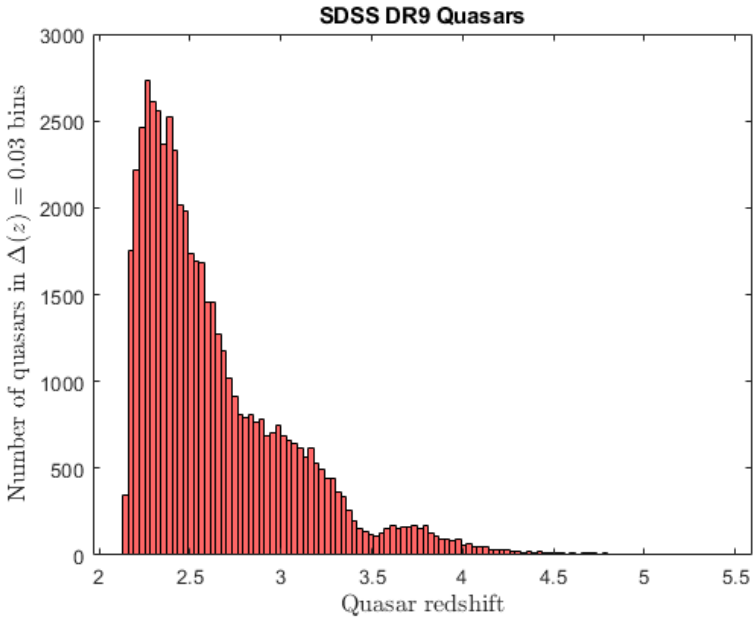

Figure 2.5: Redshift distribution of the 54,468 high-redshift ($z > 2$) quasars used to probe the Lyman-α forest.

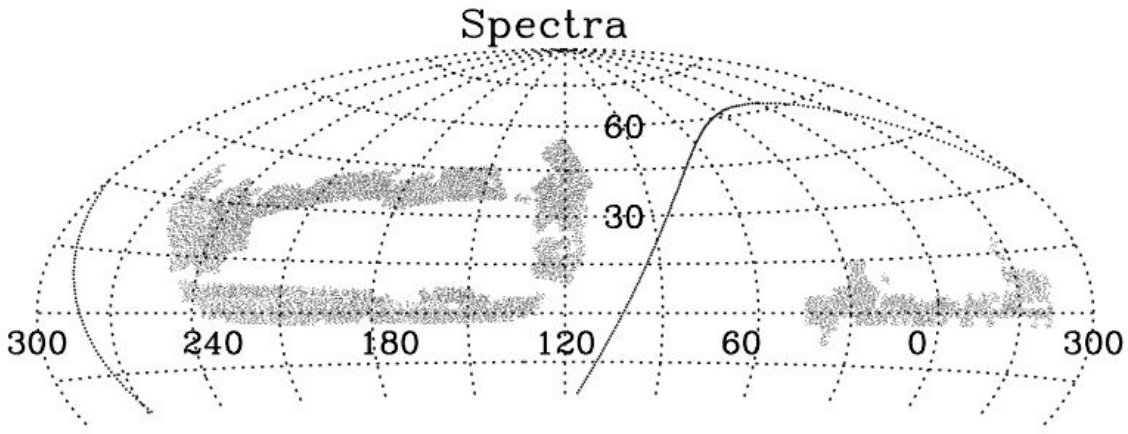

Figure 2.6: Sky distribution of all SDSS BOSS DR9 spectroscopy (Includes stars, galaxies, and quasars). The solid line represents the Galactic equatorial plane. Figure from [1]

simply applied the recommended masks and corrections to the data to obtain the full Lyman-α data set. The next sections summarize what the mask and corrections are and why they are important.

Pixel Masks

SDSS BOSS Lyman-α analysis employs a bitmask system to flag unreliable pixels which should be abandoned. Lee discusses the pipeline mask as the first bitmask system which processes and calibrates BOSS galaxy and quasar flux data by way of an automated classification and redshift measurement software called **idlspec2d**. The second bitmask system he discussed is primary source of pixel noise, the sky. The sky mask flags pixels that should be discarded due to the saturation by sky variance, which drowns out the astrophysical flux signal.

BALs and DLAs

A broad absorption line (BAL) quasars have a continuous broad absorption spectrum and are associated with high velocity active galactic nuclei (AGN) outflows from an accretion disk [15]. Lee states that BAL troughs may affect the continuum fitting and introduce intrinsic quasar absorption into the Lyman-α forest region. Therefore, 5,848 quasars were visually flagged as BAL quasars and omitted from the data set.

Some hydrogen clouds are so dense that their outer layers shield neutral hydrogen in the interior from radiation. These high density regions will cause the absorption to plummet. This type of region is known as a damped Lyman alpha (DLA) region. These collapsed DLA regions therefore do not effectively trace the underlying dark matter fluctuations on large scales. Also, according to Lee, each

DLA produces large damped absorption profiles that pollute large portions of sightlines ($\Delta v \gtrsim 5000$ km s^{-1}). Thus, it is desirable to exclude DLAs from any large-scale Lyman-α forest analysis.

Noise Corrections

It is difficult to estimate continua when the spectra contain a lot of noise. The idlspec2d software tool provided a noise estimation σ_p which took into account each pixel in each spectrum. The overall noise estimate was refined by application of a combination of several noise correction factors including individual exposure correction (cor_{exp}), co-addition correction (cor_{add}), and flux-dependent correction (cor_{flux}). The total correction to the pixel noise is then

$$cor_{tot}(\lambda, f) = cor_{exp} \times cor_{add}(\lambda) \times cor_{flux}(f, \lambda) \tag{2.1}$$

To produce a more accurate noise estimate, the pipeline noise estimate is divided by the overall noise correction

$$\sigma_{cor} = \frac{\sigma_p}{cor_{tot}(\lambda, f)} \tag{2.2}$$

where f is a modified weighted flux for a given pixel. For more details on the noise correction procedure see [14].

Continuum Fitting

The spectrographs collect photon counts within specific channels, they do not obtain the actual spectrum. The actual spectrum of the source can be determined by direct comparison of an observed spectrum with a set of models or an empirical library of objects with known characteristics on a pixel-by-pixel basis. In the case of estimating quasar continua, typical methods include models which perform a

power-law extrapolation from $\lambda_{rest} > 1216\,\text{Å}$. However, due to uncertain blue-end spectroscopy in BOSS and a break in the quasar continuum at $\lambda_{rest} > 1200$ Å, continuum extrapolations are unreliable. Consequently, estimation of the continuum must be accomplished using information in the Lyman-α forest itself.

Extraction of the transmitted Lyman-α flux is accomplished by dividing the observed flux by an estimate of the intrinsic quasar continuum. A two-step process using a modified version of the mean-flux regulated principal component analysis (MF-PCA) technique was used to obtain a continuum estimate. An initial PCA fit to predict the shape of the Lyman-α forest continuum, and a step to ensure the continuum amplitude is consistent with published constraints on the Lyman-α forest mean-flux.

CHAPTER 3

Wavelet Packet Power Spectrum

Our primary method of analysis uses wavelet theory, the basis of which are wavelets. Like the Fourier transform, we use the DWPT to separate a density fluctuation field into individual components. Since Parseval's Theorem for energy conservation holds for wavelet transforms, we can use those DWPT coefficients to compute the power spectrum.

3.1 Introduction to Wavelet Analysis

3.1.1 Wavelets and Wavelet Systems

A *wavelet* is a "small wave" which has its energy concentrated in time. This allows analysis of transient, non-stationary, or time-varying phenomena. In addition, it allows simultaneous time and frequency analysis [5]. Figure 3.1 displays an example wavelet known as the Morlet wavelet, which has finite energy focused between location -3 and $+3$.

A function or signal $f(t)$ and an orthogonal set of basis functions ψ_i can be expressed as a linear combination

$$f(t) = \sum_i a_i \psi_i(t) \tag{3.1}$$

where i is an integer index for an infinite or finite sum and a_i are expansion coefficients. For orthogonal basis, the inner product of the expansion set functions is

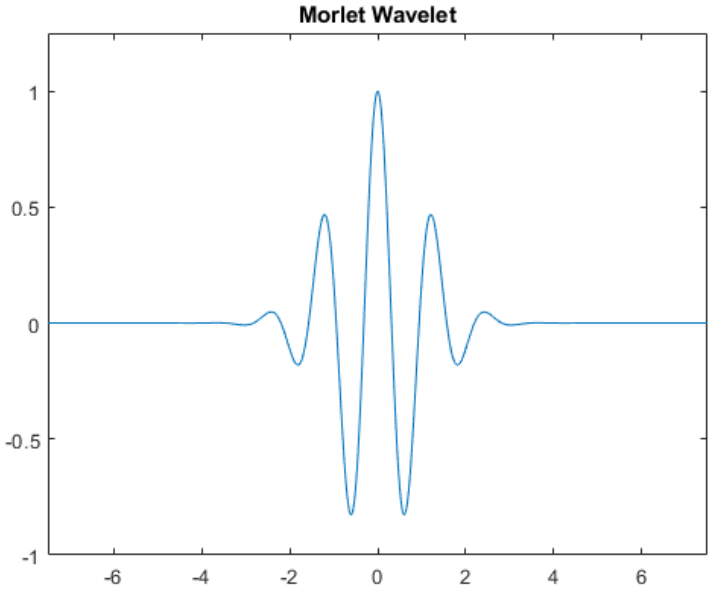

Figure 3.1: Example of a wavelet.

zero when the indices are not equal

$$\langle \psi_k(t), \psi_i(t) \rangle = \int \psi_k(t), \psi_i(t) dt = 0 \qquad k \neq i, \tag{3.2}$$

and the coefficients of the expansion can be calculated by the inner product

$$a_k = \langle f(t), \psi_k(t) \rangle = \int f(t) \psi_k(t) dt \tag{3.3}$$

For the case of a wavelet expansion, a two-parameter system can be developed from equation (3.1) to produce

$$f(t) = \sum_k \sum_j a_{j,k} \psi_{j,k}(t) \tag{3.4}$$

where j, k are integer indices and $\psi_{j,k}$ are wavelet expansion functions that form an orthogonal basis. This provides the basis for building wavelet systems which can map a function into a two-dimensional array of coefficients. The j index is the scaling index, and k is the translation index. Combined with the summations, this

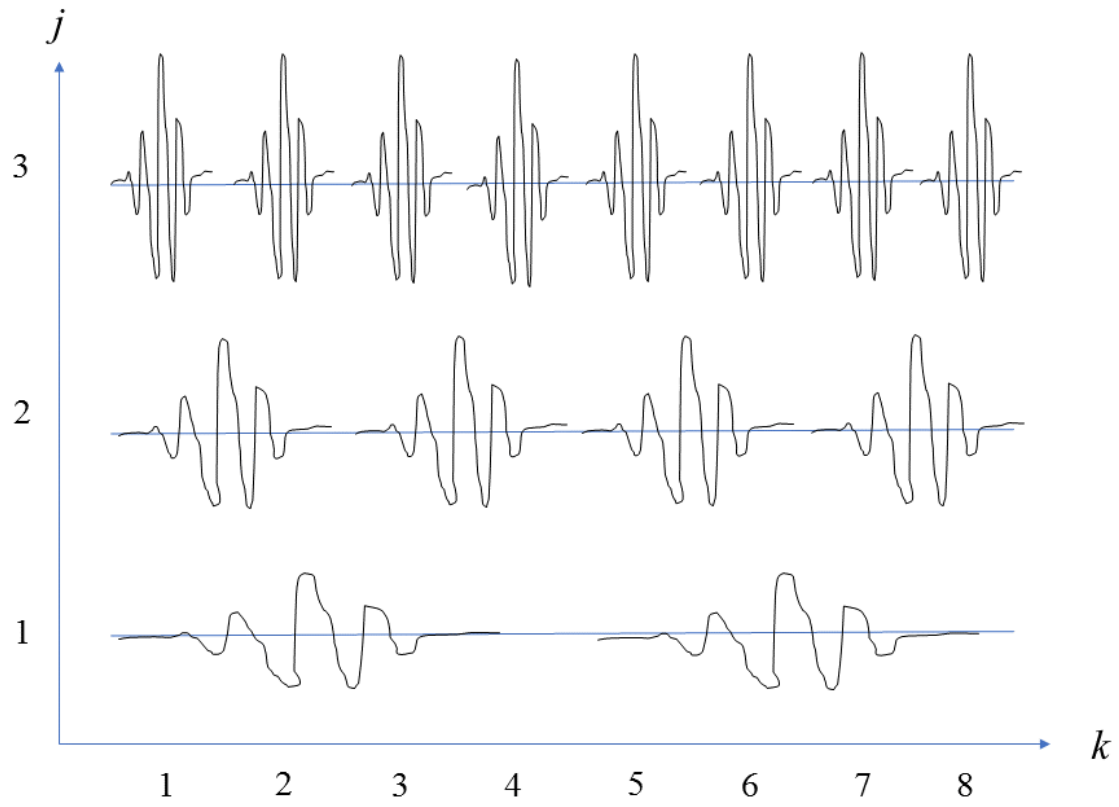

Figure 3.2: Translation and scaling of a wavelet ψ

system moves the wavelet across the signal (which is called *translation*) sampling different portions of it. It also can adjust or change the scale of the wavelet to enable approximations at different resolutions.

