


2016

Applying Hierarchical Clustering to Broad Absorption Line Profiles for Quasar Classification

Nathalie Chantal Marie Thibert
Western University, nthibert@uwo.ca

Follow this and additional works at: https://ir.lib.uwo.ca/ungradawards_2016

 Part of the [External Galaxies Commons](#), and the [Other Astrophysics and Astronomy Commons](#)

Recommended Citation

Thibert, Nathalie Chantal Marie, "Applying Hierarchical Clustering to Broad Absorption Line Profiles for Quasar Classification" (2016). *2016 Undergraduate Awards*. 4.
https://ir.lib.uwo.ca/ungradawards_2016/4

Applying Hierarchical Clustering to Broad
Absorption Line Profiles for Quasar Classification

Nathalie Chantal Marie Thibert

Applying Hierarchical Clustering to Broad Absorption Line Profiles for Quasar Classification

Nathalie Chantal Marie Thibert

Supervised by:

Dr. Sarah Gallagher and Dr. Mark Daley

Honors Thesis (Physics 4999E)

Submitted: March 28, 2016

Re-submitted: April 20, 2016

Department of Physics and Astronomy

The University of Western Ontario

Abstract

The region immediately surrounding an actively accreting supermassive black hole at the centre of a massive galaxy, the accretion disk, produces an enormous amount of radiation resulting in a luminous, short-lived phenomenon called a quasar. About 20% of quasars show broad, blue-shifted absorption features in their UV spectra, indicative of an outflowing wind from the accretion disk. These winds can remove angular momentum from the accretion disk, thereby contributing to the growth of the central black hole. Understanding these winds will help us to better constrain the details of how black holes grow during the quasar phase. The structures of the absorption features are sensitive to the properties (ionization state, velocity profile, and thickness) of the winds. Consequently, the broad absorption line profiles of these objects show great diversity in depth and velocity width. Using a sample of 1,084 broad absorption line quasar spectra from the Sloan Digital Sky Survey, we apply an agglomerative hierarchical clustering algorithm to group spectra by similar C IV absorption line shapes. For each cluster, we compose median spectra and compare the shapes of the C IV broad absorption lines with the properties of prominent, broad emission lines. In agreement with results in the literature, low-velocity, deep troughs are found preferentially in objects for which the radiation from the accretion disk is more energetic. The link between broad absorption line properties and those of emission lines holds promise for allowing us to constrain the structure and dynamics of the outflowing winds.

Chapter 1

Introduction

1.1 Observational Properties of Quasars

The centres of massive galaxies are home to supermassive black holes that, during a short phase of the host galaxy's lifetime, grow through the gravitational infall of matter onto a hot *accretion disk*. During this phase, the central black hole grows as a luminous *quasar*. The light emitted from the central accretion disk is sometimes enough to outshine the stellar light from the entire host galaxy by a factor of 100 or more [Peterson, 1997]. Since accretion disks are on the order of only a few light-days across¹, the above model for the central engine of a quasar was proposed to account for its size relative to its energy output. In quasars, masses of supermassive black holes typically exceed $M_{\text{SMBH}} \approx 10^8 M_{\odot}$ and their luminosities are on the order of $L_{\text{quasar}} \approx 10^{46} \text{ erg s}^{-1}$ [Peterson, 1997].

Quasars are the most luminous subclass of a more general type of phenomena called *active galactic nuclei*. First observed in radio surveys, quasars are characterized not only by their radio properties, but also by small angular sizes (star-like appearance), continuum variability, large ultraviolet (UV) fluxes, large redshifts, and broad emission lines [Peterson, 1997].

Quasars are a valuable tool for astronomers wishing to study the evolution of galaxies over cosmic time because they are so luminous and, as a result, can be

¹1 light-day = $2.54 \times 10^{15} \text{ cm}$

detected at high cosmological redshifts. Redshift z is defined as

$$z \equiv \frac{\lambda_{\text{obs}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}}, \quad (1.1)$$

where λ_{obs} and λ_{emit} are the observed and emitted wavelengths, respectively. Cosmological redshift is a direct consequence of the expansion of the Universe, causing the light emitted from distant objects to be shifted to longer wavelengths as it travels to Earth. Rearranging Equation 1.1, we can calculate the rest-frame (emitted) wavelengths of a spectrum to be

$$\lambda_{\text{emit}} = \frac{\lambda_{\text{obs}}}{1 + z}. \quad (1.2)$$

Variations in the continuum of a quasar spectrum can be observed on timescales as short as a few days [Peterson, 1997], helping to constrain the size of the continuum source and supporting the idea that quasars are powered by accretion onto a supermassive black hole.

Quasars show both broad and narrow emission lines in their spectra. These emission lines originate from distinct regions outside the accretion disk, called the *broad line region* and *narrow line region*, respectively. We cannot observe the accretion disk directly, but we can use these regions to probe the underlying continuum with use of emission and absorption lines. The broad emission lines are Doppler-broadened by bulk motions of individual clouds and can have velocity widths on the order of $\Delta v_{\text{FWHM}} \approx 500 \text{ km s}^{-1}$ up to $\Delta v_{\text{FWHM}} > 10^4 \text{ km s}^{-1}$ [Peterson, 1997].

1.2 The Quasar Spectral Energy Distribution

Multi-wavelength studies have been used to compose spectral energy distributions (SEDs) of quasars from radio wavelengths ($\sim 10 \text{ cm}$) all the way to the X-ray ($> 1.2 \text{ keV}$). Figure 1.1 shows the SEDs for populations of both *radio-loud*² and *radio-quiet* quasars.

Each frequency regime in the figure corresponds to a different physical driver

²These objects show strong emission from relativistic particle jets.

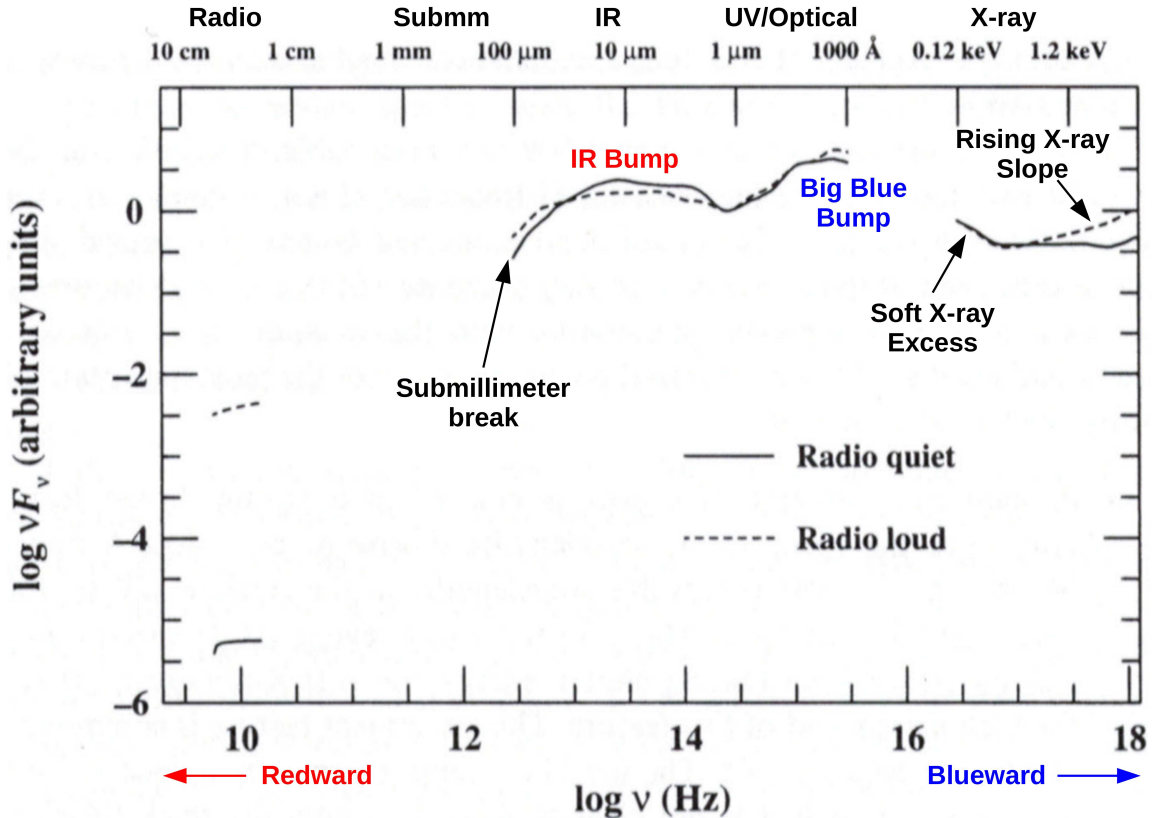


Figure 1.1: Mean rest-frame SEDs of 18 radio-loud (dashed line) and 29 radio-quiet (solid line) normal quasars. The SEDs are normalized at $1.25 \mu\text{m}$. (Figure 4.1 in Peterson 1997 as adapted from Figure 10 in Elvis et al. 1994 and annotated here).

within the quasar system. According to Elvis et al. [1994], the continuum emission of a quasar can be well-modeled by the composition of two power law components, one for the hard (high-energy) X-ray component and the other for the $1\text{--}100 \mu\text{m}$ infrared (IR) band. These two power laws intersect at about 1 keV and are superposed with several “continuum features”. These features include: (1) the *submillimeter break*, (2) the *IR bump*, (3) the *big blue bump*, (4) the *soft X-ray excess*, and (5) a rising X-ray slope in radio-loud objects.

The main sources of flux at radio wavelengths can be attributed to (non-thermal) synchrotron emission from either a compact central source or extended emission from lobes of material from the host galaxy interacting with the surrounding intergalactic medium. Relativistic particle jets can also contribute to the radio continuum flux.

The IR bump and the minimum occurring at $1 \mu\text{m}$ can be attributed almost exclu-

sively to thermal emission by dust. In this interpretation, the UV/optical continuum heats dust that is hundreds of light-days from the central source [Peterson, 1997]. The dust grains absorb the continuum emission from the accretion disk and emit thermal IR photons as they cool.

The UV/optical portion of the spectrum (the big blue bump) comes mostly from thermal emission by the accretion disk. Since the accretion disk is not all the same temperature, we cannot treat it as a single blackbody at a given temperature. The temperature depends on the radial position within the disk. We must integrate the “local” luminosity over all possible disk radii to obtain the total luminosity of the accretion disk [Netzer, 2006]. That is, the UV/optical continuum is most likely dominated by emission from a continuous distribution of blackbodies at different temperatures.

At energies of about 0.1–1 keV, it is assumed that particles in a hot medium *inverse Compton scatter* photons emitted by the disk, thereby increasing their energy and producing a simple power law [Kembhavi and Narlikar, 1999]. At the highest energies (>1 keV), a hot corona contributes to the hard X-ray portion of the spectrum [Netzer, 2006].

The wavelength range of this work will be the rest-frame UV (see Figure 1.2). In addition to the continuum flux, there are also several broad and narrow emission features present. Values for the line centres of prominent emission lines in the rest-frame UV quasar spectrum are summarized in Table 1.1.

1.3 Broad Absorption Line Quasars

In general, quasar spectra can be further classified as those that show broad absorption features (*broad absorption line quasars*) and those that do not (*normal quasars*). Although broad absorption line quasars are relatively rare, with approximately 10–20% of quasars showing broad absorption lines in their spectra [Morris, 1988, Weymann et al., 1991, Gibson et al., 2009], they provide useful insights into the quasar system. Specifically, broad absorption line quasar spectra can be used as a probe of

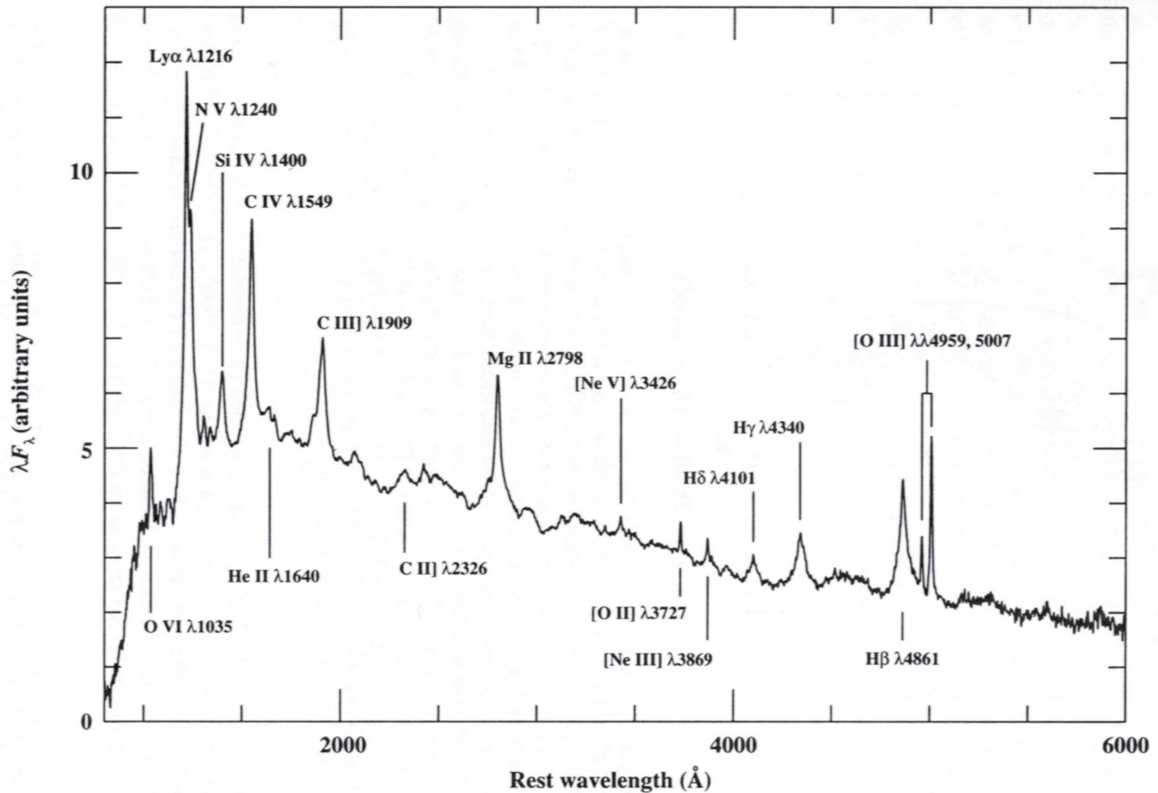


Figure 1.2: Mean rest-frame UV/optical spectrum of over 700 normal quasars from the Large Bright Quasar Survey. The 6000 Å mark corresponds to ~ 2 eV. Image taken from Peterson [1997] as adapted from Francis et al. [1991].

the central regions and the gas immediately surrounding it. Both radiation pressure and gas pressure become important in such high-energy environments and gas can be expelled in high-velocity outflows called *winds* [Murray et al., 1995, Gibson et al., 2009]. These winds can be accelerated to velocities of $\sim 0.1c$ [Murray et al., 1995]. Spectral features (emission and absorption) are Doppler broadened by the bulk motions of this high-velocity gas. Gas moving toward the observer gives rise to broad absorption lines that are blue-shifted (i.e., shifted to shorter wavelengths) from the rest-frame wavelength of their corresponding emission lines. These shifts can be as high as $25,000 \text{ km s}^{-1}$ away from the emission line centre.

Several suggestions have been made regarding the origin of broad absorption line quasars. For example, Surdej and Hutsemekers [1987] suggested that broad absorption line quasars form a distinct population, and that about 10% of all quasars con-

Table 1.1: Rest-frame wavelengths of prominent emission features in the rest-frame UV quasar spectrum (values taken from Table 2 in Vanden Berk et al. 2001).

Emission Line	Rest-frame Wavelength at Line Centre (\AA)
C II	1335.30
Si IV	1396.76
C IV	1549.06
He II	1640.42
O III]	1663.48
Al III	1857.40
Si III]	1892.03
C III]	1908.73

tain the material responsible for broad absorption line winds, which covers the central source of the quasar. In contrast, Turnshek et al. [1988], among others, suggested that viewing angle is important in the definition of broad absorption line quasars and that we can only see the broad absorption line winds 10% of the time depending on the angle of inclination of the quasar with respect to the observer. The more generally accepted approach put forth by Morris [1988] suggests that all quasars have outflowing winds, but different *covering fractions* of the central source by the broad absorption line region material allow us to see broad absorption line quasars only a fraction of the time. This broad absorption line region material was suggested by Murray et al. [1995] to be a “wind” and not individual “clouds” since a wind will naturally produce smooth line profiles. More recently, however, some authors [Baldwin et al., 1996, Nenkova et al., 2002, Elitzur and Shlosman, 2006] have suggested that the wind is not entirely homogeneous, but rather clumpy. In addition, Weymann et al. [1991] showed that broad absorption line quasars and normal quasars are in fact drawn from the same parent population and do *not* form two distinct classes of objects. For example, they showed that the emission line and continuum properties of broad absorption line and normal quasars are very similar, with few exceptions.

This further supports Morris [1988]’s interpretation of a distribution of different broad absorption line covering fractions.

Gallagher and Everett [2007] presented a schematic of a stratified model for the quasar accretion disk and outflowing wind system and discussed the physical drivers of the outflowing wind. In this model, the wind from the accretion disk is split into three components, all corresponding to different distance scales within the quasar system and different wavelength regimes. Figure 1.3 shows this model (Figure 2 from Gallagher and Everett 2007). Closest to the central black hole is where the most energetic (X-ray) photons are emitted. In this model, there is a high column density, ionized X-ray absorbing (“shielding”) gas at radii of 10^{15-16} cm between the inner accretion disk and the observer [Murray et al., 1995]. Further out in the accretion disk, about 10^{17} cm from the central black hole, a radiatively driven broad absorption line wind is present [Gallagher and Everett, 2007]. We can detect this wind spectroscopically at UV wavelengths. In this case, gas is pushed vertically out of the accretion disk and is accelerated radially outward from the source of UV continuum emission by radiation pressure [Murray et al., 1995]. At about $10^{18.5}$ cm from the black hole, a dusty outflow can be seen at IR wavelengths. The focus of this work is to investigate the spectral contributors due to the broad absorption line wind.

Figure 1.4 shows examples of both normal (“Non-BALs”) and broad absorption line (“HiBALs + LoBALs”) quasar rest-frame UV spectra. At longer wavelengths, the two composites match up well, but the broad absorption line spectrum has a significant flux deficit at shorter wavelengths [Reichard et al., 2003, Gibson et al., 2009], giving rise to a “redder” continuum.

One key feature of the broad absorption line spectrum is in the wavelength range from about 1400–1600 Å. The *blue wing* of the C IV emission feature has a large portion of its flux cut out by a broad, blue-shifted C IV absorption line, causing the C IV emission to appear asymmetric. Recall that this blue-shift is a result of the high-velocity winds being driven up and out of the accretion disk [Murray et al., 1995]. In addition, there is some velocity structure to the C IV broad absorption line.

In a qualitative sense, we can look at the spectra of most quasars and visually

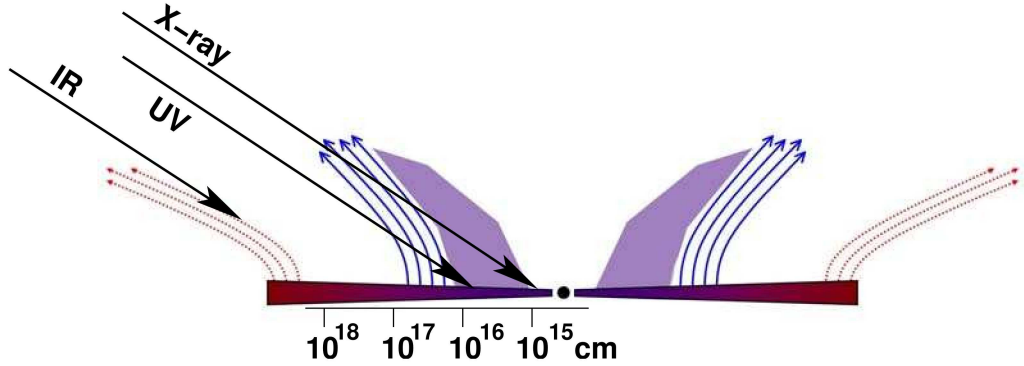


Figure 1.3: Schematic of the black hole, accretion disk, and outflowing wind system in a quasar (Figure 2 from Gallagher and Everett 2007). The black dot is the black hole and the purple-to-red gradient-coloured wedges are the accretion disk viewed edge-on. Solid purple blocks are the shielding gas, solid blue lines are the broad absorption line wind, and dotted orange lines are the dusty outflow. The black arrows show possible lines of sight through these components and point to the approximate position of origin of the continuum radiation in each wavelength regime.

classify them as broad absorption line or normal quasars. Weymann et al. [1991] introduced a quantitative way to separate broad absorption line quasars from normal quasars. They defined a *balnicity index*, BI , to quantify the amount by which an object's spectrum is affected by broad absorption lines. The balnicity index is a continuous measure of the strength of the broad absorption line features and it is defined using the C IV absorption feature. The spectra of objects with higher values of the balnicity index are more affected by broad absorption lines, whereas those with lower values of the balnicity index are less affected by broad absorption lines. A value of 0 for the balnicity index refers to a normal quasar with no broad absorption lines in its spectrum. The balnicity index is given by

$$BI \equiv \int_{3000}^{25,000} \left[1 - \frac{f(V)}{0.9} \right] C dV, \quad (1.3)$$

where V is the velocity displacement from the line centre and $f(V)$ is the continuum-normalized flux. The parameter C is equal to zero when the quantity in brackets is negative. It is equal to 1 when that same quantity has been positive (i.e., when the spectrum has fallen at least 10% below the continuum) for a velocity range of

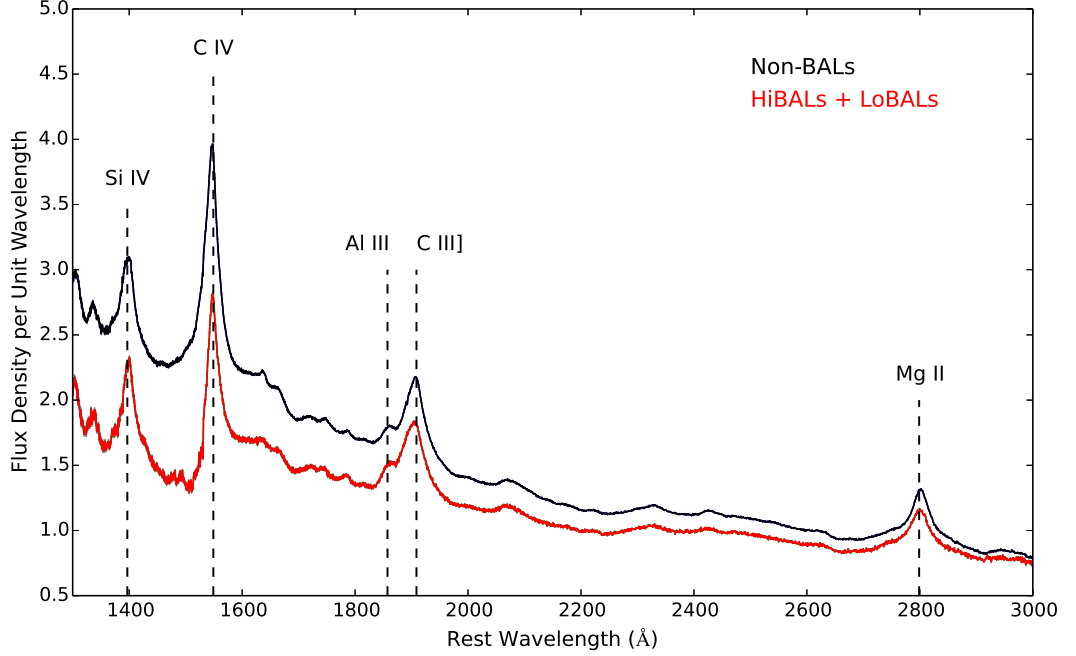


Figure 1.4: Composite spectra of normal and broad absorption line samples taken from the SDSS Early Data Release. Similar to Figure 1 from Reichard et al. [2003]. The black (red) solid line is the composite spectrum from a sample of normal (broad absorption line) quasars and the vertical black dashed lines mark vacuum wavelengths of prominent emission features (see Table 1.1).