Figure 3.2 is a graphical portrayal of the scaling and translation of an individual wavelet. The ψ term is an arbitrary wavelet which could be Haar, Daubachies, Morlet, etc. The location of the wavelet can move along the horizontal axis by changing the value of the index k. Adjustment of the k index allows the wavelet expansion to definitively analyze a specific location of events in time or space. On the vertical axis, the shape of the wavelet is controlled by the value of

the index j, which corresponds to changing the scale of the wavelet. The shape of the wavelet changes in scale as the index j changes. Higher resolution or higher detail is achieved by increasing the scale. Larger scales (higher values of j) produce a taller, thinner wavelet. A thinner wavelet coincides with smaller steps in time for the wavelet duration. It is this two-dimensional procedure of thinner wavelet and smaller transnational steps that produce higher resolution levels. As alluded to earlier in this paragraph, in designing the wavelet system, the shape of the mother wavelet can be changed to better suit analysis of the specific signal under study. This dual parameterization consisting of scaling and translation functions in wavelet systems allows localizing of the signal in both the time and frequency domains.

The basis functions of wavelets are produced by two entities, the scaling function, also known as the *father wavelet* $\phi(t)$, and the mother wavelet $\psi(t)$ [19]. The scaling function has the general form

$$\phi_k = \phi(t-k) \qquad k \in \mathbb{Z} \tag{3.5}$$

The mother wavelet is based on the scaling function and produces a two-dimensional characterization which parameterizes time (or space) k and frequency (or scale) j and has the form

$$\psi_{j,k} = 2^{j/2}\psi(2^j t - k) \qquad j,k \in \mathbb{Z} \tag{3.6}$$

3.1.2 Multiresolution Analysis

Fundamental to wavelet analysis is *multiresolution analysis* (MRA) which is decomposition of a signal (for example a quasar flux signal) into subsignals of different size resolution levels [25]. MRA is the core of wavelet analysis and depends on the requirement of nesting of the vector spaces spanned by the scaling functions. In

terms of signal processing, this means that the space containing high-resolution signals will also contain those of lower resolution. Another way to describe this feature is that the elements in a space are just scaled versions of the elements in the next space. Burrus et al.[5] describes how the MRA version of the scaling function can be written as a shifted scaling function $\phi(2t)$

$$\phi(t) = \sum_n h(n)\sqrt{2}\,\phi(2t-n), \quad n \epsilon \mathbb{Z} \tag{3.7}$$

The term $h(n)$ represent a sequence of real or complex numbers called the scaling function coefficients. The $\sqrt{2}$ maintains the norm of the scaling function with a scale of two. The choice of the coefficient $h(n)$ is what defines the design of the wavelet system.

The mother wavelet or wavelet function is a weighted sum of the shifted scaling function defined in equation (3.7)

$$\psi(t) = \sum_n h_1(n)\sqrt{2}\,\phi(2t-n), \quad n \epsilon \mathbb{Z} \tag{3.8}$$

where $h_1(n)$ are a set of wavelet coefficients. By orthogonality the wavelet coefficients are related to the scaling function coefficients by

$$h_1(n) = (-1)^n h(1-n) \tag{3.9}$$

The function in equation (3.8) is what produces the general mother wavelet function for a class of expansion functions in equation (3.6).

3.2 Discrete Wavelet Transform

The goal of most expansions of a function or signal is to have the coefficients of the expansion provide more useful information about the signal than can be gleaned from the signal itself. To efficiently represent continuous functions requires finite, or

discrete approximations. Using the general scaling and mother wavelet expansion functions from equations (3.5) and (3.6), any function or signal $g(t)$ can be written as a series expansion in terms of the scaling function and mother wavelets

$$g(t) = \sum_k c_{j_0}(k)\phi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_j(k)\psi_{j,k}(t) \tag{3.10}$$

where the first part of equation (3.10) is the low resolution, coarse approximation or trend. The second component of equation (3.10) is the high resolution detail or fluctuation. The value chosen for j_0 determines the coarsest scale spanned by $\phi_{j_0,k}(t)$. Therefore, j_0 could be zero, or any other integer. If no scaling function is used, j_0 would be set to negative infinity. The coefficients of this wavelet expansion constitute the *discrete wavelet transform* (DWT) of the signal $g(t)$. If the wavelet system is orthogonal, the coefficients can be calculated by the following inner products

$$c_j(k) = \langle g(t), \phi_{j,k}(t) \rangle$$

$$d_j(k) = \langle g(t), \psi_{j,k}(t) \rangle$$

The orthonormality of the scaling functions and wavelet functions is important because it allows application of Parseval's theorem. The Parseval relation for the norm of $g(t)$ shows how the energy of the signal $g(t)$ relates to each component and their corresponding wavelet coefficients. Parseval's theorem can be expressed as a general wavelet expansion of equation (3.10) [5]

$$||g(t)||^2 = \sum_{t=-\infty}^{\infty} |g(t)|^2 = \sum_{n=-\infty}^{\infty} |c(n)|^2 + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} |d_j(k)|^2 \tag{3.11}$$

This equation provides confirmation that the energy of a signal is distributed to the coefficients of the expansion. In other words, the discrete wavelet transform process conserves energy by preserving the energy in the time domain in the wavelet domain [9]. See Appendix A for an explicit proof of Parseval's Theorem for discrete wavelet transforms.

3.2.1 Haar Transform

One of the strengths of wavelet transforms is the multitude of wavelets that exists. This allows one to select the wavelet type which works best for the particular situation being analyzed. To better understand the discrete wavelet transform (DWT), this section details Haar wavelets. Developed by Alfred Haar in 1910, Haar wavelets are the most basic type of wavelet.

Haar wavelets make up the basis for the Haar transform. The Haar transform is the basic model for all other transforms. The scaling function for the Haar wavelet is a straightfoward step function that is equal to 1 on the interval [0,1) and zero everywhere else.

$$\phi(t) = \begin{cases} 1 & \text{if } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

Figure 3.3a is a graphical representation of this step function. Equations (3.8) and (3.9) require that the wavelet function take the form

$$\psi(t) = \begin{cases} 1 & \text{for } 0 < t < 0.5 \\ -1 & \text{for } 0.5 < t < 1 \\ 0 & \text{otherwise} \end{cases}$$

In our project, the data is in the form of discrete signals with length N whose values occur at discrete instances in time or space. To simplify, we assume that the time increment separating successive discrete values is always the same. The term *equally spaced sample values* or just *sample values* will be the descriptor for this class of discrete signal values. For processing, N must be a positive even integer, in other words, a power of 2. If N is not a power of 2, zeroes can be added to the end of the signal to make it a power of 2. The Haar transform process decomposes a discrete

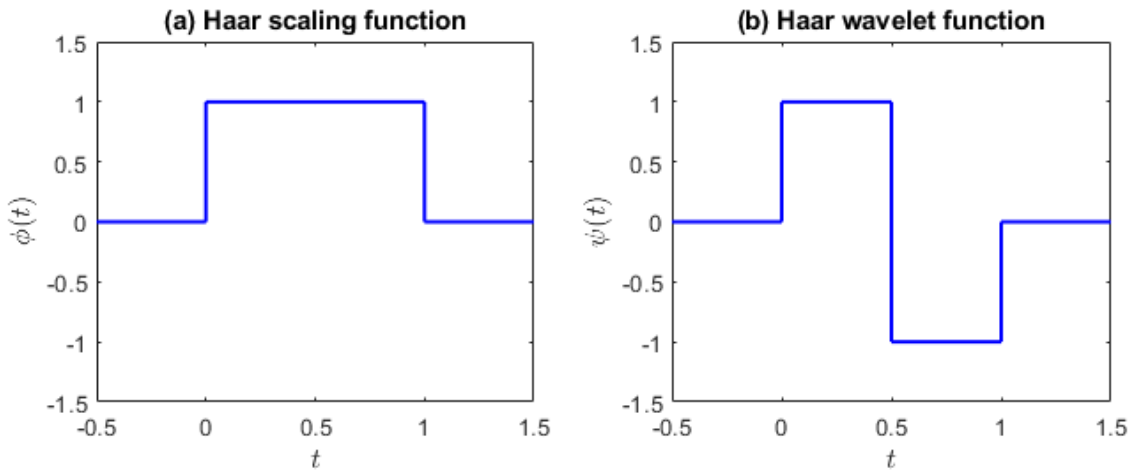

Figure 3.3: Plot (a) on the left is the Haar scaling function $\phi(t)$. Plot (b) on the right is the Haar wavelet $\psi(t)$. Both functions are defined on interval [0,1).

signal into two subsignals, each half the length of the original. One subsignal is a running average, or trend. The other subsignal is a running difference, or fluctuation.

As an example, we will perform a 1-level Haar transform on the discrete signal s =[2 6 7 11 15 5 4 2] with length $N = 8$, a positive integer. We'll start with the first average, or trend $A_1 = (a_1, a_2, a_3, a_{N/2})$. To compute the first coefficient, a_1 we take the average of the first pair of values in the signal, then multiply by $\sqrt{2}$. The reason for the multiplication by $\sqrt{2}$ is to make sure the Haar transform conserves energy in the signal. To proceed, we take the average of the next pair of values in the signal and again multiply by $\sqrt{2}$. This pattern of generating the average coefficients is represented by the following equation

$$A_m = \frac{s_{2m-1} + s_{2m}}{2}\sqrt{2} = \frac{s_{2m-1} + s_{2m}}{\sqrt{2}} \tag{3.12}$$

for $m = 1, 2, 3, ..., N/2$.

From this it is easy to see that

$$a_1 = \frac{2+6}{\sqrt{2}} = 4\sqrt{2}$$

$$a_2 = \frac{7+11}{\sqrt{2}} = 9\sqrt{2}$$

$$a_3 = \frac{15+5}{\sqrt{2}} = 10\sqrt{2}$$

$$a_4 = \frac{4+2}{\sqrt{2}} = 3\sqrt{2}$$

So the first trend subsignal $A_1 = (4\sqrt{2}, 9\sqrt{2}, 10\sqrt{2}, 3\sqrt{2})$.

The next subsignal is called the difference or first fluctuation of the signal s which will be designated $D_1 = (d_1, d_2, d_3, d_{N/2})$. The calculation of the difference coefficients differs from the average coefficient calculation in that we take half the difference of each pair of values in the signal then multiply by $\sqrt{2}$.

$$D_m = \frac{s_{2m-1} - s_{2m}}{2}\sqrt{2} = \frac{s_{2m-1} - s_{2m}}{\sqrt{2}} \tag{3.13}$$

for $m = 1, 2, 3, ..., N/2$.

The difference coefficients for the first fluctuation are computed as follows

$$d_1 = \frac{2-6}{\sqrt{2}} = -2\sqrt{2}$$

$$d_2 = \frac{7-11}{\sqrt{2}} = -2\sqrt{2}$$

$$d_3 = \frac{15-5}{\sqrt{2}} = 5\sqrt{2}$$

$$d_4 = \frac{4-2}{\sqrt{2}} = \sqrt{2}$$

So the first fluctuation subsignal is $D_1 = (-2\sqrt{2}, -2\sqrt{2}, 5\sqrt{2}, \sqrt{2})$.

Now that we have completed a 1-level Haar transform, the procedure can be repeated to perform multi-level Haar transforms. The 1-level Haar transform produced the first average A_1 and the first difference $D1$. To perform the 2-level Haar transform we compute a second average A_2 and a second difference $D2$, but only on the first trend A_1. The previous difference D_1 is copied into the next level to keep the signal the same size. Using equation (3.12) it is straight forward to calculate the second average $A_2 = (a_1, a_2)$, where the coefficients a_1 and a_2 are unique to this second Haar level.

$$a_1 = \frac{4\sqrt{2} + 9\sqrt{2}}{\sqrt{2}} = 13$$

$$a_2 = \frac{10\sqrt{2} + 3\sqrt{2}}{\sqrt{2}} = 13$$

Therefore $A_2 = (13, 13)$.