$>2000 \text{ km s}^{-1}$. The index is in units of km s^{-1} and is positive for broad absorption line quasars. Gibson et al. [2009] modified this definition slightly and introduced a *modified balnicity index*, BI_0 , given by

$$BI_0 \equiv \int_0^{25,000} \left[1 - \frac{f(V)}{0.9} \right] C dV. \quad (1.4)$$

Defining BI and BI_0 becomes especially useful in borderline cases where spectra may be affected by poor resolution or low signal-to-noise ratios, thereby causing classification by visual inspection to be uncertain [Weymann et al., 1991].

1.4 Motivation for this Project

As mentioned in the previous section, broad absorption line quasars show broad, blue-shifted absorption features in their spectra indicative of high-velocity outflowing material along the line of sight. Broad absorption line winds are important because they may carry away angular momentum from the central regions of the quasar, allowing for accretion to occur. The extra energy may be injected into the host galaxy by the winds. Studying the properties of broad absorption lines could help us to understand how the evolution of the quasar system is affected by the winds.

Studies of broad absorption line *troughs* have been carried out in the hopes of constraining the properties of these systems and the material that comprises them. For example, Gibson et al. [2009] used a sample of broad absorption line quasars to investigate the properties of broad absorption line and normal quasars. Their results implied that although broad absorption line and normal quasars are not intrinsically different classes of objects, their spectra still show different properties.

The properties of broad absorption line quasars is the interest of this work. Specifically, we will investigate the C IV broad absorption line feature in 1,110 broad absorption line quasars from a subsample of the catalog used in Gibson et al. [2009]. C IV broad absorption lines show great diversity in shape, examples of which are shown in Figure 1.5. The different broad absorption line features show varying depths, widths, and velocity shifts from the C IV emission line centre. In addition they also show different structure within the broad absorption line.

Gibson et al. [2009] stated that trough shapes can be determined by the geometry of the outflowing material. In addition, the outflow structure is different for different ionization states and for different species. Parameters not directly associated with the C IV broad absorption lines can also have an effect on the shapes of these troughs (see, for example, Baskin et al. 2013).

We would like to understand why the C IV broad absorption lines have different shapes. Might there be a way to classify these troughs based on their shapes? What, then, could we say about the mechanisms (physical, geometric, or otherwise) govern-

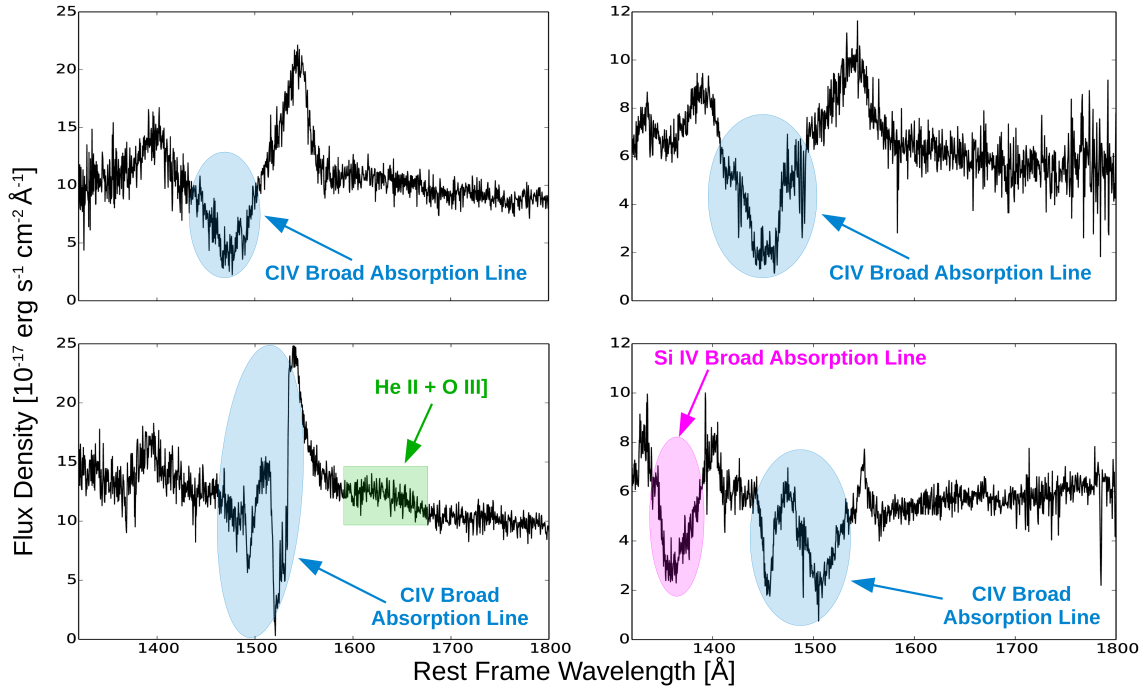


Figure 1.5: Examples of rest-frame UV broad absorption line quasar spectra from the SDSS DR5. Prominent features in each of the spectra are labelled.

ing each of these different shapes? To answer the above questions, we will employ machine learning clustering techniques to separate broad absorption line quasars into a number of groups based on their C IV trough shape. We will then examine the properties of the quasars within the different clusters and compare our results to previous work.

Chapter 2

Catalog Details and Data Reduction

2.1 Description of the Data

The quasars used in this study are drawn from the broad absorption line quasar catalog of Gibson et al. [2009], which itself is drawn from the Fifth Data Release of the Sloan Digital Sky Survey (SDSS DR5, Adelman-McCarthy et al. 2007)¹. The SDSS is a sky survey taken by the 2.5m Sloan Foundation Telescope, a ground-based telescope at Apache Point Observatory in New Mexico. The Fifth Data Release mapped one-quarter of the entire sky and collected photometric (imaging) and spectroscopic data for galaxies, quasars, and stars. Spectroscopic data were made available for 1,048,960 objects, 90,611 of which were classified as quasars.

Gibson et al. [2009] placed constraints on the SDSS DR5 quasar catalog to only select objects that contain broad absorption lines. We place a number of additional constraints on the Gibson et al. [2009] catalog to select objects that contain the C IV broad absorption line.

The first constraint is on the range of redshifts for the objects in our sample. The observed-frame wavelength coverage of the spectra in the full DR5 catalog (3800–9200 Å) limits the number of objects for which we can measure the C IV broad

¹<http://classic.sdss.org/dr5/>

absorption line. Gibson et al. [2009] suggests a redshift range of $1.68 < z < 4.93$ be imposed on their full catalog of 5039 objects such that when shifted into the rest frame, the C IV broad absorption line still falls within the SDSS bandpass. Figure 2.1 shows the distribution of redshifts for the broad absorption line quasars in our sample.

In addition to the redshift constraint, we require that there be complete wavelength and flux density coverage in the rest-frame wavelength range 1320–1800 Å. This ensures a sufficient data range on either side of the C IV broad absorption line. The second constraint is a signal-to-noise ratio cutoff of ≥ 9 so that our spectra will all have roughly the same amount of noise. The third constraint is a modified balnicity index of $BI_0 > 0$ (see Equation 1.4); i.e., a C IV broad absorption line is present in the spectrum.

After these restrictions, our initial sample includes the UV spectra of 1,110 broad absorption line quasars, along with additional quantities from the Gibson et al. [2009] catalog. These quantities include the modified balnicity index BI_0 as described in Section 1.3, and the minimum and maximum outflow velocities for the C IV broad absorption line, v_{min} and v_{max} , respectively. The values of v_{min} and v_{max} correspond to the minimum and maximum wavelengths, respectively, of the broad absorption line as defined using BI_0 . Physically, v_{min} and v_{max} give an idea of the width of the broad absorption line as well as the range of velocities to which the wind can be accelerated.

2.2 Spectral-fitting Procedure

2.2.1 The Original Reduction Pipeline

After downloading the spectra from the SDSS DR5 archive, we use the redshifts provided in the headers of the spectrum files and Equation 1.2 to convert to rest-frame wavelengths. To isolate the C IV broad absorption line for further study, we model the continuum and emission lines immediately around the C IV broad absorption line so these features may be removed. Exploratory fitting of spectra was carried

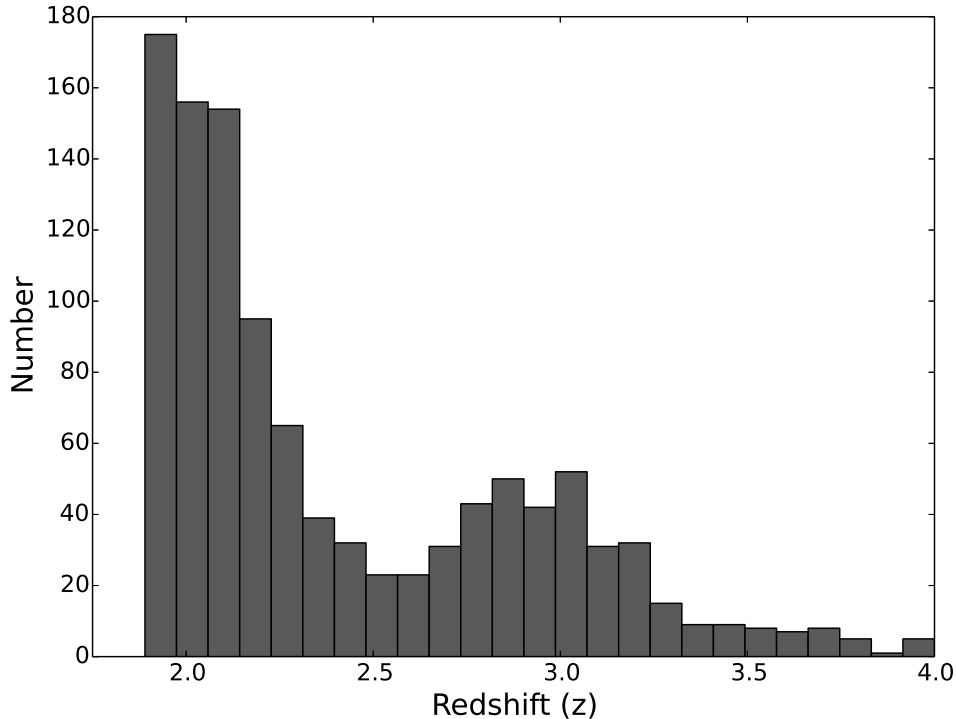


Figure 2.1: Distribution of cosmological redshifts for the 1,110 broad absorption line quasars in our sample. The data were arbitrarily chosen to be divided into 25 bins.

out using Rob Dimeo’s wrapper function to the IDL `mpfit` routine, **Peak Analysis (PAN)**². Written by Craig Markwardt³, the IDL `mpfit` routine employs a Levenberg-Marquardt non-linear least squares algorithm to fit a pre-defined model to the data [Markwardt, 2009]. It is an iterative procedure that finds the local minimum of the sum of squared errors between the data and the model.