It follows that the second difference $D_2 = (d_1, d_2)$ is computed using equation (3.13)

$$d_1 = \frac{4\sqrt{2} - 9\sqrt{2}}{\sqrt{2}} = -5$$

$$d_2 = \frac{10\sqrt{2} - 3\sqrt{2}}{\sqrt{2}} = 7$$

Therefore $D_2 = (-5, 7)$.

The process continues until you reach a trend node of size 1. The full Haar transform decomposition of signal s is shown in table (3.1).

2	6	7	11	15	5	4	2
$4\sqrt{2}$	$9\sqrt{2}$	$10\sqrt{2}$	$3\sqrt{2}$	$-2\sqrt{2}$	$-2\sqrt{2}$	$5\sqrt{2}$	$\sqrt{2}$
13	13	-5	7	$-2\sqrt{2}$	$-2\sqrt{2}$	$5\sqrt{2}$	$\sqrt{2}$
$13\sqrt{2}$	0	-5	7	$-2\sqrt{2}$	$-2\sqrt{2}$	$5\sqrt{2}$	$\sqrt{2}$

Table 3.1: 3-level Haar transform. The first row represents the original signal [2 6 7 11 15 5 4 2] being processed. Each row is a different Haar level. So, the second row is Haar level 1, the third row is Haar level 2, and the fourth row is Haar level 3, the final level for this transform. For the subsequent rows, the black numbers are the average coefficients (trend) and the red numbers are the difference coefficients (fluctuation).

The second row of table 3.1 represents a Haar level 1 transform, or one level of decomposition. The frequency node in level 1 (second row) containing the four black numbers represents the low-pass (low frequency) coarse component of the approximation. The black numbers are referred to as the *average* of that frequency. The frequency node containing the four red numbers in the level 1 transform represent the high-pass (high frequency) detail version of the approximation. The red numbers can also be thought of as the local *fluctuations* at that frequency. The subsequent row, row 3 is a Haar level 2 transform. This produces an average that is coarser than the previous level, resulting in a frequency node containing two black numbers [13 13]. The two red numbers in row 3 of the Haar level 2 transform represent a narrower frequency node and thus more detailed approximation than the previous level.

3.3 Discrete Wavelet Packet Transform

The DWPT is the primary analysis method used in this project. Both wavelet transforms and wavelet packet transforms are time-frequency tools which decompose a discrete signal in the time-frequency domain to obtain a good resolution in

time as well as frequency. At the first level, both tools decompose the signal into coefficients, the low frequency trend and the high frequency fluctuation. However, the difference between these two is the way they decompose the signal after the first level. The wavelet transform decomposes only the low frequency trend components at each level, whereas the wavelet packet transform decomposes both the low frequency trend and the high frequency fluctuation components at each level. As a result, the wavelet packet transform produces better resolution when the signal also contains high frequency information. The decomposition of the signal into multiple frequency nodes can be stopped at any level. Increasing the number of levels of decomposition increases the frequency resolution allowing the user to customize this setting to the needs of the application. Figure 3.4 shows a level-3 decomposition which generates 8 wavelet packet coefficients. As mentioned above, it is viable to use the results from level 1 or level 2 of the decomposition tree. The choice of level to use is determined by the needed frequency resolution.

As an example, we will decompose the signal s =[2 6 7 11 15 5 4 2] from section §3.2.1 using the DWPT method.

Both wavelet transforms and wavelet packet transforms use wavelets as a basis for signal processing. This example uses the Haar wavelet as the basis for the calculation. The DWPT is calculated by performing a first level Haar transform on all subsignals, both the trends and fluctuations. The first level Haar transform and fist level DWPT are identical. The second level is where the decomposition is noticeably different. For the second level, start with the first node of the level one transform from table 3.1: $[4\sqrt{2}\ \ 9\sqrt{2}\ \ 10\sqrt{2}\ \ 3\sqrt{2}]$, and perform another level 1 Haar transform. This produces the first two nodes in the second level. The first node contains the trend coefficients [13 13], and the second node contains the fluctuation coefficients [-5 7]. Next, take the second node of the level one

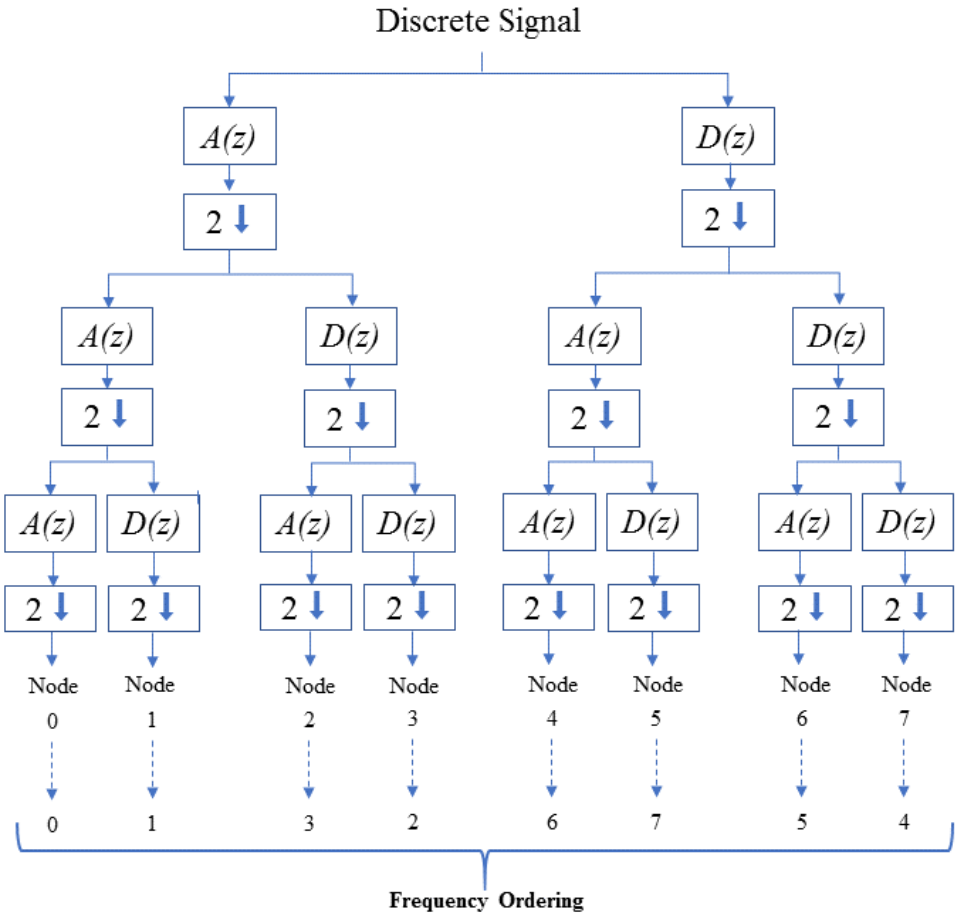

Figure 3.4: Wavelet packet tree for three levels of wavelet packet decomposition. In this case $A(z)$ and $D(z)$ represent the average (trend/low pass) and difference (fluctuation/high pass) filters, respectively. The down arrows represent the down sampler, or decimator by 2. After decomposition, the coefficients of the wavelet packet transform in natural order. Revision from natural order to the correct frequency order is done by applying grey code permutation.

transform from table 3.1: $[-2\sqrt{2} \quad -2\sqrt{2} \quad 5\sqrt{2} \quad \sqrt{2}]$, and perform another level 1 Haar transform. This produces the first third and fourth nodes in the second level. The third node contains the trend coefficients [-4 6], and the second node contains the fluctuation coefficients [0 4]. The full signal decomposition using the DWPT is shown in table 3.2.

2	6	7	11	15	5	4	2
$4\sqrt{2}$	$9\sqrt{2}$	$10\sqrt{2}$	$3\sqrt{2}$	$-2\sqrt{2}$	$-2\sqrt{2}$	$5\sqrt{2}$	$\sqrt{2}$
13	13	-5	7	-4	6	0	4
$13\sqrt{2}$	0	$\sqrt{2}$	$-6\sqrt{2}$	$\sqrt{2}$	$-5\sqrt{2}$	$2\sqrt{2}$	$-2\sqrt{2}$

Table 3.2: 3-level Discrete Wavelet Packet Transform using Haar wavelets. The first row represents the original signal [2 6 7 11 15 5 4 2] being processed. Each row is a different Haar level. For the subsequent rows, the black numbers are the average coefficients (trend) and the red numbers are the difference coefficients (fluctuation). Each level is grouped by frequency nodes. The final level is in natural order.

It is worth noting that when performing a DWPT, the frequency nodes in the final level of decomposition are not automatically in frequency order. Returning to the example in table 3.2, the nodes in the last row are in what is called *natural order*. To get the nodes in frequency order, a gray code ordering algorithm must be applied.

The frequency ordering is applied to the last "requested" level of decomposition. The term "requested" level is used because the decomposition can be stopped at any level, depending on the level of resolution you're looking for. The signal in 3.2 was fully decomposed for a total of 3 levels. The frequency order is then determined by running the graycode algorithm, supplying the level as the argument n. In this case where $n = 3$ the graycode method returns [0 1 3 2 6 7 5 4]. This vector contains the order of indexes which will result in all frequency nodes being

in frequency order, with the index starting at zero. Applying this to the final level of decomposition of the signal in table 3.2, the frequency order becomes

$$[13\sqrt{2} \mid 0 \mid -6\sqrt{2} \mid \sqrt{2} \mid 2\sqrt{2} \mid -2\sqrt{2} \mid -5\sqrt{2} \mid \sqrt{2}]$$

What if the decomposition is set to stop after the second level? Two levels of DWPT processing of signal s produces four frequency nodes, each containing two coefficients. This is shown in row 3 of table 3.2 and copied again below

$$[13 \;\; 13 \mid -5 \;\; 7 \mid -4 \;\; 6 \mid 0 \;\; 4]$$

At level 2 the frequency order can be determined by setting $n = 2$ and running the graycode function yielding [0 1 3 2]. Use this gray code ordering to put the level 2 frequency nodes in order to obtain

$$[13 \;\; 13 \mid -5 \;\; 7 \mid 0 \;\; 4 \mid -4 \;\; 6]$$

The MATLAB wavelet toolbox contains a discrete wavelet packet transform function called **dwpt**, which automatically performs the frequency ordering.

3.4 Wavelet Transform vs Fourier Transform

As discussed in §1.2.1 Fourier analysis is the traditional, older technique to perform spectral analysis. Another tool for spectral analysis is the DWPT which was introduced in this chapter. The DWPT technique of computing the power spectrum has some distinct advantages over its Fourier counterpart.

First, for the Fourier transform the basis functions are limited to sines and cosines. However, when using the DWPT there are an infinite number of wavelets which can be chosen as basis functions. This provides the ability to choose the wavelet which best approximates the signal being analyzed. Second, the Fourier

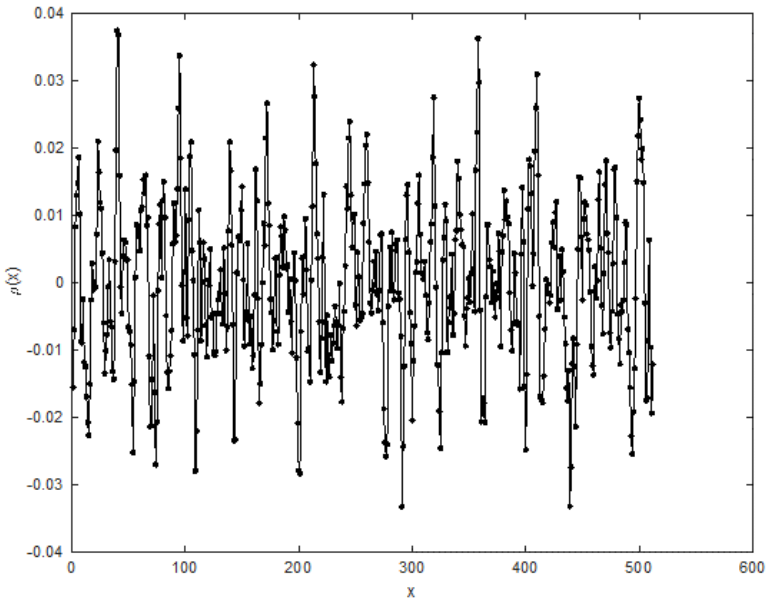

Figure 3.5: Sample spectrum with Gaussian random noise added.

transform maps into a 1-dimensional array of coefficients while the DWPT maps into a 2-dimension array of coefficients. This makes the DWPT localized in both space (or time) as well as frequency. The result of this is that errors are also localized. Contrarily, the Fourier transform is localized in frequency only. This causes errors to be propagated throughout the entire signal approximation. Next, we will explore an example which demonstrates the advantages in signal processing that the DWPT has over the conventional Fourier method.

Figure 3.5 displays a sample spectrum produced with simulated data containing Gaussian random noise (0 mean, unit variance) with a sampling frequency of 1 Hz. The sample is based on the function with a power spectrum,

$$P(k) = \frac{k}{(1+k^4)} \tag{3.14}$$

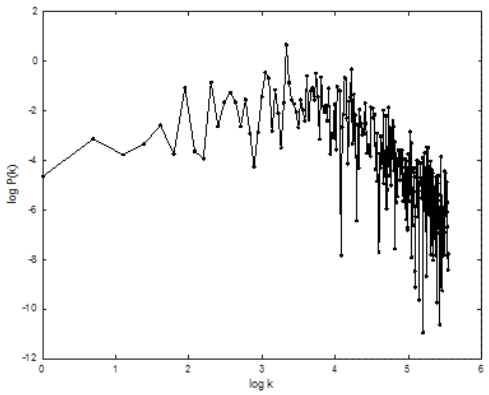

Figure 3.6: Fourier power spectrum of sample.

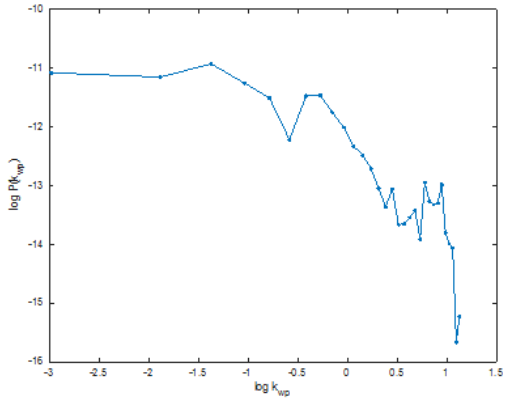

Figure 3.7: Wavelet packet power spectrum of sample.

Applying the Fourier power spectrum and the wavelet power spectrum to the sample signal produced the result in Figures 3.6 and 3.7. The shape of the curves is similar, but when using the wavelet method there is much less noise. The x-axis in the Fourier case is traditionally known as frequency or scale. The Fourier method has higher frequency resolution, however at the expense of picking up a lot of noise in the process. For the wavelet power spectrum, the frequency is 1/node, where node is a wavelet packet node, or frequency band. The frequency bands go from 0 to π, and are evenly distributed on each node.

In figures 3.8 and 3.9 60 bad data points have been introduced into the original sample and the values where changed to the mean of the signal $+5\sigma$. After computing the power spectrum using both methods, the general shape of the curve is maintained. However, the Fourier version has changed by 2 orders of magnitude while the wavelet version remains relatively unchanged. Since the basis functions of the Fourier methods are infinite (sine and cosine waves), any errors will propagate throughout the entire estimate, which includes all frequencies. The DWPT method has the advantage of using finite wavelets as basis functions and being localized in both space(time) and frequency. As a result, errors in the data are lim-

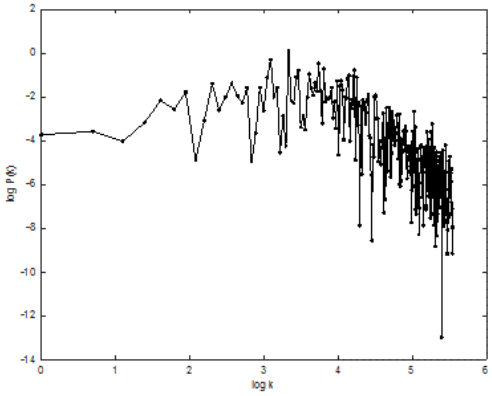

Figure 3.8: Fourier power spectrum of sample with 60 bad data points added.

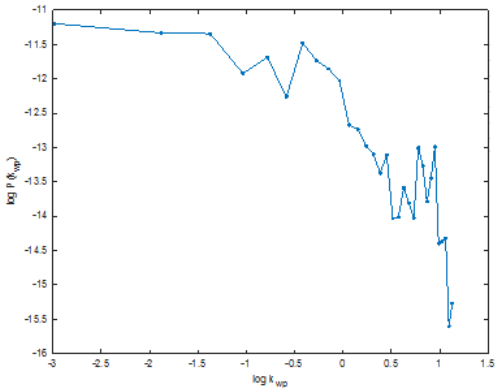

Figure 3.9: Wavelet packet power spectrum of sample with 60 bad data points added.

ited the branch of descendant frequencies in the wavelet packet tree (see Figure 3.4).

The Fourier transform has better frequency resolution than the DWPT, but this comes at the expense of picking up high frequency noise that will tend to "wash out" the transform. The DWPT does not have this issue because it is capturing information over a frequency bin.

3.5 Power Spectrum Estimation

In this section we provide an overview of the entire process of computing the power spectrum. We describe how we applied the recommended data corrections from the value added SDSS catalogs to ensure extraction of only Lyman-α spectra from the BOSS DR9 dataset. The data binning algorithm is fully explained followed by the details of calculating the energy and power spectrum from the DWPT.

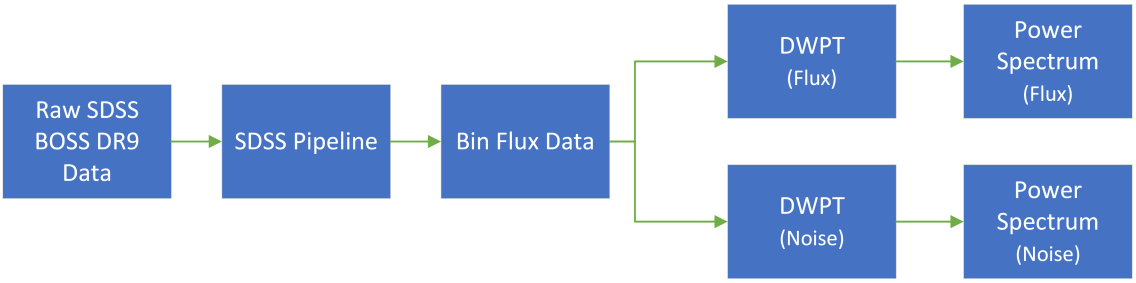

Figure 3.10: An overview flowchart of our process to calculate the wavelet packet power spectrum.

3.5.1 Process Overview

Figure 3.10 is a flowchart of our process. We begin with the SDSS BOSS D9 data described in §2.3.2 and apply the SDSS recommended masks and corrections from the SDSS value added catalogs described in §3.5.2. We then bin the flux data by redshift, while adjusting the bin size due to the expanding universe (see §3.5.3). The data is then grouped by redshift range as described in §3.5.3. We then apply the DWPT described in §3.3 to the flux and noise data. Finally, in §3.5.4 we use the DWPT transform results to calculate the wavelet packet power spectrum described in §3.5.

3.5.2 Application of pipeline corrections to DR9 dataset

The quasar spectral data is provided in multiple Flexible Image Transport System (FITS) files. These individual "speclya" spectral files have a naming convention which embeds the following information: plate number, fiber ID, and Julian date. These three pieces of information are needed to identify the redshift of the quasar. The speclya files are accompanied by additional mask and corrections as listed in table 3.3. The value-added masks and corrections have not been applied to the pipeline flux or variances. To extract Lyman-α forest transmission and

corresponding pixel error we perform the following data clean up procedure on each FITS file.

The quasar's redshift is not in the same file as the rest of the quasar spectral data, so we needed to use the plate number, fiber ID, and Julian date to locate the redshift for the quasar in the **sortedkeys** file. Once the redshift is determined, we verify that the spectra is indeed a Lyman-alpha absorption spectra from a hydrogen cloud. In this data set, the maximum allowed Lyman-α forest interval is 1041Å-1185Å restframe with respect to redshift. Redshift is given by the formula

$$z \equiv \frac{\lambda_{ob} - \lambda_{em}}{\lambda_{em}} \tag{3.15}$$

where λ_{ob} is the *observed wavelength*, and λ_{em} is the wavelength emitted from the source. Solving for λ_{ob} yields

$$\lambda_{ob} = z\lambda_{em} + \lambda_{em} \tag{3.16}$$

We define the Lyman-α forest region by the range $1041 < \lambda_{rest} < 1185$ Å. This range was chosen because it avoids quasar Lyman-α (1216) and Lyman-β (1026) emission lines where continuum fitting is more difficult due to large gradients (by 8000 km/s and 5000 km/s respectively). Also, the chosen range avoids the quasar proximity zone on the red end, to avoid contamination of the power spectrum by astrophysical effects in the vicinity of the quasar [6].

Using equation 3.16 the minimum and maximum allowed observable wavelength values for the Lyman-α forest can be computed as follows

$$\lambda_{min} = z_\alpha(1041\,\text{Å}) + 1041\,\text{Å}$$

$$\lambda_{max} = z_\alpha(1185\,\text{Å}) + 1185\,\text{Å}$$

Where z_α is the redshift of the quasar. This range was then used to filter out quasar absorption spectra that are not due to Lyman-α.

Table 3.3: Spectral Products in HDU 1 of 'SPECLYA' Product Offering

Standard Pipeline Products	
FLUX	Coadded calibrated flux density in units of $10^{-17} erg\,s^{-1}\,cm^{-2}\,\mathring{A}^{-1}$
LOGLAM	Logarithm base 10 of wavelength in angstroms
IVAR	Inverse variance of the flux
AND_MASK	AND mask
OR_MASK	OR mask
WDISP	Wavelength dispersion in dloglam units
SKY	Subtracted sky flux in units of $10^{-17} erg\,s^{-1}\,cm^{-2}\,\mathring{A}^{-1}$
MODEL	Pipeline best model fit used for classification and redshift
Value-Added Products	
MASK_COMB	Combined mask incorporating pipeline AND_MASK, sky-line masks, and DLA masks
DLA_CORR	Flux correction for known DLAs
NOISE_CORR	Pipeline noise corrections
CONT	Estimated quasar continuum in 1040-1600 Årestframe, in units of $10^{-17} erg\,s^{-1}\,cm^{-2}\,\mathring{A}^{-1}$

In addition to the primary SDSS photometry and spectroscopy data, several value-added catalogs (VACs) were included in the BOSS DR9 release. The VACs contain masks and corrections must be applied to the fluxes and variances before analyzing the data to ensure unrelated fluxes are filtered out. The following corrections were applied according to instructions provided by SDSS[1]. DLA_CORR is the flux correction for known DLAs. CONT is the estimated quasar continuum in 1040-1600 Å restframe in units of $10^{-17} ergs/s/cm^2/$Å. The corrected Lyman-alpha

[1] `https://dr16.sdss.org/datamodel/files/BOSS_LYA/cat/speclya.html`

transmission $\Psi_{Ly\alpha}$ can be calculated using the formula

$$\Psi_{Ly\alpha} = FLUX\left(\frac{DLA\_CORR}{(CONT)(RESIDCORR)}\right) \tag{3.17}$$

where RESIDCORR is interpolated from file residcorr_v5_4_45.dat, which is available for download on the SDSS website. We find the last RESIDCORR value in the RESIDCORR column that is less than or equal to the current wavelength. We also take into account that the wavelengths in the SDSS data are $\log_{10}$(wavelength), with the wavelength in units Å.

3.5.3 Data binning

After filtering unrelated fluxes as described near the end of §3.5.2, the true Lyman-α fluxes are averaged for each bin. For the binning algorithm we first create four bin groups by redshift ranges $2.0 < z < 2.2$, $2.2 < z < 2.4$, $2.4 < z < 2.6$, and $2.6 < z < 2.8$. Because the universe is expanding, the bin size will change according to the redshift. This is accounted for in each redshift group by calculating the bin size based on the median redshift value of the group. As an example, for redshift group with range $2.2 < z < 2.4$, the median redshift is 2.3. Using equation (1.11) calculate the observed wavelength that corresponds to the median redshift of the group range

$$\lambda_{obs} = \lambda_{rest}(1+z) = 1216\text{Å}(1+2.3) = 4013 \text{ Å}$$

The relationship between the observed wavelength λ_0, changing emitted wavelength $d\lambda$, and changing velocity dv is

$$dv = \frac{c\, d\lambda}{\lambda_{obs}} \tag{3.18}$$

Here $dv = 69\, km/s$, which corresponds to the pixel resolution of the spectrographs. All pixels have the same resolution. Using equation (3.18) the bin size for redshift

range $2.2 < z < 2.4$ is

$$d\lambda = \frac{(dv)(\lambda_{obs})}{c} = \frac{(69/10^{-13})\,(4013)}{(3 \times 10^5/10^{-13})} = 0.9229 \text{ Å}$$

The 10^{-13} factors are converting the values from km/s to Å/s.