Our model to fit the region around the C IV broad absorption line (1320–1800 Å) consists of four components: (1) a first-order polynomial for the continuum, (2) a single Gaussian for the Si IV emission line, (3) a single Gaussian for the broad component of the C IV emission line, and (4) a single Gaussian for the narrow component of the C IV emission line. We use Gaussian profiles (as opposed to Lorentzian) because the emission lines are Doppler broadened due to bulk motions of the gas. We use 2 Gaussians to fit the C IV emission line profiles because they are often complex

²<http://www.ncnr.nist.gov/staff/dimeo/panweb/pan.html>

³<http://cow.physics.wisc.edu/~craigm/idl/idl.html>

and asymmetric. The `mpfit` routine takes initial guesses for the parameters of a *composite model*, which is the addition of the four components described above. In our case, there are 11 parameters: the slope and intercept for the continuum, and the area under the Gaussian, the line centre, and full width at half maximum (FWHM) for each of the three Gaussian profiles.

The initial guesses for the parameters of the linear continuum model are calculated using flux densities and wavelengths in the ranges 1320–1340 Å and 1690–1800 Å and the Python function `numpy.ma.polyfit`. All 1,110 continuum fits are visually inspected. The initial guesses for the Gaussian parameters are estimated by performing manual fits to ten randomly selected spectra using the PAN interface. The estimated initial parameters are summarized in Table 2.1.

The model produced by the fitting routine is sensitive to the initial guesses for the emission line centres. Richards et al. [2011] showed that C IV emission lines have a wide range of possible line centres and that blueshifting in the C IV emission line is common (see Figure 2.2). We must thus allow the C IV line centres to vary over the course of the routine to account for this difference across objects, but not by much so as to avoid major displacements of the line to other parts of the spectrum. We place constraints on the amount by which both the C IV narrow and broad component line centres can vary by considering the possible range of values for C IV emission line blueshifts. Looking at Figure 2.2, velocity shifts of -1500 km s^{-1} and $+500 \text{ km s}^{-1}$ from the C IV emission line centre ($\sim 1549 \text{ Å}$) reasonably capture most of their sample. Converting these velocities into wavelengths, we find that the C IV emission line centre can vary from about 1541–1551.6 Å if given the above velocity shifts. Furthermore, we require the areas of all three Gaussian emission lines to remain positive so as to eliminate the possibility of fitting absorption features or negative dips in the spectrum due to noise. The constraints on each parameter of the model are included in Table 2.1.

Before the spectra are fit, we mask out features that do not contribute to the composite model of “linear continuum + two emission lines”. These features include the Si IV broad absorption line (when present), the C IV broad absorption line, and

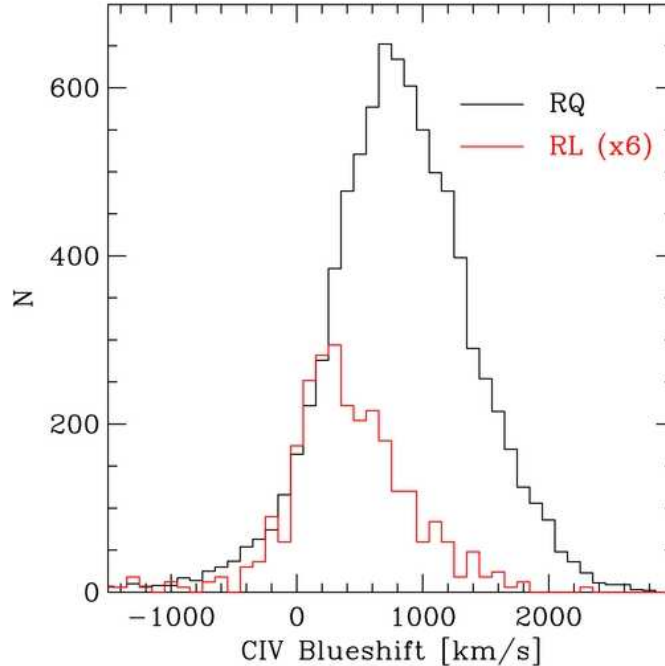


Figure 2.2: Figure 2 from Richards et al. [2011]. Range of C IV emission line centre blueshifts in a sample of $\sim 30,000$ quasars from the SDSS DR7. The red and black histograms represent radio-loud and radio-quiet quasars, respectively. Here, positive values denote blueshift, whereas negative values denote redshift.

the He II + O III] complex just redward (i.e., toward longer wavelengths) of the C IV emission line. To mask the He II + O III] complex, we exclude all data in the wavelength range 1605–1690 Å. This range has been verified by visual inspection of all spectra. To exclude the Si IV and C IV broad absorption lines, we apply unique masks to each spectrum based on the v_{min} and v_{max} values for each of Si IV and C IV given in the Gibson et al. [2009] catalog. In the case where no Si IV v_{min} and v_{max} values are given, no mask is applied for the Si IV broad absorption line. All objects have v_{min} and v_{max} values listed for the C IV broad absorption line.

The above method describes the original version of the spectral-fitting pipeline applied to the 1,110 broad absorption line quasar spectra in our sample (see Figure 2.3). Output was given in the form of plots to be used for visual inspection, arrays of the model values at each wavelength value, and the 11 parameters of the model. Normalized residuals were calculated using $\text{resid} = (\text{model} - \text{flux}) / \text{err}$, where err is the noise in the spectrum, i.e., standard deviation, in the same units as the flux density.

Table 2.1: Initial guesses and constraints for the parameters of the composite model to the emission lines and continuum around the C IV broad absorption line.

Parameter	Initial Guess	Constraints
Continuum Slope	Unique to each spectrum	
Continuum Intercept	Unique to each spectrum	
Si IV Area	$67 \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ \AA}^{-1}$	≥ 0
Si IV Centre	1396 \AA	
Si IV Width	20 \AA	
C IV (narrow) Area	$120 \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ \AA}^{-1}$	≥ 0
C IV (narrow) Centre	1545 \AA	$\in [1541.0, 1551.6] \text{ \AA}$
C IV (narrow) Width	15 \AA	
C IV (broad) Area	$60 \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ \AA}^{-1}$	≥ 0
C IV (broad) Centre	1545 \AA	$\in [1541.0, 1551.6] \text{ \AA}$
C IV (broad) Width	45 \AA	

All 1,110 spectra along with their models and residuals were visually inspected and categorized based on the “goodness of fit”. We define a spectrum with a “good fit” to be one for which the model overlaps most of the data in the wavelength range around the C IV broad absorption line (1400–1550 \AA) and for which the residuals in that same range fall within $[-1, +1]$ of the zero mark. This definition excludes the parts of the spectrum in that range that correspond to the C IV broad absorption line itself. Using this technique, 500 of the 1,110 models were visually categorized as “good fits” and the other 610 spectra needed some modification to this procedure. Figure 2.4 shows an example of a composite model fit to a broad absorption line quasar spectrum using the original pipeline. Figure 2.5 shows the residuals calculated for the model fit in Figure 2.4.

2.2.2 Modifications to the Pipeline

The original pipeline discussed in the previous section was modified to produce better fits for the remaining 610 spectra. We construct two different modifications to the original pipeline. We apply manual masking in the wavelength range 1320–1800 Å to all 610 of the remaining spectra. We also change the constraints for some spectra to allow for major shifts in the emission lines with respect to the expected line centres.

In summary, we were able to successfully fit 1,084 of the 1,110 spectra in our sample using the above modifications. We perform all of our analyses in the upcoming sections using the sample of 1,084 spectra. We summarize the modifications to the original pipeline and the number of objects fit using each modification in Table 2.2.

Table 2.2: Summary of the modifications to the original pipeline along with the number of spectra fit using each modification.

Pipeline Used	Description of Pipeline	# of Spectra Fit
Original Pipeline (OP)	See Section 2.2.1	500
Modification 1 (M1)	OP + Manual Masking	+ 458
Modification 2	M1 + Expanded C IV Emission Line Range	+ 126
No Pipeline Used		– 26
Total for subsequent analysis:		= 1,110 – 26 = 1,084

2.2.3 Normalization and Resampling

After fitting the spectra, we normalize them to the local continuum and emission lines by dividing the data by the composite model (i.e., linear continuum + 3 Gaussian profiles). Normalizing the spectra in this way allows for all data that is closely fit by the model to lie around 1.0 and any other features to show up as deviations from 1.0 on the normalized flux scale.

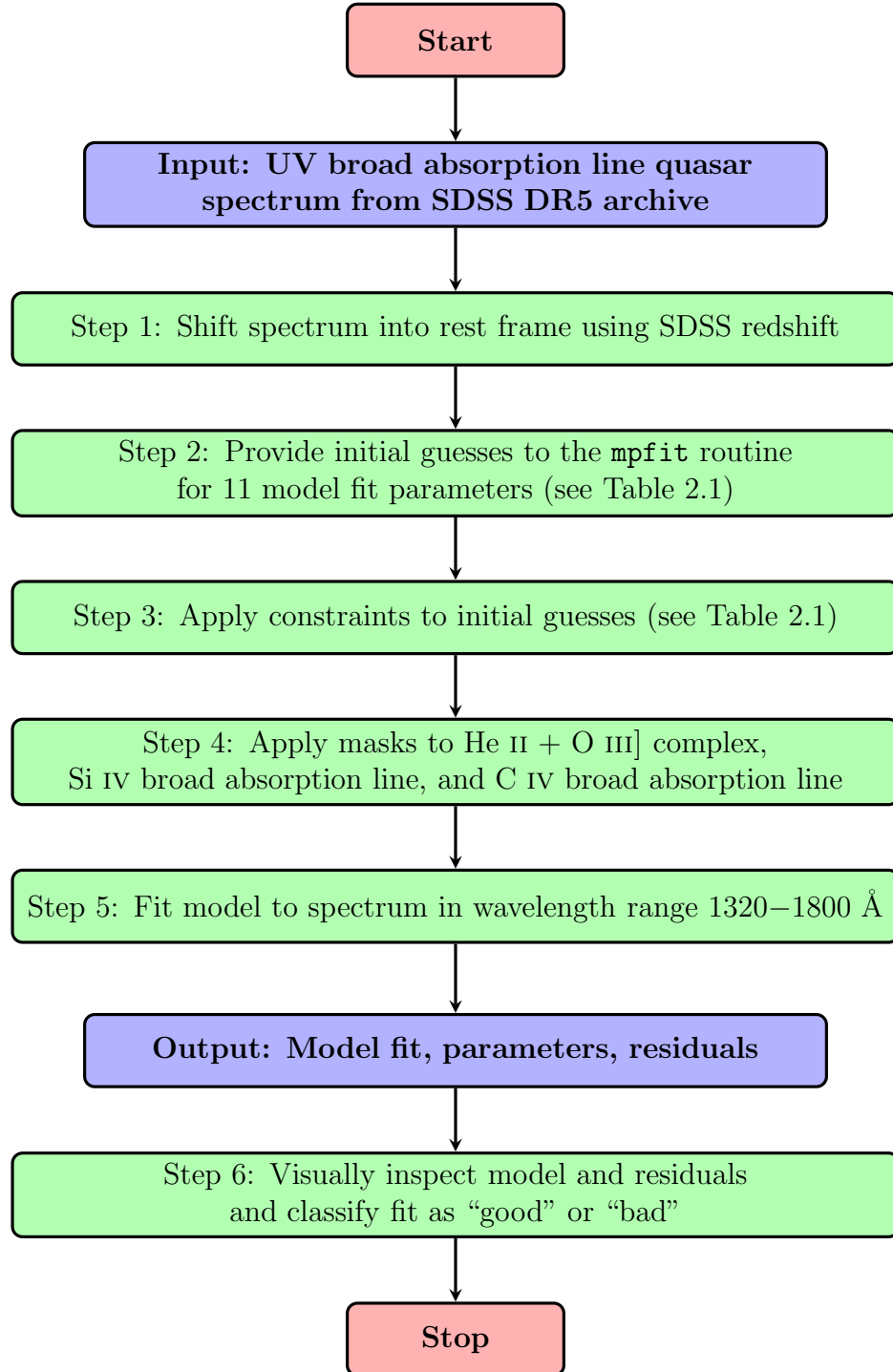


Figure 2.3: Summary of the “original pipeline” used to fit 500 of the 1,110 broad absorption line spectra. Modifications to this pipeline are described in Section 2.2.2.