Next in the binning process is calculating the number of bins in the redshift range N_{bz}.

$$N_{bz} = \frac{\lambda_{max} - \lambda_{min}}{d\lambda} \tag{3.19}$$

Where λ_{max} and λ_{min} are the maximum and minimum wavelength of the redshift range respectively. Therefore, for range $2.2 < z < 2.4$, the total number of bins in the range is

$$N_{bz} = \frac{4134 - 3891}{0.9229} \sim 263 \text{ bins} \tag{3.20}$$

Next we determine which of the N_{bz} bins the quasar flux should be placed in, which we call the bin location $binL$. The bin location is defined as

$$binL = BG_{start} + \frac{(\lambda_{obs} - \lambda_{min})}{d\lambda} \tag{3.21}$$

Where BG_{start} is index of the first bin of the redshift bin group range, and λ_{obs} is again the quasar observed wavelength. Now that we know the size of the bins in the range and which bin the quasar belongs to, the quasar's flux is added to the total flux in that bin. This process is repeated until all quasar fluxes are properly binned, resulting in bins containing the total flux contribution from all quasars in each bin.

3.5.4 Wavelet Packet Power Spectrum Estimation

To measure the one-dimensional power spectrum, the quasar absorption spectrum are cast in velocity units, where all pixels in are the same size $\Delta v = 69$ km/s. We

therefore use velocity instead of observed wavelength in our results.

As described in section 3.2, Parseval's theorem also applies to the wavelet packet transform. Therefore, the power (P_m) of a wavelet packet node m (or frequency node) can be defined as

$$P_m(watt) = \frac{E_m(Joule)}{2^j} \tag{3.22}$$

Where E_m is the energy of the wavelet packet node.

$$E_m = \sum_{i=1}^{N/2^j} node\_coef(i)^2 \tag{3.23}$$

where $node_coef(i)$ represent the individual DWPT coefficients of the frequency node. The value j is the desired decomposition level which makes 2^j the number of frequency nodes (also called wavelet packet nodes or sampling nodes) at level j. This leads to a definition for the power spectrum density in the m^{th} frequency band (PSD_m) [9]

$$PSD_m(watt\,/\,Hz) = \frac{P_m(watt)}{2^j} \tag{3.24}$$

CHAPTER 4

Analysis of the Lyman-α Forest

In this chapter, we will present the results, analyze our data, and compare to previous results.

4.1 Results

Using the procedures described in §3.3 and §3.5, we compute the wavelet packet power spectrum and group the results into four redshift bins (see figure 4.1). We have plotted the dimensionless quantity $\Delta^2(k) \equiv \pi^{-1}\, k\, P(k)$, which is the contribution to the variance per unit ln k [18]. The horizontal axis is the wavenumber based frequency space $2\pi/v$ in units of $[km/s]^{-1}$ with the highest $k - mode$ possible defined by the Nyqvist-Shannon limit $k[Nyqvist] = \pi/\Delta v$.

The error bars represent the standard error, with the variance found as the spectral noise in the Lyman-α flux data. Constant comoving bin size of 69 km/s produced 1,030 bins. Each Lyman-α forest flux was analyzed and placed in the correct bin according to its redshifted wavelength. We then take the average flux for the entire bin, and obtain the variance from the average. The wavelength bin size changes depending on the redshift in order to keep a constant velocity bin with of 69 km/s.

The results of the applying our wavelet packet power spectrum tool to the SDSS DR9 dataset are plotted in Figure 4.1. The k values on the x-axis represent

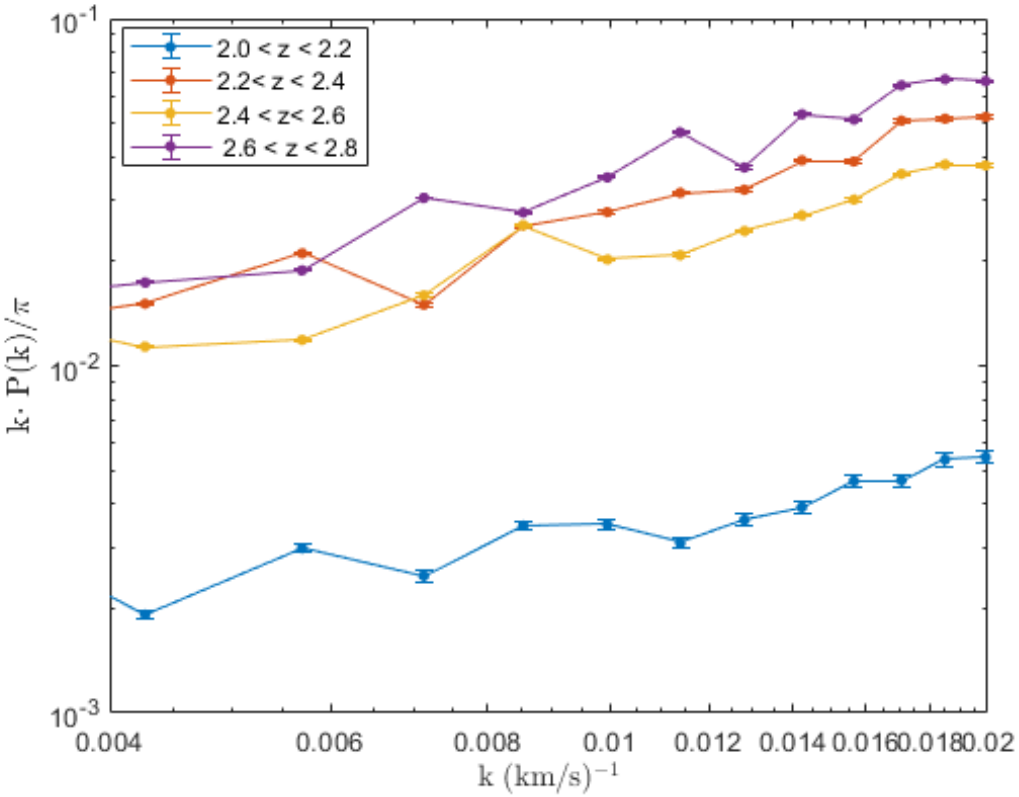

Figure 4.1: One-dimensional Lyman-α forest wavelet packet power spectrum. Error bars are the standard deviation of the power spectrum of the noise from the SDSS DR9 Lyman-α flux data produced by quasars. The lines show the true power value. Redshift range, from bottom to top, are: blue line $2.0 < z < 2.2$, red line $2.2 < z < 2.4$, yellow line $2.4 < z < 2.6$, purple line $2.6 < z < 2.8$.

the size of the fluctuations in velocity space. The power spectrum on the y-axis represents $\Delta^2(k) \equiv \pi^{-1} k P(k)$, where the power spectrum density $P(k)$ is multiplied by $k\pi^{-1}$ to make it a dimensionless contribution to the variance. The plot is logarithmic and the fluctuations follow a power law, so small scales are towards the right side of the graph where there are many fluctuations. There are fewer fluctuations for larger scales, which are towards the left side of the graph.

Aside from the recommended pipeline noise and mask corrections, we did not filter out any data. We used the entire dataset of 54,468 quasars. We then let the wavelet packet power spectrum analyze the data, filter out the noise, and tell us the story of density fluctuations and patterns of large-scale structure.

Figure 4.1 shows our wavelet packet power spectrum results. Previous studies have shown there is an increase in power as you go through smaller and smaller scales on this logarithmic plot. This is consistent with the hierarchical model of structure formation where many small-scale structures merged to form larger scale structures. The figure also shows that power generally increases with increasing redshift. It is known that the flux power spectrum evolves strongly with redshift. As discussed in detail by (Croft et al.) [7] this is due to the increased optical depth at high redshift.

Note that while the data from redshift range $2.4 < z < 2.6$ follow the general trend of $2.0 < z < 2.2$ and $2.6 < z < 2.8$, it has anomalous features when compared to these two redshift ranges. We hypothesize that a reason for this can be explained by Figure 4.2, which shows the redshift distribution of the high redshift quasars. At low redshift between 2 and 2.2, the number density of quasars does not change much. However, from redshift 2.25 to 2.6 there is a sharp drop off. For the middle redshift bins the number of quasars contributing sharply decreases.

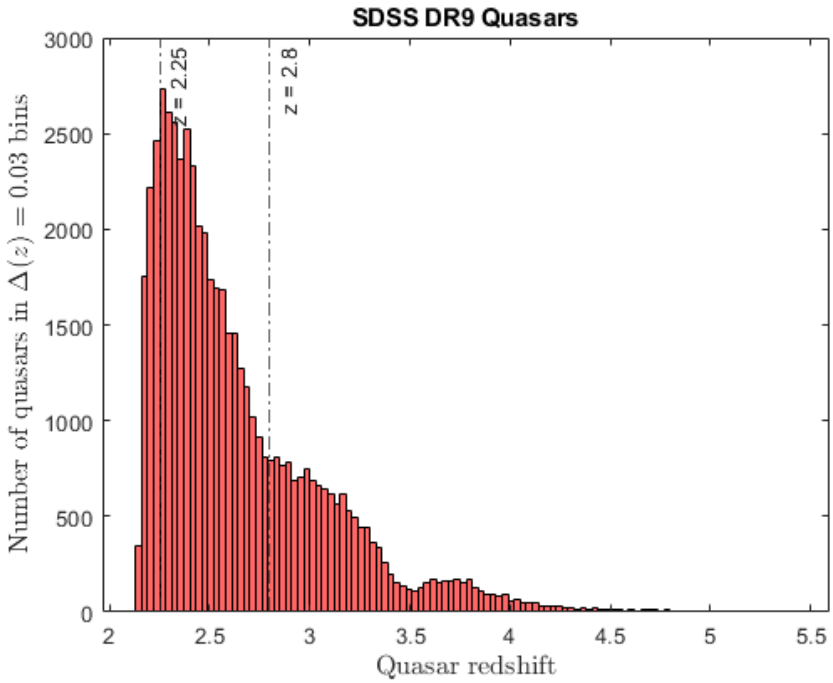

Figure 4.2: This is a copy of Figure 2.5 with vertical lines marking the redshift bin locations $z = 2.25$ and $z = 2.8$ to guide the reader's eye to the location where the number density of quasars changes the most drastically.

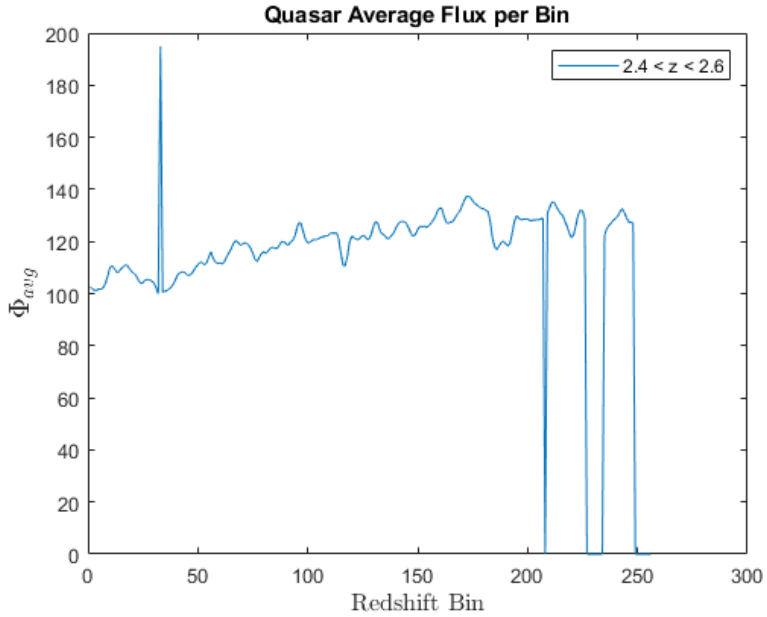

Figure 4.3: Binned Lyman-α forest flux data for redshift range $2.4 < z < 2.6$. There is inconsistent data, both high and low flux, which probably does not represent the sample. At (Bin= 33, $\Psi_{avg} \sim 195$) there is a spike in flux that much higher than all others. Towards the far right end of the graph there are several areas were the flux abruptly drops to zero.

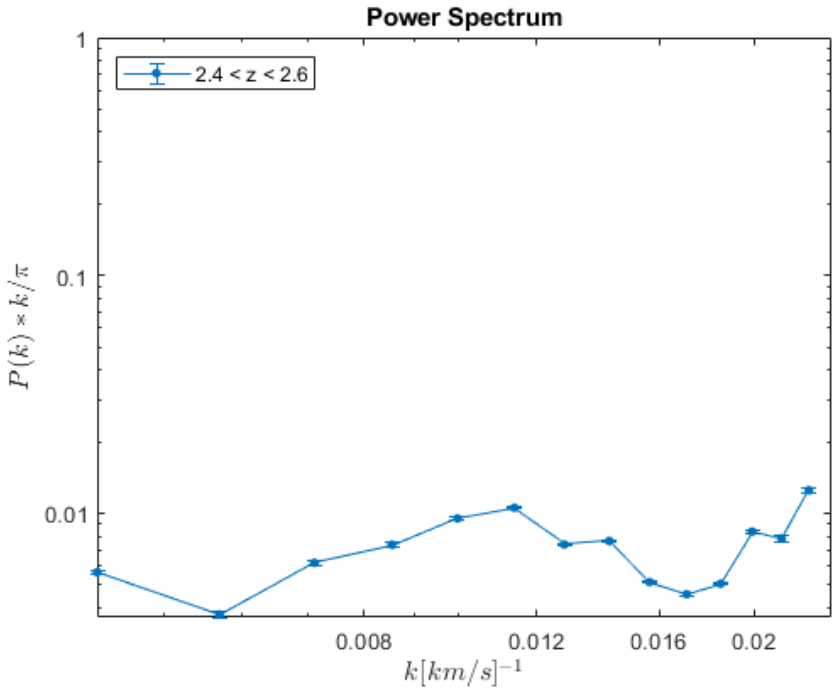

Figure 4.4: Power spectrum of Lyman-α forest flux data for redshift range $2.4 < z < 2.6$.

To rephrase it, the number of quasars contributing to the flux in each bin changes drastically in a small redshift interval. Nevertheless, the wavelet packet power spectrum still captures the general trend without having to sacrifice any of the data. The power increases as you go to smaller scales and the amplitude of the power increases as a function of redshift.

To further illustrate this point, we isolate the flux data in range 2.4-2.6 and show the average flux per bin in Figure 4.3. Note that there are outliers in the data. At bin 33 for example, the flux which is averaging around 100, suddenly jumps to 195. While at bins greater than 200 there are several sections where no flux is being contributed at those bins. Still, we calculate the wavelet packet power spectrum even though this data contains questionable points. The results

are shown in Figure 4.4. Once again, at the smallest frequencies there is a general increase in power from $k = 0.0100$ to $k = 0.0228$. The resulting power spectrum in Figure 4.4 shows that the wavelet packet power spectrum is still able to recover despite the data being potentially problematic. Even with questionable outliners in the data, any effect they have stays local and the power spectrum is not affected too much.

We did not investigate which quasar files were causing these outliers in the data. The purpose of this thesis was to determine if the wavelet packet power spectrum could recover regardless of any potential problems in the data. We need to point out that in all published results of the power spectrum as will be discussed shortly, the majority of the Lyman-α forest is discarded due to problems in the data.

4.2 Previous Results

4.2.1 Palanque-Delabrouille et al. 2013

Palanque-Delabrouille et al. 2013 [21] developed a method based on the Fourier transform to measure the one-dimensional power spectrum of the transmitted flux in the Lyman-α forest. They applied their method to a subset of the SDSS-III BOSS DR9 dataset. Of the over 50,000 spectra available in BOSS DR9, only 13,821 quasar spectra where selected based on high quality, high signal-to-noise ratio (S/N), and good spectral resolution. Their results spanned redshift bins from $z = 2.2$ to $z = 4.4$ and are shown in Figure 4.5.

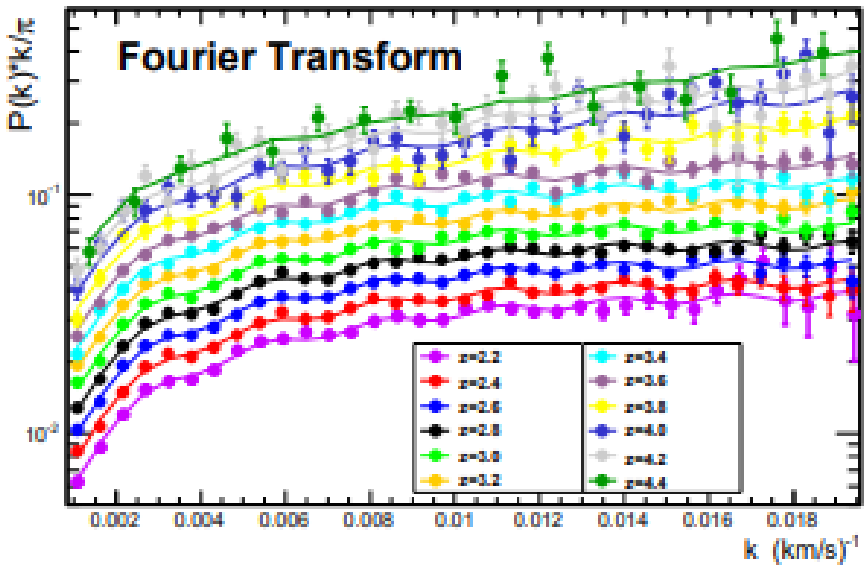

Figure 4.5: One-dimensional Lyman-α forest power spectrum obtained with the Fourier transform method. Figure from Palanque-Delabrouille et al. 2013 [21]

The final power spectrum data is rebinned onto a predefined grid in k-space (frequency), giving equal weight to the different Fourier modes that enter each bin. This acts to minimize the size of the error bars. The power spectrum calculation includes several corrections which take into account pixel noise, the impact of the spectral resolution of the spectrograph. Finally, an estimate of the uncorrelated background due to metal absorption in the Lyman-α forest is completed by estimating the background component in sidebands positioned at longer wavelengths than the Lyman-α forest region.

4.2.2 Chabanier et al. 2019

(Chabanier et al. 2019) [6] measured the one-dimensional power spectrum using a Fast Fourier transform (FFT) method. They applied their method to quasar spectra from SDSS-III BOSS and SDSS-IV eBOSS DR14 surveys. From a total

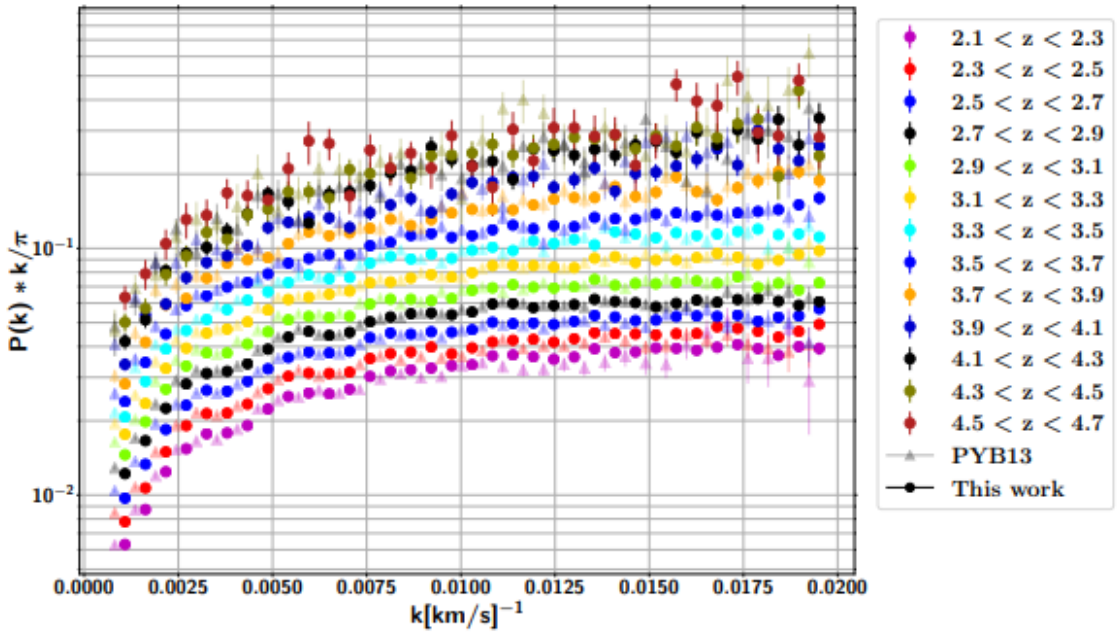

Figure 4.6: One-dimensional Lyman-α forest power spectrum obtained with the Fourier transform method. Figure from Chabanier et al. 2019 [6]

of 180,413 quasars, only 43,751 quasars were selected covering 13 bins spanning redshift range $z = 2.2$ to 4.6. After binning the Lyman-α forest absorbers according to mean redshift, the final power spectrum was rebinned onto an evenly spaced grid in k(frequency)-space. This assigned equal weight to the different Fourier modes that entered each bin. This rebinning effort is what makes the error bars in Figure 4.6 small. The power spectrum calculation includes several corrections which take into account pixel noise, the impact of the spectral resolution of the spectrograph, correlated absorption from either Si $_{I}II$ or Si $_{I}I$, and uncorrelated background noise due to metal absorption. Their results are show in Figure 4.6

4.3 Comparison with our results

Our results shown in Figure 4.1 are consistent with the previous results of (Palanque-Delabrouille et al. 2013) [21] and (Chabanier et al. 2019) [6]. The data lines up particularly well with the velocity space on the horizontal axis. Our power spectrum result on the vertical access is ~ 0.1 lower in magnitude than the two previous results. However, the relative shape of the graph matches up very well with both of the prior studies. Our results begin to flatten out after $k \sim 0.018$ similar to both of the previous studies. This provides evidence that the wavelet packet power spectrum is a valid technique to extract the power spectrum. Unlike the previous studies, our method requires less data manipulation. The wavelet packet method localizes noise and errors and allows us to use the entire dataset.

The rebinning of the power spectrum data in frequency space by both (Palanque-Delabrouille et al 2013) and (Chabanier et al. 2019) contributed to the small error bars in their results. In our results, our small error bars are due to the fact that the wavelet packet power spectrum took care of a lot of the noise for us, keeping our results closer to expected values.

4.4 Conclusion

In this thesis we developed a method of detecting LSS by measuring the one-dimensional power spectrum of the transmitted flux in the Lyman-α forest. This was a proof of concept which demonstrated that the wavelet packet power spectrum is a valid technique to extract the power spectrum. Proving that Parseval's Theorem for energy conservation holds for the DWPT provided a preliminary indication that this method had promise. Further investigation revealed that the wavelet packet power spectrum, which is powered by the DWPT, is able to capture

the important aspects of the power spectrum, but it does not do as good a job as Fourier at capturing the high frequency information. Unlike Fourier transform methods however, the wavelet packet power spectrum minimizes error propagation by containing noise within a limited number of frequency nodes rather than dispersing the error throughout the entire signal. It was also shown that despite the presence of irregular data or outliers, the wavelet packet power spectrum is still able to recover and produce meaningful results.

The decision was made to use the SDSS BOSS DR9 Lyman-α forest dataset because it was fully vetted, contained the high redshift data necessary for studying LSS, and was packaged with value-added products to assist in Lyman-α forest analysis. After using the value added products to extract the quasar Lyman-α forest lines we computed the wavelet packet power spectrum. The wavelet packet transform handles noisy data efficiently, which minimized data manipulation in our workflow. Contrary to previous studies, we did not filter out any data. Aside from the application of the recommended SDSS corrections to the data, we used the entire dataset of 54,468 quasars. We then let the wavelet packet power spectrum analyze the data, filter out the noise, and tell us the story of density fluctuations and patterns of large-scale structure.

Our results agree with previous studies which indicate increasing power moving towards smaller scales on the logarithmic plot. This is consistent with the *bottom-up* scenario where the first objects to form are the smallest with galaxies forming first, followed by clusters, then superclusters.[22] The trend of increasing power with increasing redshift also provides evidence to support there is increased optical depth at high redshift as designated by (Croft et al.).

4.5 Future Research

This study was a proof of concept so we limited the scope to only the SDSS DR9 dataset. By comparison to previous studies which used the traditional Fourier method we have established the proof of concept was successful. The first evident follow-up research path is to apply our wavelet packet power spectrum tool to the latest SDSS dataset which was produced by the extended Baryonic Oscillation Spectroscopic Survey (eBOSS) collaboration. The sixteenth and final eBOSS data release (SDSS DR16) was made available to the public in December of 2019. It contains all data from eBOSS and its predecessor, BOSS, providing 210,005 quasars with $z > 2.10$ that are used to measure Lyman-α absorption [17].