Next, we resample each spectrum to the same wavelength grid such that they can be compared to one another. We define a grid from 1320–1800 Å with linear

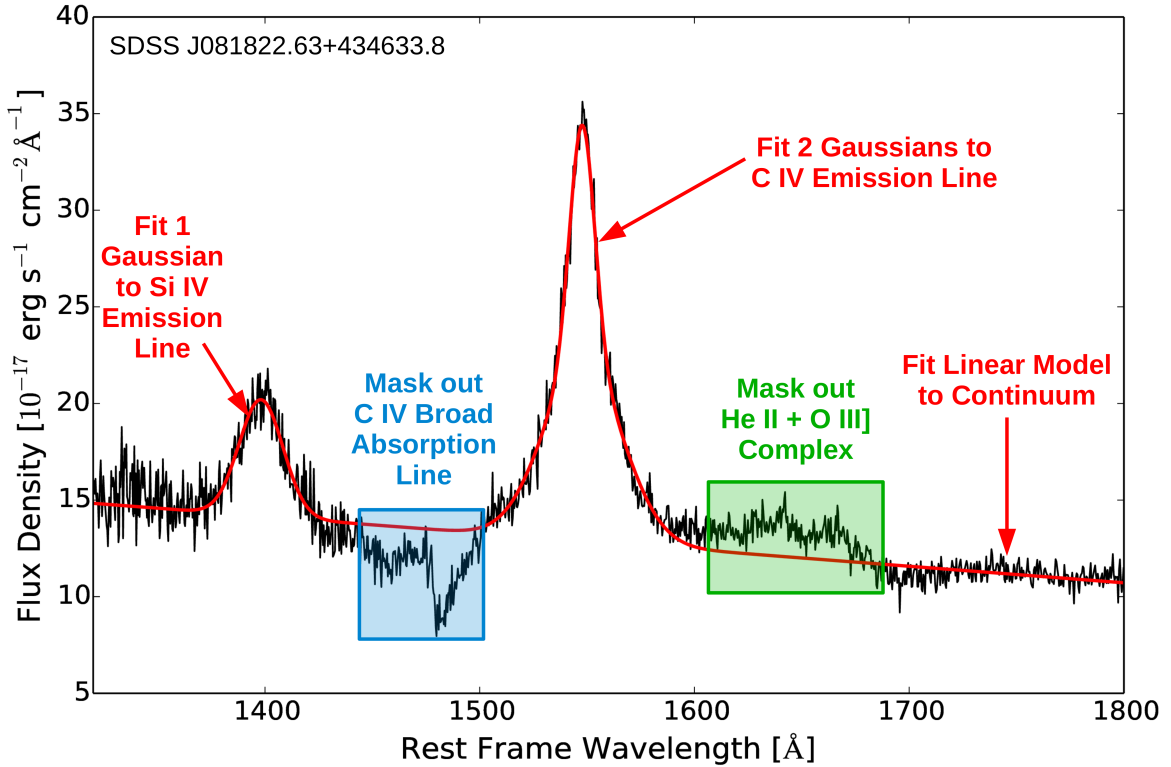


Figure 2.4: Composite model fit to a broad absorption line quasar spectrum. The black line is the data and the solid red line is the model fit using the original pipeline. There is no Si IV absorption in this spectrum, so there is no masking in the region blueward of the Si IV emission line.

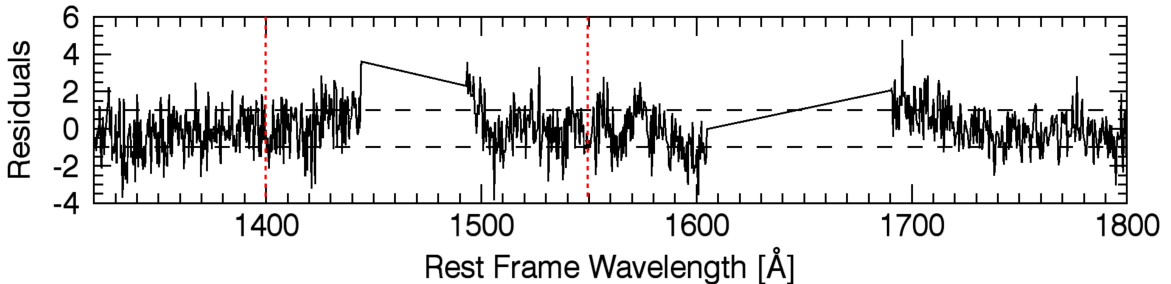


Figure 2.5: Residuals for the model fit in Figure 2.4. Black dashed lines mark ± 1 standard deviation above and below the zero mark. Red dotted lines mark 1400 and 1550 Å, the region inside which the residuals are considered when assessing goodness of fit. Missing values correspond to the masked regions in the fit (here, we mask the C IV broad absorption line and the He II + O III] complex).

spacing of 0.3 \AA , similar to the spectral resolution of the SDSS spectra. When each spectrum is resampled to this grid, a linear interpolant is used to infer a normalized flux value at each of the sampled points in the new grid. We use the Python function

`numpy.interp` to perform the resampling.

We truncate the spectra to the wavelength range 1400–1550 Å, which includes the C IV broad absorption line, the red wing of the Si IV emission line, and the blue wing of the C IV emission line. This is done so we are only considering the range immediately around the C IV broad absorption line in our analysis. In the next chapter, we will consider the methods used to analyze the 1,084 normalized, resampled spectra.

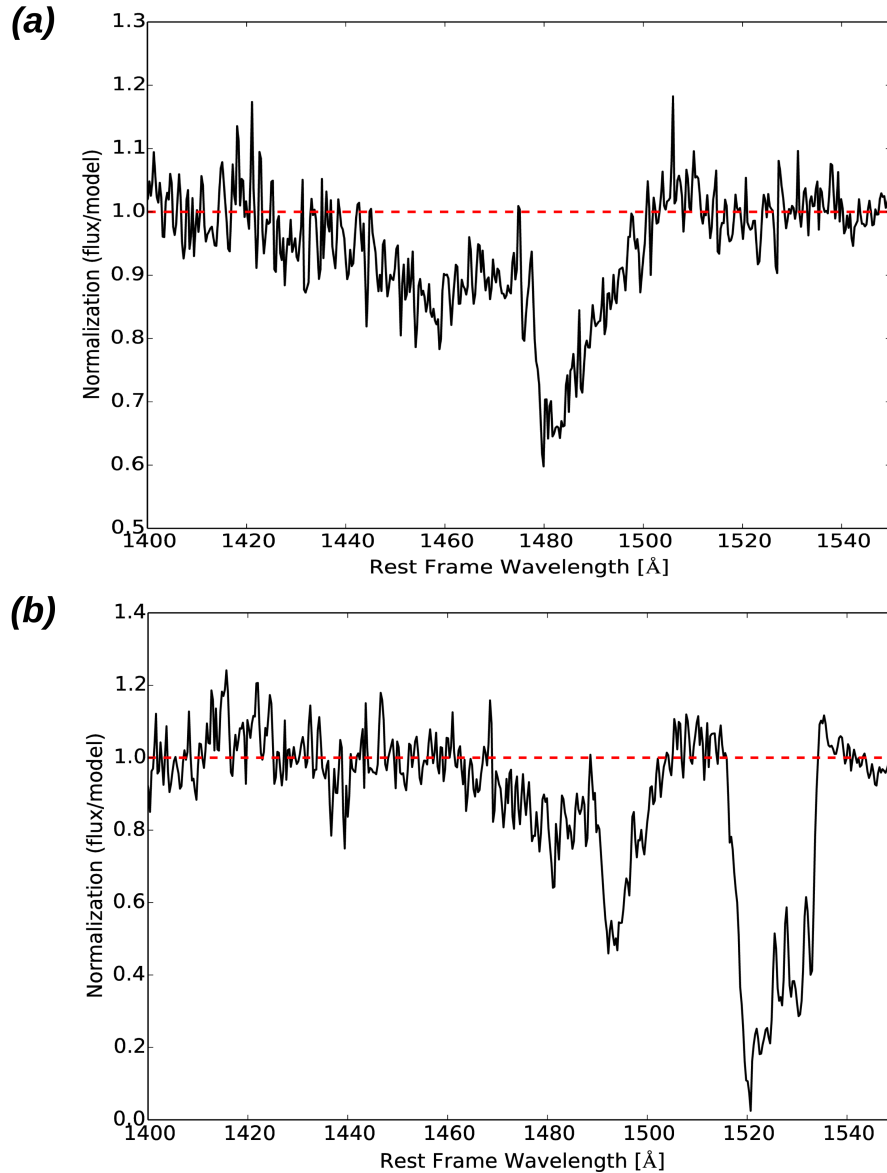


Figure 2.6: Normalized, resampled spectra produced using the method described in Section 2.2.3. The red, dashed line corresponds to a data-to-model ratio of 1. (a) Same object as in Figure 2.4. (b) Another object for comparison.

Chapter 3

Application of Machine Learning

3.1 The Similarity of Spectra

3.1.1 Pearson Product-moment Correlation Coefficients

The purpose of this study is to assess the similarity between broad absorption line quasar spectra based on the shapes of their C IV broad absorption lines in the wavelength range of the normalized, resampled spectra (1400–1550 Å). How a quantitative measure of the similarity between two spectra can be provided begins with the concept of *correlation*. In this work, we use *Pearson product-moment correlation coefficients*. The pairwise Pearson product-moment correlation coefficient between spectrum X and spectrum Y is defined as follows [Kaufman and Rousseeuw, 1990]

$$r_{XY} \equiv \frac{\sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (3.1)$$

where x_i and y_i are the normalized flux densities in wavelength bin i of spectrum X and Y , respectively, and $\bar{a} = \bar{x}, \bar{y}$ is the mean value of the normalized flux density for spectrum A with $A = X, Y$. The Pearson product-moment correlation coefficient can take values between -1 and $+1$. Values close to $+1$ (-1) are given to pairs of spectra that show a strong positive (negative) correlation with one another. Values

close to zero correspond to weak pairwise correlations between spectra.

Intuitively, r_{XY} takes the normalized flux densities in each of n wavelength bins for both spectrum X and Y and makes an ordered pair. We plot these n ordered pairs in the “flux-flux” plane and assess how well the data points can be fit to a straight line. If they all fall perfectly on a line with positive slope, the Pearson coefficient is exactly equal to $+1.0$ and the two spectra correlate perfectly with one another. Conversely, the more scattered the points are, the closer the Pearson coefficient is to 0.0 . If, however, the value of the Pearson coefficient is exactly -1.0 , the points all fall perfectly on a line with negative slope. The slope of the line does not contribute to the value of the Pearson coefficient. What is important is how closely all of the points fit a straight line. Since we are interested in grouping together similar spectra, no matter the sign of the correlation, we take the absolute value of the Pearson product-moment correlation coefficient when conducting our analyses.