It is possible that the wavelet packet power spectrum can be used as a data clean up tool to prepare datasets for further processing. Since the wavelet localizes by time and frequency, it can isolate where bad data points are. The wavelet packet power spectrum can therefore be used as a way to denoise data as part of the pipeline effort similar to wavelet coefficient thresholding.

Additional research efforts could include using a different binning technique. Our study used a rounding technique to handle cases that were at the borders of bins. It could prove fruitful to investigating other ways to ensure data fluxes at the borders of redshift bins are placed in the bin that most realistically represents the sample. Also, including and plotting additional redshift ranges could produce more lines of investigation. Experimenting with different decomposition levels of the DWPT to see how or if they change the power spectrum result is worth pursuing. Finally, experimenting with different kinds of wavelets when running the DWPT or using the maximally overlapped discrete wavelet transform may yield other interesting ways to analyze LSS.

Appendices

APPENDIX A

Proof of Parseval's Theorem for Discrete Wavelet Transforms

Start with the general statement of the series expansion in terms of the scaling function and wavelet equation (3.10) copied here

$$g(t) = \sum_k c_{j_0}(k)\phi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_j(k)\psi_{j,k}(t) \tag{3.7}$$

where the first part of equation (3.10) is the low-resolution, coarse approximation or trend. The second component of equation (3.10) is the high-resolution detail or fluctuation. The value chosen for j_0 determines the coarsest scale spanned by $\phi_{j_0,k}(t)$. Therefore, j_0 could be zero, or any other integer. If no scaling function is used, j_0 would be set to negative infinity.

Starting with equation (3.10), let N be the length of the signal while J is the desired level of signal decomposition. Therefore, $N/2^J$ is the length of the wavelet packet node, or the number of coefficients per node at the current decomposition level.

$$g(t) = \sum_{k=1}^{N/2^J} c_{J,k}(t)\phi_{J,k}(t) + \sum_{j=1}^{J} \sum_{k=1}^{N/2^J} d_j(k)\psi_{j,k}(t) \tag{3.7}$$

integrate the norm square of $g(t)$

$$\int |g(t)|^2 dt = \int \left(\sum_{k=1}^{N/2^J} c_{J,k}(t)\phi_{J,k}(t) + \sum_{j=1}^{J} \sum_{k=1}^{N/2^J} d_j(k)\psi_{j,k}(t) \right) \left(\sum_{k=1}^{N/2^J} c^*_{J,k}(t)\phi^*_{J,k}(t) + \sum_{j=1}^{J} \sum_{k=1}^{N/2^J} d^*_j(k)\psi^*_{j,k}(t) \right) dt$$

$$\int |g(t)|^2 dt = \sum_{k=1}^{N/2^J} c_{J,k}\, c^*_{J,k} \int \phi_{J,k}(t)\, \phi^*_{J,k}(t)\, dt \; + \sum_{k=1}^{N/2^J} c_{J,k} \sum_{j=1}^{J} d^*_{j,k} \int \phi_{J,k}(t)\, \psi_{j,k}(t)\, dt + \sum_{k=1}^{N/2^J} c^*_{J,k} \sum_{j=1}^{J} d_{j,k} \int \psi_{j,k}(t)\, \phi^*_{J,k}(t)\, dt \; + \sum_{j=1}^{J} \sum_{k=1}^{N/2^J} d_{j,k}\, d^*_{j,k} \int \psi_{j,k}(t)\, \psi^*_{j,k}(t) dt \tag{A.1}$$

The norm of the wavelet is typically normalized to one, resulting in the following

$$\int \phi_{J,k}(t)\, \phi^*_{J,k}(t)\, dt = 1$$

$$\int \psi_{j,k}(t)\, \psi^*_{j,k}(t)\, dt = 1$$

The wavelet and scaling functions are orthogonal over both the translation and scale. Therefore, we have the following relations

$$\int \phi_{J,k}(t)\, \psi_{j,k}(t)\, dt = 0$$

$$\int \psi_{j,k}(t)\, \phi^*_{J,k}(t)\, dt = 0$$

Equation (A.1) then becomes

$$\int |g(t)|^2 dt = \sum_{k=1}^{N/2^J} c_{J,k}\, c^*_{J,k}(1) + \sum_{k=1}^{N/2^J} c_{J,k} \sum_{j=1}^{J} d^*_{j,k}(0) + \sum_{k=1}^{N/2^J} c^*_{J,k} \sum_{j=1}^{J} d_{j,k}(0) + \sum_{j=1}^{J} \sum_{k=1}^{N/2^J} d_{j,k}\, d^*_{j,k}(1)$$

$$\int |g(t)|^2 dt = \sum_{k=1}^{N/2^J} |c_{J,k}|^2 + \sum_{j=1}^{J} \sum_{k=1}^{N/2^J} |d_{j,k}|^2 \tag{A.2}$$

This result shows that the energy of a signal is distributed to the coefficients of the expansion and *Parseval's Theorem* holds for the wavelet transform.

APPENDIX B

LyAlphaDataExtractor Program

```
function LyAlphaDataExtractor
%LYALPHADATAEXTRACTOR Extracts the relevant Lyman Alpha data from SDSS catalogs
%   LyAlphaDataExtractor extracts the relevant Lyman Alpha data from the
%   SDSS value added catalogs. This data can then be combined into a single
%   array, or used as is to study spatial correlations or other studies.
%
% Outputs:  Mat file containing the Lyman Alpha transmitted fluxes
%           treated appropriately as described in Lee et al, astro/ph
%           1211.5146v2

tic

datadir = 'D:\sdss\';
% get list of all fits files in directory and subdirectories
dirinfo = dir(fullfile(datadir, '**\*.*'));  %get list of files and folders in any subfolder
```

```
16 filelist = dirinfo(~[dirinfo.isdir]);  %remove folders from
       list
17 dirlen = length(filelist);
18
19 %dirlist =importdata('/Users/jpando/Research/Cosmology/LSS/
       SDSS/data/BOSSLyaDR9_spectra/fisica_files.txt');
20 %dirlen = length(dirlist);
21 %m = matfile('/Users/jpando/Research/Cosmology/LSS/SDSS/data
       /Binned_Data/Weighted_Binned_Data.mat','Writable',true);
22 m = matfile('C:\Users\jspero\Desktop\sdss\
       Weighted_Binned_Data2.mat','Writable',true);
23
24 % deltalambda = 1;  % binsize. Set to values described in
       astro/ph 1812.03554
25 quasarcount = 0;
26 for dr= 1:dirlen
27     if rem(dr,5000)== 0
28         fprintf('working on file %g\n',dr)
29     end
30     binflux = zeros(1,1050);
31     %specfile = dirlist{dr};
32     %specfile = specfile(3:end);
33     %fname = strcat('/Users/JPANDO/Research/Cosmology/LSS/
           SDSS/data/BOSSLyaDR9_spectra/',specfile);
34
35     data = readflux(dr,filelist(dr).name);
36     quasarcount = quasarcount+1;
37
```

```
38      % Get Lyman-alpha redshift for all quasars
39      z = (data(:,1) - 1216)./1216;
40      zend = length(z);
41
42      % Binning algorithm to calculate total flux in each bin.
43      % deltalambda is the bin width (in angstroms)
           corresponding to
44      % velocity space 69 km/s.
45      % set to z ~ 2
46      for zo = 1:zend
47
48          if z(zo)>=2 && z(zo)<2.2
49              lambdamin=3648;
50              %deltalambda is the bin size for this redshift
                   range
51              deltalambda = 0.8671; %lambda_mid = 3770 (use
                   2.1 in z=(lambda_0 - lambda_em)/lambda_em)
52              binL = round( (data(zo,1) - lambdamin)/
                   deltalambda);
53              if binL > 0
54                  binflux(binL) = binflux(binL) + data(zo,2);
55              else
56                   binL =1;
57                   binflux(binL) = binflux(binL) + data(zo,2);
58              end
59          elseif z(zo)>=2.2 && z(zo)<2.4
60              lambdamin =3891;
61              deltalambda = 0.9229; %lambda_mid = 4013
```

```
62                 binL = 281+round( (data(zo,1) - lambdamin)/
                       deltalambda);
63                 binflux(binL) = binflux(binL) + data(zo,2);
64
65             elseif z(zo)>=2.4 && z(zo)<2.6
66                 lambdamin = 4134;
67                 deltalambda = 0.9789; %lambda_mid = 4256
68                 binL = 545+round( (data(zo,1) - lambdamin)/
                       deltalambda);
69                 binflux(binL) = binflux(binL) + data(zo,2);
70
71             elseif z(zo)>=2.6 && z(zo)<2.8
72                 lambdamin = 4377;
73                 deltalambda = 1.0347;%lambda_mid = 4499
74                 binL = 794 +round( (data(zo,1) - lambdamin)/
                       deltalambda);
75                 binflux(binL) = binflux(binL) + data(zo,2);
76
77             end
78         end % for zo = 1:zend
79
80         m.binflux(dr,1:1050) =binflux;
81 end % for end dr= 1:dirlen
82 toc
83 disp(quasarcount)
84 end
```

APPENDIX C

LymanAlphaBinner Program

```
function LymanAlphaBinner
% LymanAlphaBinner builds the array to analyze by wavelet packet power
% spectrum.  Array contains all the data in average bins along with the
% variance of each bin.
%
% Outputs: 2 column file. Column 1 contains the average at that particular
%          bin, column 2 contains the variance at that bin
tic

datadir = 'C:\Users\jspero\Desktop\sdss\';
fluxdat=[];
%save('/Users/jpando/Research/Cosmology/LSS/SDSS/data/Binned_Data/Weighted_flux_Data.mat','fluxdat');
save(strcat(datadir,'Weighted_flux_Data2.mat'),'fluxdat');
% m = matfile('/Users/jpando/Research/Cosmology/LSS/SDSS/data/Binned_Data/Weighted_Binned_Data.mat');
%load('/Users/jpando/Research/Cosmology/LSS/SDSS/data/Binned_Data/Weighted_Binned_Data.mat');
load(horzcat(datadir,'Weighted_Binned_Data2.mat'));
```

```
% disp(size(binflux))
% pause
% nzz = sum(m.binflux~=0,1); % used to eliminat bins with no
    data

tmpmu=zeros(1029,1);
v = zeros(1029,1);
figure
% parfor n = 1:1029   % using parallel toolbox to speed up
   operations.
for n = 1:1029
    if rem(n,500)==0
     fprintf('working on bin %g\n',n)
    end

      [r,~] = find(binflux(:,n)); %-4
    if isempty(r)
        continue
    end
    nz = zeros(length(r),1);
    for nn=1:length(r)
        nz(nn) = binflux(r(nn),n);
    end
     tmpmu(n) = mean(nz);
     v(n) = var(nz);
    clear r nz
end

```

```
43 plot(tmpmu)
44 fluxdat = [tmpmu v];
45 %save('/Users/jpando/Research/Cosmology/LSS/SDSS/data//
       Binned_Data/Weighted_flux_Data.mat','fluxdat','-append');
46 save(strcat(datadir,'Weighted_flux_Data2.mat'),'fluxdat','-
       append');
47 toc
48
49 end
```