We can arrange the pairwise absolute Pearson coefficients in matrix form to construct a *correlation matrix*. For our sample, we have 1,084 spectra against which we would like to compare those same 1,084 spectra. Our correlation matrix will have dimensions 1084×1084 . It will have values of 1.0 along the diagonal since self-correlations will return 100% similarity and it will be symmetric about the diagonal. The entry at position (x, y) for $x, y \in [1, 2, 3, \dots, 1084]$ will correspond to the absolute value of the pairwise Pearson coefficient between spectrum X and spectrum Y . The correlation matrix will have the following form:

$$\mathbf{R} \equiv \begin{bmatrix} 1.0 & r_{1,2} & r_{1,3} & \dots & r_{1,1084} \\ r_{2,1} & 1.0 & r_{2,3} & \dots & r_{2,1084} \\ r_{3,1} & r_{3,2} & 1.0 & \dots & r_{3,1084} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1084,1} & r_{1084,2} & r_{1084,3} & \dots & 1.0 \end{bmatrix}. \quad (3.2)$$

3.1.2 Distances Between Points

Another useful measure of the similarity between two data sets is *distance*. Abstractly, we treat each spectrum as a data point in some parameter space. If we take the “distance” between two spectra, similar spectra are closer to one another, whereas dissimilar spectra are further apart. We define the pairwise *Pearson distance* between spectrum X and spectrum Y to be [Xu and Wunsch]

$$d_{XY} \equiv 1 - |r_{XY}|. \quad (3.3)$$

Similar to the Pearson coefficient, we can arrange the pairwise Pearson distances in a matrix called the *distance matrix*. In this case, our (Pearson) distance matrix will have the following form:

$$\mathbf{D} \equiv \mathbf{1} - |\mathbf{R}| = \begin{bmatrix} 0.0 & d_{1,2} & d_{1,3} & \dots & d_{1,1084} \\ d_{2,1} & 0.0 & d_{2,3} & \dots & d_{2,1084} \\ d_{3,1} & d_{3,2} & 0.0 & \dots & d_{3,1084} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{1084,1} & d_{1084,2} & d_{1084,3} & \dots & 0.0 \end{bmatrix}, \quad (3.4)$$

where $|\mathbf{R}|$ represents the absolute value of the correlation matrix (taken elementwise).

An alternative measure of distance to the Pearson distance is the Euclidean distance. To take the Euclidean distance between two spectra, we must once again imagine the abstract parameter space inside which we plot our spectra as data points. In the Euclidean case, the pairwise distances between spectra are no longer the Pearson distances, but rather the Euclidean distances between the rows (or columns) of the Pearson distance matrix. For example, the pairwise Euclidean distance between Spectrum 1 and Spectrum 2 is given by

$$d_{1,2}^E = \sqrt{(0.0 - d_{2,1})^2 + (d_{1,2} - 0.0)^2 + (d_{1,3} - d_{2,3})^2 + \dots + (d_{1,1084} - d_{2,1084})^2}. \quad (3.5)$$

This pairwise distance is added to elements 1,2 and 2,1 of the new Euclidean distance

matrix. The full Euclidean distance matrix will have the following form:

$$\mathbf{D}_E = \begin{bmatrix} 0.0 & d_{1,2}^E & d_{1,3}^E & \cdots & d_{1,1084}^E \\ d_{2,1}^E & 0.0 & d_{2,3}^E & \cdots & d_{2,1084}^E \\ d_{3,1}^E & d_{3,2}^E & 0.0 & \cdots & d_{3,1084}^E \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{1084,1}^E & d_{1084,2}^E & d_{1084,3}^E & \cdots & 0.0 \end{bmatrix}. \quad (3.6)$$

This 1084×1084 Euclidean distance matrix contains the information about the similarity between spectra, but is not ordered in any way. We must thus sort, or *cluster*, this matrix in order to obtain meaningful results.

3.1.3 Distances Between Clusters

When clustering the distance matrix, we must introduce the concept of *cluster distance*. There are several ways to compute the distances between two clusters. In this work we use a method called *complete linkage* [McQuitty, 1960]. Consider two clusters of points in an arbitrary parameter space. We would like to measure the distance between those two clusters using the complete linkage method. To do this we find the two points, one in each cluster, for which the pairwise Euclidean distance between the points is a *maximum*. This is then defined as the cluster distance, d_C . Say, now, that we merge those two clusters into a single cluster. By using complete linkage, we ensure that each point in the new, merged cluster is reasonably close to every other point ($d \leq d_C$), thereby creating a more “spherically shaped” cluster. This measure of cluster distance is especially useful in our case since we would like the clusters of spectra to contain objects that are all reasonably similar to each other.

3.2 Agglomerative Hierarchical Clustering

In the previous sections, we discussed how the similarity between spectra is quantified with distance. Using these distance measures as our starting point, we can now

describe the application of a clustering algorithm to the Euclidean distance matrix (Equation 3.6).

The algorithm used in this work is called *agglomerative hierarchical clustering* [Xu and Wunsch]. What this means is that the algorithm takes a bottom-up or merging (agglomerative) approach to the clustering, and the results correspond to all possible clusterings (hierarchical). To visualize this, imagine we have a data set in 2 dimensions with each data point corresponding to an ordered pair in the x, y -plane (see Figure 3.2(a)). We would like to cluster the data based on their x - and y -values.

To do this, we begin by taking the pairwise Pearson product-moment correlation coefficients and placing them in a correlation matrix similar to Equation 3.2, but with dimensions equal to the number of data points in the x, y -plane (here, there are 20). Next, we calculate the Pearson distance matrix (see Equation 3.4). To take the pairwise Euclidean distances between each of the points, we use the rows of the Pearson distance matrix and the 20-dimensional version of Equation 3.5. This gives us a 20×20 Euclidean distance matrix. It is this matrix that will be used in the clustering algorithm.

The agglomerative hierarchical clustering algorithm begins by taking each data point and placing it in its own singleton cluster. The pairwise Euclidean distances between all of the clusters are calculated. By constructing the distance matrix, we have already done this for the first iteration. Using these distances, the two closest clusters are merged and our first iteration of the algorithm is complete. At the second iteration, we recompute the pairwise distances between clusters and update the distance matrix accordingly. Recall that the methods used to calculate the distances between individual points (Section 3.1.2) and between clusters (Section 3.1.3) are different. To calculate the distance between clusters we use complete linkage as described in Section 3.1.3. We again choose the two closest clusters and merge them. This corresponds to permuting the rows of and columns of the distance matrix to bring similar clusters closer together. The above method is iterated until all of the points are merged into a single cluster. See Figure 3.1 for an overview of the algorithm.

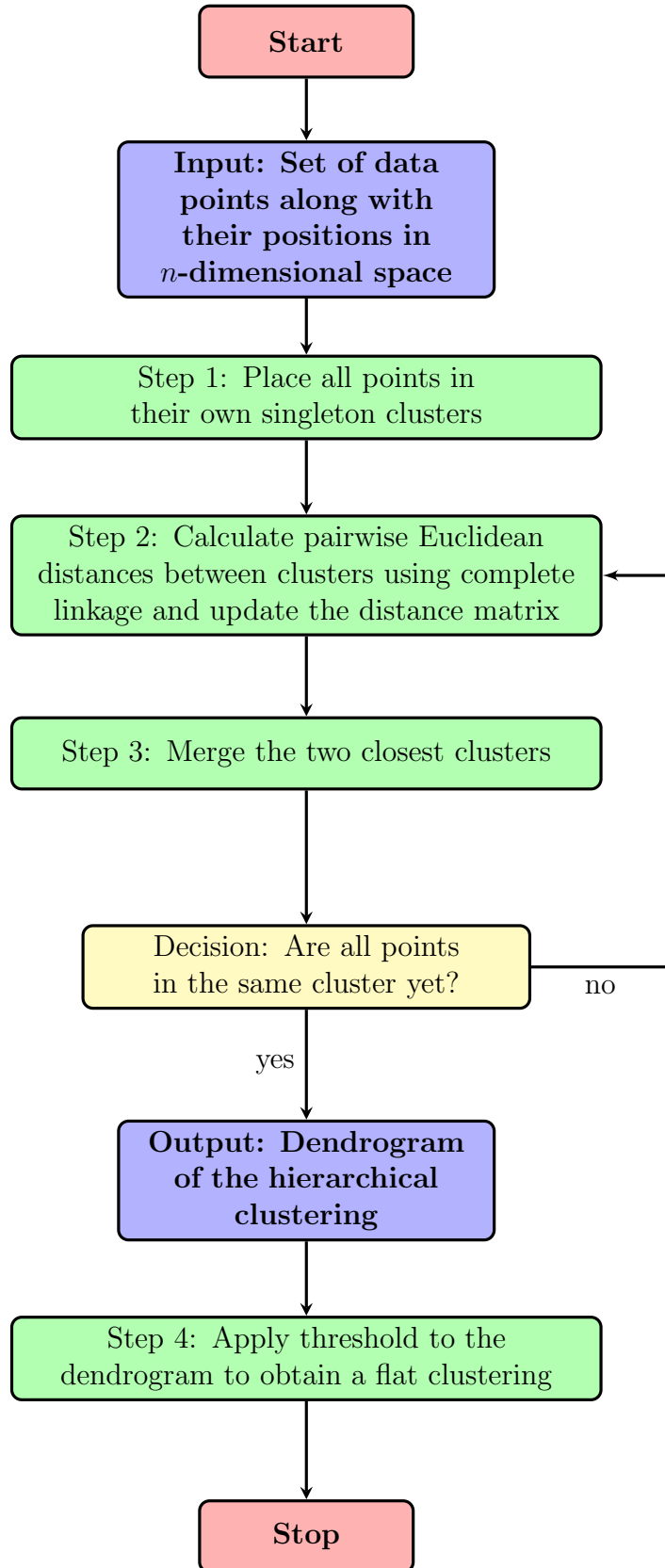


Figure 3.1: Summary of the agglomerative hierarchical clustering algorithm used to cluster an $n \times n$ Euclidean distance matrix.