APPENDIX D

WP_Power Program

```
function WP_Power(level)
%WP_POWER 1-D wavelet packet power spectrum
%   WP_POWER(LEVEL) uses the discrete wavelet packet transform function
%   DWPT to compute the 1-D power spectrum of a Sloan Digital Sky Survey
%   (SDSS) dataset. The argument @level determins the terminal (final-level)
%   nodes of the discrete wavelet packet transform (DWPT).
%
%   The SDSS dataset contains two columns of redshift binned data:
%   1) Total flux per bin, and 2) Total noise per bin.
%
%   For plotting the power spectrum, the data is group by redshift ranges.
%   The average flux in each redshift group is calculated, and that is used
%   to find the variance of each flux file from SDSS.

% Directory containing SDSS lyman-alpha flux files.
dataDir = 'C:\Users\jspero\Desktop\sdss\';
```

```
17 load(horzcat(dataDir,'Weighted_flux_Data.mat'));
18 %dataDir = 'C:\Users\jspero\Google Drive (pero.jason@gmail.
     com)\Education\DePaul\Courses\PHY 480 - Research\depaul-
     cosmo-research-project\sandbox\';
19 %load(horzcat(dataDir,'Weighted_flux_Data_all.mat'));
20
21 n = 0;
22 DeltaV = 69; % km/s
23 % Used to create four redshift bin groups which will be
     plotted later
24 binGroupSize =[281,544,792,1029];
25 f1 = figure;
26 for zz =1:4
27  %if (zz == 1 | zz == 2 | zz==4)
28  if (zz == 3)
29     % Average flux for entire bin group
30      avgflux= mean(nonzeros(fluxdat(n+1:binGroupSize(zz),1)))
           ;%,'omitnan');
31
32     % Variance from average flux for each data file
33 %      delta = nonzeros((fluxdat(n+1:binGroupSize(zz),1)-
     avgflux)./avgflux);
34      delta = nonzeros(fluxdat(n+1:binGroupSize(zz),1)-avgflux
           );
35      delta = delta./avgflux;
36
37     % Standard deviation computed from the noise vector
38      sigma = sqrt(fluxdat(n+1:binGroupSize(zz),2));
```

```
39      sigma=sigma./avgflux;
40
41      % Standard error
42      sigma = sigma/sqrt(length(fluxdat(n+1:binGroupSize(zz)))
            );
43
44      % Perform discrete wavelet packet tranform on the flux
            and noise data
45      [nodecoef,bkv,plevels,freq,en] = dwpt(delta,'db1','Level
            ',level);
46      [snodecoef,bkv,plevels,freq,sigen] = dwpt(sigma,'db1','
            Level',level);
47
48      % Number of sampling nodes produced from the DWPT
            decomposition
49      num_Nodes = length(nodecoef);
50
51      % Creates separate values for total energy of all
            frequency (sampling)
52      % nodes for both flux and standard deviation
53      totalenergy= 0;
54      totalsigenergy=0;
55      for nodelevel=1:num_Nodes
56          totalenergy = totalenergy +sum(nodecoef{1,nodelevel
                }.^2);
57          totalsigenergy = totalsigenergy +sum(snodecoef{1,
                nodelevel}.^2);
58      end
```

```
59
60      % We want energy per frequency
61      physfreq = pi/2*linspace(1/(DeltaV*num_Nodes),1/DeltaV,
            num_Nodes);
62      physfreq = physfreq';
63      fwp = physfreq(2) - physfreq(1);
64
65      % Calculate the power spectrum
66      normfactor=1./(2^num_Nodes*fwp);
67      pwrspec=totalenergy*cell2mat(en).*normfactor;% see eqs
            (15,16) Arianandа et al
68      spwrspec=totalsigenergy*cell2mat(sigen).*normfactor;
69      pwrspec=pwrspec(2:end);
70      spwrspec=sqrt(spwrspec(2:end));
71      physfreq=physfreq(2:end);
72
73
74      % Plot of power spectrum
75      % multiply power spectrum by k/pi like in page 30 of
            0405013.pdf
76      % paper plots log data on a log scale
77      figure(f1)
78      errorbar(log(physfreq),log(pwrspec.*(physfreq/pi)),
            spwrspec./pwrspec,'.-','MarkerSize',12);
79  %      errorbar(physfreq,pwrspec.*(physfreq/pi),spwrspec);
80      set(gca,'Xscale','log','Yscale','log')
81
82      % Update axis labels to show the exponential values
```

```
    instead of the
% logorithmic values.
xtick_vals = [-6.266 -5.991 -5.298 -4.893 -4.605 -4.382
    -3.912];
xtick_labels = num2cell(xtick_vals);

ytick_vals = [-4.605 -2.303 0];
ytick_labels = num2cell(ytick_vals);

% Convert axis value to its exponential equivalent. Also
    , put into cell
% array of character vectors format for the xticklabels
    function.
% Since the values in the plot are already logged,
    actually, natural
% logged, we just need to take the exponential of the
    current value to
% get the updated axis label value.
%   ln(newlabel) = current_axis_value, so
%   newlabel = exp(current_axis_value)
for i=1:length(xtick_labels)
    xtick_labels{1,i} = num2str(round(exp(xtick_labels
        {1,i}),4));
end
for i=1:length(ytick_labels)
    ytick_labels{1,i} = num2str(round(exp(ytick_labels
        {1,i}),4));
end
```

```
103
104     % Tick values are set to the actual log values. The tick
            labels are
105     % updated to reflect the exponential
106      xticks(xtick_vals)
107      xticklabels(xtick_labels)
108      yticks(ytick_vals)
109      yticklabels(ytick_labels)
110
111      hold on
112 end % //end if (zz == 1 | zz==4)
113
114      n = binGroupSize(zz);
115 end
116
117 % Make x-axis 1.5 times larger than the y and z axes
118 pbaspect([1.5 1 1])
119
120 % Adujst axis to ensure all data points are visible with a
       small white
121 % space cushion on either side and top and bottom of the x
       and y axes.
122 axis([-6 -3.70 -5.3 -1.2])
123 %axis([-6 -3.70 -5.3 -1])
124 ylabel('$P(k)*k/\pi$','Interpreter','latex','FontSize',12)
125 xlabel('$k[km/s]^{-1}$','Interpreter','latex','FontSize',12)
126 %legend('2.0 < z < 2.2','2.2 < z < 2.4','2.4 < z < 2.6','2.6
       < z < 2.8','Location','northwest')
```

```
127 %legend('2.0 < z < 2.2','2.6 z < 2.8','Location','northwest
      ')
128 legend('2.0 < z < 2.2','2.2 < z < 2.4','2.6 < z < 2.8','
      Location','northwest')
129 end
```

REFERENCES

[1] Christopher P Ahn et al. "THE NINTH DATA RELEASE OF THE SLOAN DIGITAL SKY SURVEY: FIRST SPECTROSCOPIC DATA FROM THE SDSS-III BARYON OSCILLATION SPECTROSCOPIC SURVEY". In: *The Astrophysical Journal Supplement Series* 203.2 (Nov. 2012), p. 21. ISSN: 1538-4365. DOI: 10.1088/0067-0049/203/2/21. URL: http://dx.doi.org/10.1088/0067-0049/203/2/21.

[2] L A Barnes. *Studying galaxy formation through Lyman alpha in emission and absorption (Doctoral Thesis)*. 2010. URL: https://doi.org/10.17863/CAM.15990.

[3] R Barnett et al. "The spectral energy distribution of the redshift 7.1 quasar ULAS J1120+0641". In: *Astronomy & Astrophysics* 575 (Feb. 2015), A31. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201425153. URL: http://dx.doi.org/10.1051/0004-6361/201425153.

[4] Jill Bechtold. *Shadows of creation: quasar absorption lines and the genesis of galaxies.* July 1997. URL: https://www.thefreelibrary.com/Shadows+of+creation%3a+quasar+absorption+lines+and+the+genesis+of...-a019897527.

[5] Charles Burrus, R Gopinath, and H Guo. "Introduction to Wavelets and Wavelet Transform—A Primer". In: *Recherche* 67 (June 1998).

[6] Solène Chabanier et al. "The one-dimensional power spectrum from the SDSS DR14 Lyα forests". In: *Journal of Cosmology and Astroparticle Physics* 2019.07 (July 2019), p. 17. ISSN: 1475-7516. DOI: 10.1088/1475-7516/2019/07/017. URL: http://dx.doi.org/10.1088/1475-7516/2019/07/017.

[7] Rupert A C Croft et al. "Toward a Precise Measurement of Matter Clustering: Lyα Forest Data at Redshifts 2–4". In: *The Astrophysical Journal* 581.1 (Dec. 2002), pp. 20–52. ISSN: 1538-4357. DOI: 10.1086/344099. URL: http://dx.doi.org/10.1086/344099.

[8] A Day. *An Analysis of Astrophysics and Fundamental Physics from the Lyman-alpha Forest*. 2013. URL: https://escholarship.org/uc/item/3mx4f1bs.

[9] Dyonisius Dony Ariananda, Madan Kumar Lakshmanan, and Homayoun Nikookar. "An Investigation of Wavelet Packet Transform for Spectrum Estimation". In: *arXiv e-prints* (Apr. 2013), arXiv:1304.3795.

[10] Daniel J Eisenstein et al. "SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems". In: 142.3 (Sept. 2011), p. 72. DOI: 10.1088/0004-6256/142/3/72.

[11] Antonella Garzilli, Tom Theuns, and Joop Schaye. "The broadening of Lyman-α forest absorption lines". In: *Monthly Notices of the Royal Astronomical Society* 450.2 (Apr. 2015), pp. 1465–1476. ISSN: 0035-8711. DOI: 10.1093/mnras/stv394. URL: http://dx.doi.org/10.1093/mnras/stv394.

[12] J Gunn and B Peterson. "On the Density of Neutral Hydrogen in Intergalactic Space". In: *The Astrophysical Journal* 142 (1965), pp. 1633–1636.

[13] Siyu He et al. "Learning to Predict the Cosmological Structure Formation". In: *Proceedings of the National Academy of Sciences* 116.28 (Aug. 2019), pp. 1–8. URL: https://arxiv.org/pdf/1811.06533.pdf.

[14] Khee-Gan Lee et al. "The BOSS Lyman-alpha Forest Sample from SDSS Data Release 9". In: *aj* 145.3 (Mar. 2013), p. 69. DOI: 10.1088/0004-6256/145/3/69.

[15] Wen-Juan Liu et al. "A COMPREHENSIVE STUDY OF BROAD ABSORPTION LINE QUASARS. I. PREVALENCE OF He i ABSORPTION LINE MULTIPLETS IN LOW-IONIZATION OBJECTS". In: *The Astrophysical Journal Supplement Series* 217.1 (Mar. 2015), p. 11. DOI: 10.1088/0067-0049/217/1/11. URL: https://doi.org/10.1088/0067-0049/217/1/11.

[16] Theodore Lyman. "The Spectrum of Hydrogen in the Region of Extremely Short Wave-Lengths". In: 23 (Apr. 1906), p. 181. DOI: 10.1086/141330.

[17] Hélion du Mas des Bourboux et al. "The Completed SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations with Lyα Forests". In: *The Astrophysical Journal* 901.2 (Oct. 2020), p. 153. ISSN: 1538-4357. DOI: 10.3847/1538-4357/abb085. URL: http://dx.doi.org/10.3847/1538-4357/abb085.

[18] Patrick McDonald et al. "The Lyα Forest Power Spectrum from the Sloan Digital Sky Survey". In: *The Astrophysical Journal Supplement Series* 163.1 (Mar. 2006), pp. 80–109. ISSN: 1538-4365. DOI: 10.1086/444361. URL: http://dx.doi.org/10.1086/444361.

[19] Jatan K Modi et al. "Wavelet transforms: Application to data analysis - I". In: *Resonance* 9.11 (2004), pp. 10–22. ISSN: 0973-712X. DOI: 10.1007/BF02834969. URL: https://doi.org/10.1007/BF02834969.

[20] Daniel J Mortlock et al. "A luminous quasar at a redshift of z = 7.085". In: *Nature* 474.7353 (June 2011), pp. 616–619. ISSN: 1476-4687. DOI: 10.1038/nature10159. URL: http://dx.doi.org/10.1038/nature10159.

[21] Palanque-Delabrouille Nathalie et al. "The one-dimensional Lyα power spectrum from BOSS". In: *A&A* 559 (July 2013), A85. DOI: 10.1051/0004-6361/201322130. URL: https://doi.org/10.1051/0004-6361/201322130.

[22] Barbara Ryden. *Introduction to Cosmology*. Second Edition. Cambridge University Press, 2017. ISBN: 978-1-107-15483-4.

[23] Barbara Ryden and Bradley M Peterson. *Foundations of Astrophysics*. Cambridge University Press, 2020. DOI: 10.1017/9781108933001.

[24] Stephen A Smee et al. "THE MULTI-OBJECT, FIBER-FED SPECTROGRAPHS FOR THE SLOAN DIGITAL SKY SURVEY AND THE BARYON OSCILLATION SPECTROSCOPIC SURVEY". In: *The Astronomical Journal* 146.2 (July 2013), p. 32. ISSN: 1538-3881. DOI: 10.1088/0004-6256/146/2/32. URL: http://dx.doi.org/10.1088/0004-6256/146/2/32.

[25] James S. Walker. *A Primer on WAVELETS and Their Scientific Applications*. 2nd. Boca Raton: Chapman & Hall/CRC, Taylor & Francis Group, 2008.

[26] Donald G York et al. "The Sloan Digital Sky Survey: Technical Summary". In: *The Astronomical Journal* 120.3 (Sept. 2000), pp. 1579–1587. ISSN: 0004-6256. DOI: 10.1086/301513. URL: http://dx.doi.org/10.1086/301513.