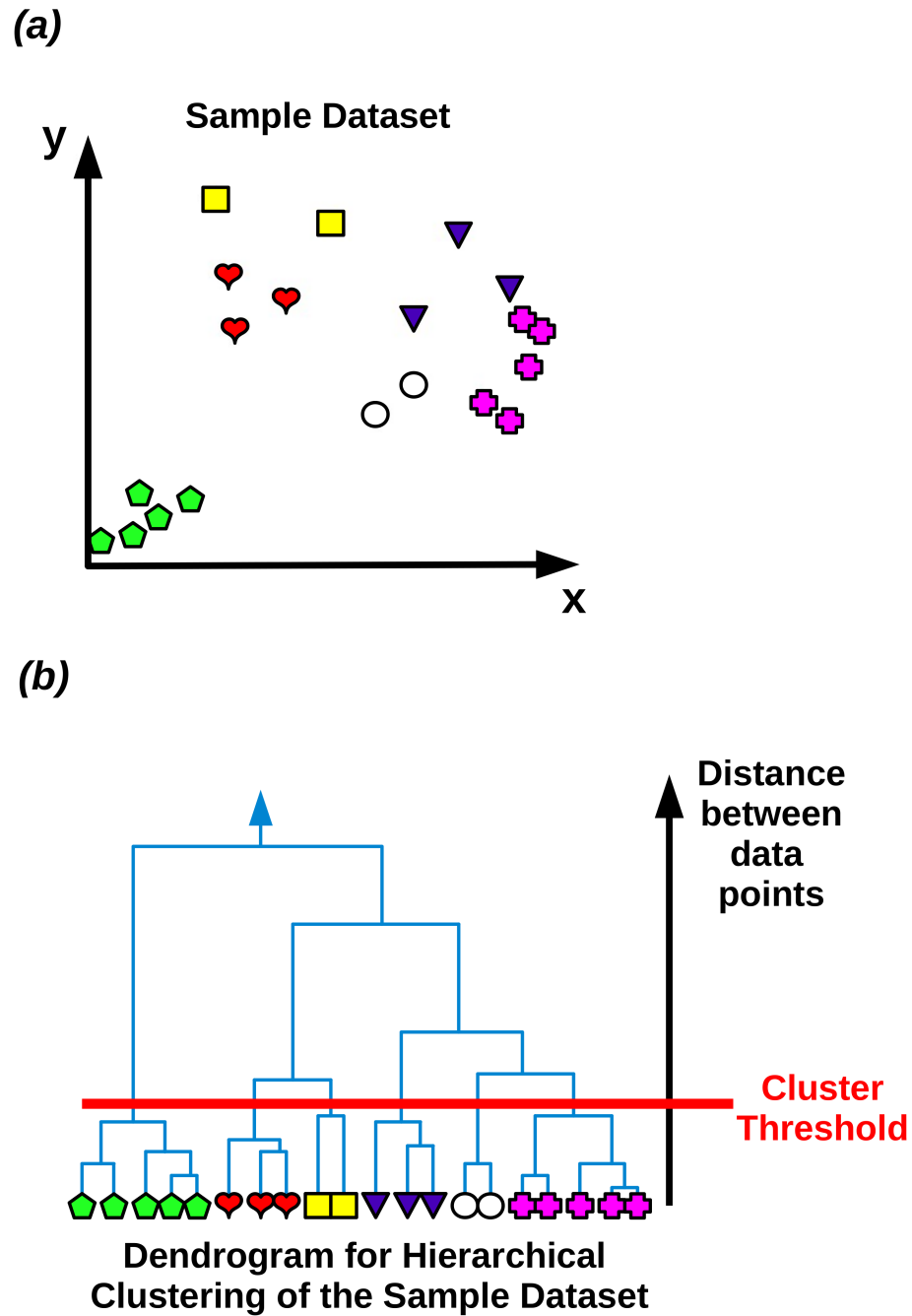


Figure 3.2: Agglomerative hierarchical clustering algorithm example (modified from Bovermann et al.). (a) Sample data set of 20 points in the x, y -plane. The colours and shapes of the data points correspond to the flat clustering obtained by applying a threshold to the dendrogram. (b) Dendrogram for the hierarchy of clusters formed after applying the algorithm. A threshold is applied that defines 6 clusters.

3.3 Interpretation of the Dendrogram

By applying the algorithm in Section 3.2 to a Euclidean distance matrix we obtain what is called a *dendrogram*. A dendrogram is a tree structure that contains the information of the cluster merging process taken by the algorithm. Figure 3.2(b) shows the resulting dendrogram from applying a hierarchical clustering algorithm to the sample data set in Figure 3.2(a). At the bottom of the dendrogram, each data point is in its own cluster and the distance between points within each of the clusters is zero. At the top of the dendrogram, all of the points are in one cluster and the distance between points within the cluster is at a maximum.

To extract information from a dendrogram we must choose a *threshold distance* (i.e., the maximum distance allowed between points within a cluster) and apply a horizontal cut. When the threshold is applied, each unique branch below the cut becomes its own cluster and a set of *flat* clusters is produced: that is, the threshold distance is the same for all clusters. If we take a top-down approach to interpreting the dendrogram, we can think of the largest cluster as being split into smaller clusters. As we move down the dendrogram, the distances between points in each of the clusters decreases or, in other words, data points within a single cluster are more alike.

The beauty of this algorithm lies in the hierarchical nature of the results it produces. We, the user, are permitted to make as many cuts as we wish and may follow the merging process, examining the results of the clustering at each new cut. In the next section, we will apply this algorithm to the 1,084 broad absorption line quasar spectra in our sample.

3.4 Clustering the Distance Matrix

We calculate the Pearson correlation matrix for the 1,084 broad absorption line quasar spectra in our sample using Equation 3.2. In Python, this corresponds to `mat.corr()` if `mat` is a 500×1084 `pandas.DataFrame` matrix containing 1,084 normalized, resampled spectra each with 500 wavelength values. Recall that we resampled our spec-

tra to a wavelength grid of 1400–1550 Å with 0.3 Å linear spacing. This gives us 500 normalized, resampled flux values for each broad absorption line quasar (since $(1550-1400)/0.3 = 500$). The Python `statsmodels.api.graphics.plot_corr` function was used to produce Figure 3.3, which shows the absolute-valued correlation matrix as a heat map.

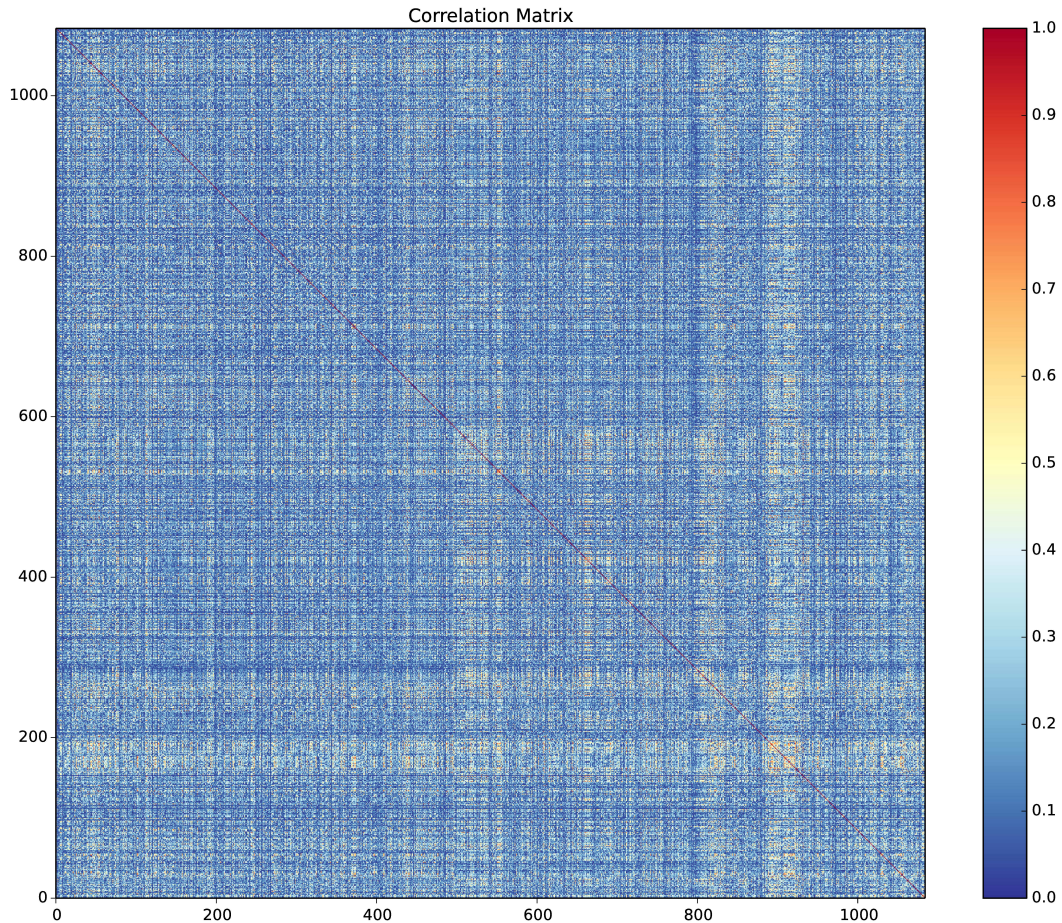


Figure 3.3: Absolute-valued Pearson correlation matrix for the 1,084 broad absorption line quasar spectra in our sample. Redder colours denote strong correlation or high similarity between spectra, whereas bluer colours denote little to no correlation or similarity between spectra. Each row (or column) of the matrix corresponds to a point in 1,084-dimensional space for a single broad absorption line quasar.

The next step is to calculate the square Euclidean distance matrix using Equation 3.5. Then, we may apply the agglomerative hierarchical clustering algorithm to obtain the clustered distance matrix and the dendrogram of the merging process.

We apply the built-in Python function `scipy.cluster.hierarchy.linkage` to

obtain the hierarchical clustering results encoded as a *linkage matrix*, Y . This function both calculates the pairwise distances between clusters at each iteration of the algorithm and sorts the distance matrix by permuting its rows and columns. We then apply the built-in `scipy.cluster.hierarchy.dendrogram` function to plot the dendrogram corresponding to the linkage matrix. Figure 3.4 shows the results of applying the hierarchical clustering algorithm to our data.

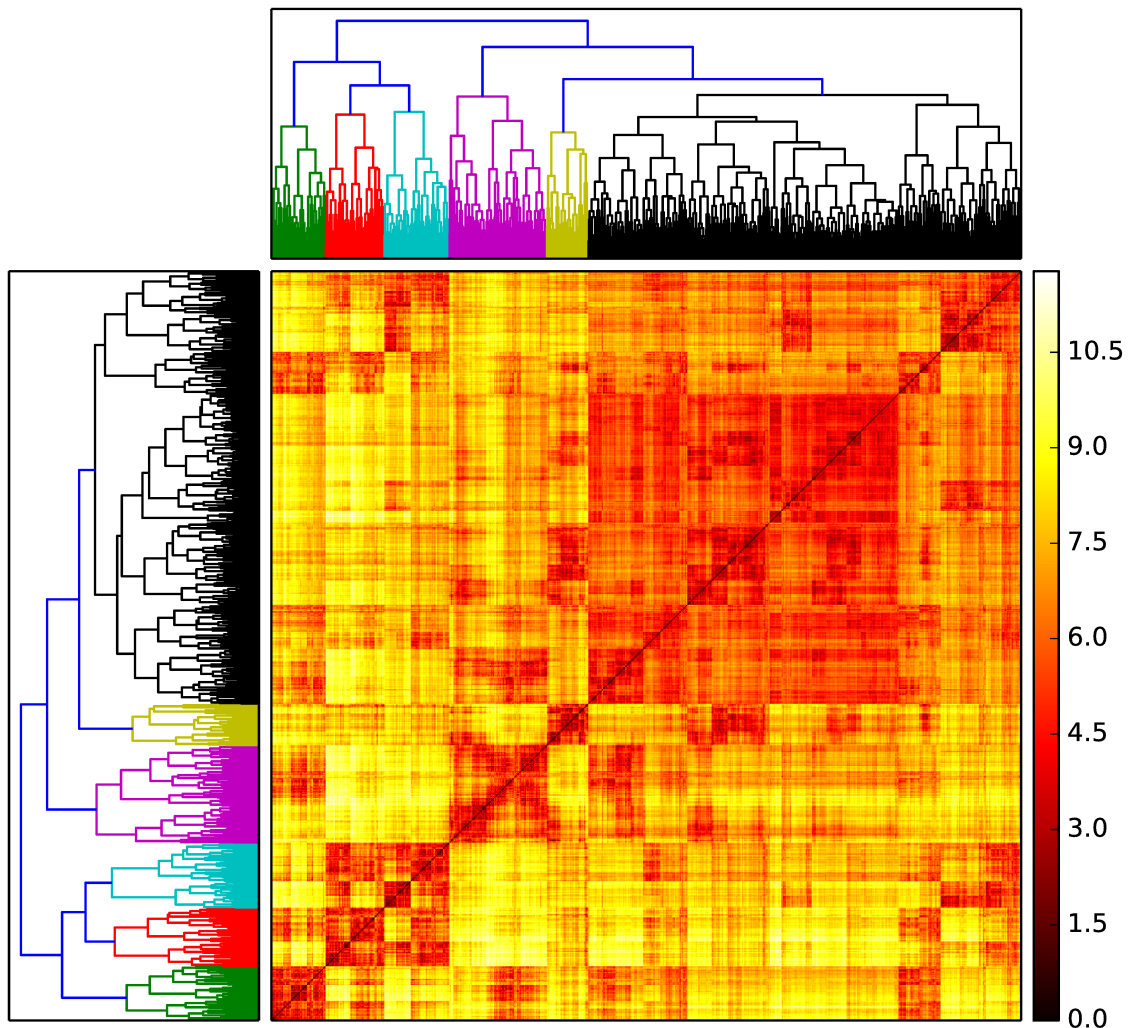


Figure 3.4: Clustered Euclidean distance matrix after the application of the agglomerative hierarchical clustering algorithm. Darker colours in the matrix correspond to smaller pairwise Euclidean distances between spectra. The two dendrograms are the same and their colours denote the flat clustering given the default distance threshold in Python.

Chapter 4

Results and Discussion

4.1 Median Composite Spectra

As shown by Figure 3.4 in Section 3.4, the application of a hierarchical clustering algorithm to our data set was successful: that is, distances between points within a cluster are low and distances between points in adjacent clusters are high. It is important to keep in mind that not all data sets are capable of being clustered. Take, for example, a data set nearly homogeneous in its features. This would likely produce one large cluster containing the majority of data points and a few smaller clusters containing outliers. In our case, however, there are several distinct clusters formed.

To analyze the results of the clustering, we begin at the top of the dendrogram and work our way down. We form flat clusterings of the data by applying the Python function `scipy.cluster.hierarchy.fcluster` to the linkage matrix, Y . We then visualize each of the clusters with a median composite spectrum.

To create a median composite spectrum of the objects within a cluster, we begin by defining a wavelength grid spanning 1100–4000 Å with linear spacing 0.3 Å. Each spectrum in the cluster is resampled to this wavelength grid. Next, we normalize each spectrum to the median value in the wavelength range ~ 1808 – 1817 Å. This is the region in the spectrum just blueward of the Al III, Si III], and C III] emission lines. Each spectrum in the cluster is $3\text{-}\sigma$ clipped to minimize outlying points such as cosmic ray spikes or excessive noise. The spectra are then median combined and

a standard deviation array is calculated for the median combined spectrum. This is done for each cluster within a flat clustering until 28 flat clusters are produced.

We then examine the median composite spectra at each of the cluster divisions and decide (visually) whether a division should be made. To make our decision we consider the following: (1) the standard deviation array of the spectrum being divided, (2) the shapes of the two spectra being formed by the division, and (3) the number of objects in the clusters formed as a result of the division. If the standard deviations around the C IV broad absorption line seem large, then a division is considered. Furthermore, if the broad absorption lines and/or the emission lines in the two spectra formed by the division look different, we make the division and continue down that branch of the dendrogram. If after a proposed division there are too few objects in the resulting clusters and the signal-to-noise ratio is too low, we do not make the division. This approach allows for a clustering to be made that is not flat; i.e., we are able to stop cluster divisions when they seem redundant and continue dividing those that seem important.

After deciding which cluster divisions should or should not be made, we conclude that the application of hierarchical clustering to our spectra produces at least 10 clusters. Figure 4.1 shows the median composite spectra of the 10 clusters in the region around the C IV broad absorption line (1320–1700 Å). Figure 4.2 shows the median composite spectra of the 10 clusters in the region around the Al III, Si III], and C III] emission lines (1800–2000 Å). Figure 4.3 shows the (not flat) dendrogram for the 10 clusters formed. For the remainder of the discussion, we refer to the spectra using the key in Table 4.1. We examine the spectra more closely in the next section.

4.2 Interpretation of the Results

We identify at least 10 different clusters of broad absorption line quasars based on the shapes of their C IV broad absorption line profiles. For clarity, we further regroup the 10 clusters into three categories based on the general shapes of their C IV broad absorption lines: (1) broad and extending to high terminal outflow velocities (Fig-

Table 4.1: Key used when referring to median composite spectra in Sections 4.2.1, 4.2.2, and 4.2.3.

Cluster Colour	Number of Objects	Key (used in text)
First Category		
Dark Raspberry	60	DR60
Light Pink	119	LP119
Black	186	B186
Second Category		
Yellow	61	Y61
Bright Pink	64	BP64
Purple	79	P79
Third Category		
Jean Blue	78	JB78
Green	117	G117
Light Turquoise	141	LT141
Deep Salmon	179	DS179

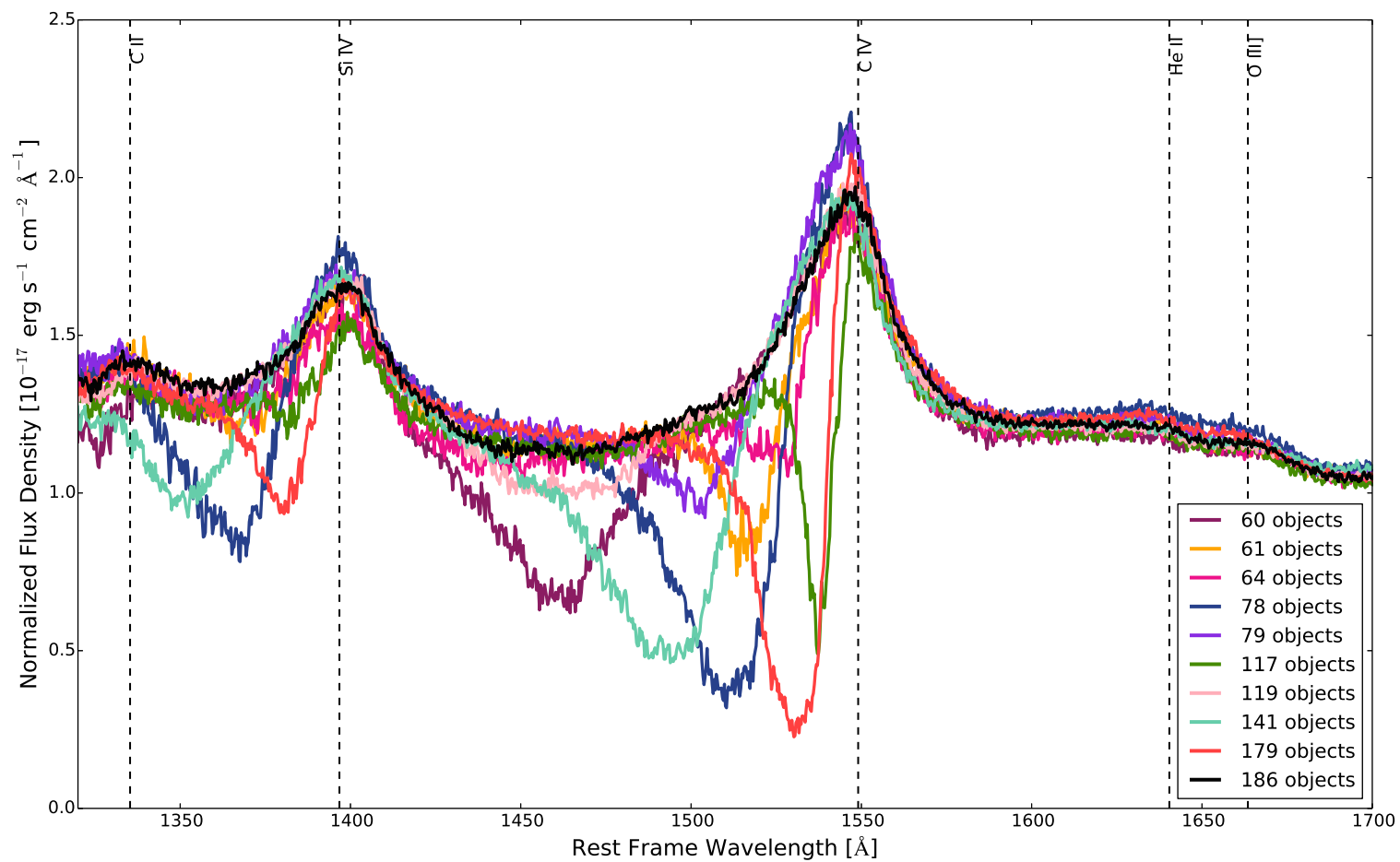


Figure 4.1: Median composite spectra for the 10 clusters. A wavelength range of 1320–1700 Å is chosen to emphasize differences in the C IV broad absorption line and the He II + O III] complex.

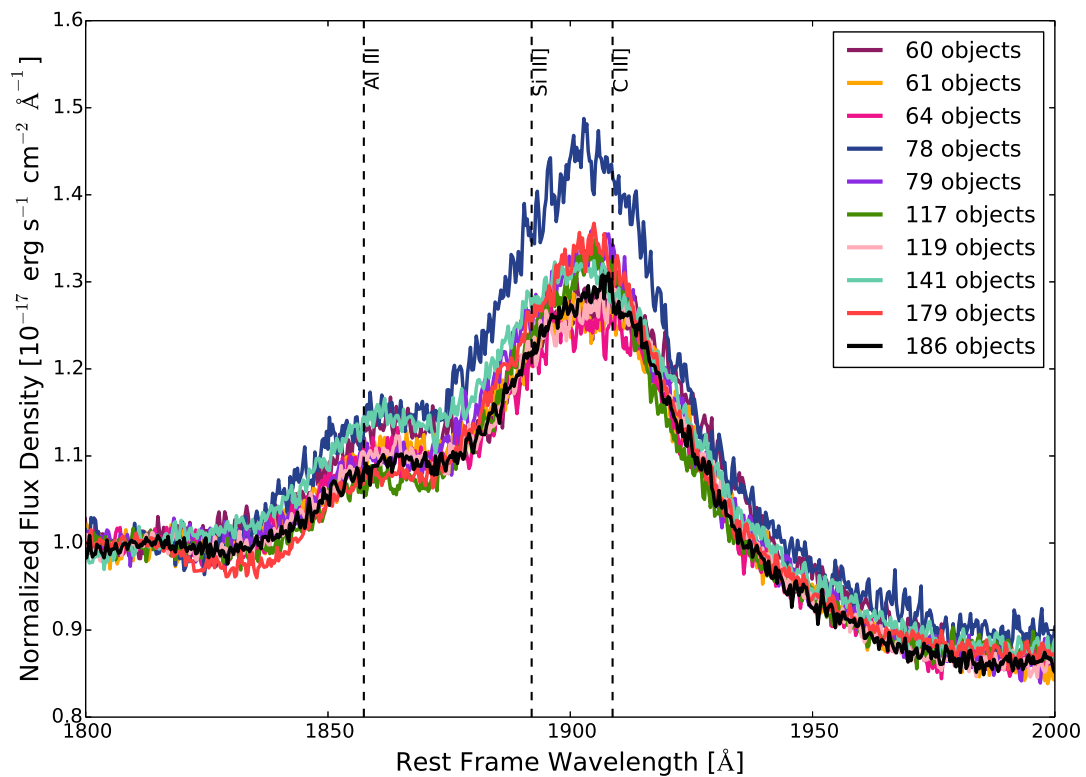


Figure 4.2: Median composite spectra for the 10 clusters. A wavelength range of 1800–2000 Å is chosen to emphasize differences in the Al III, Si III], and C III] emission lines. The cluster with the highest energy continuum is the jean blue cluster (78 objects).

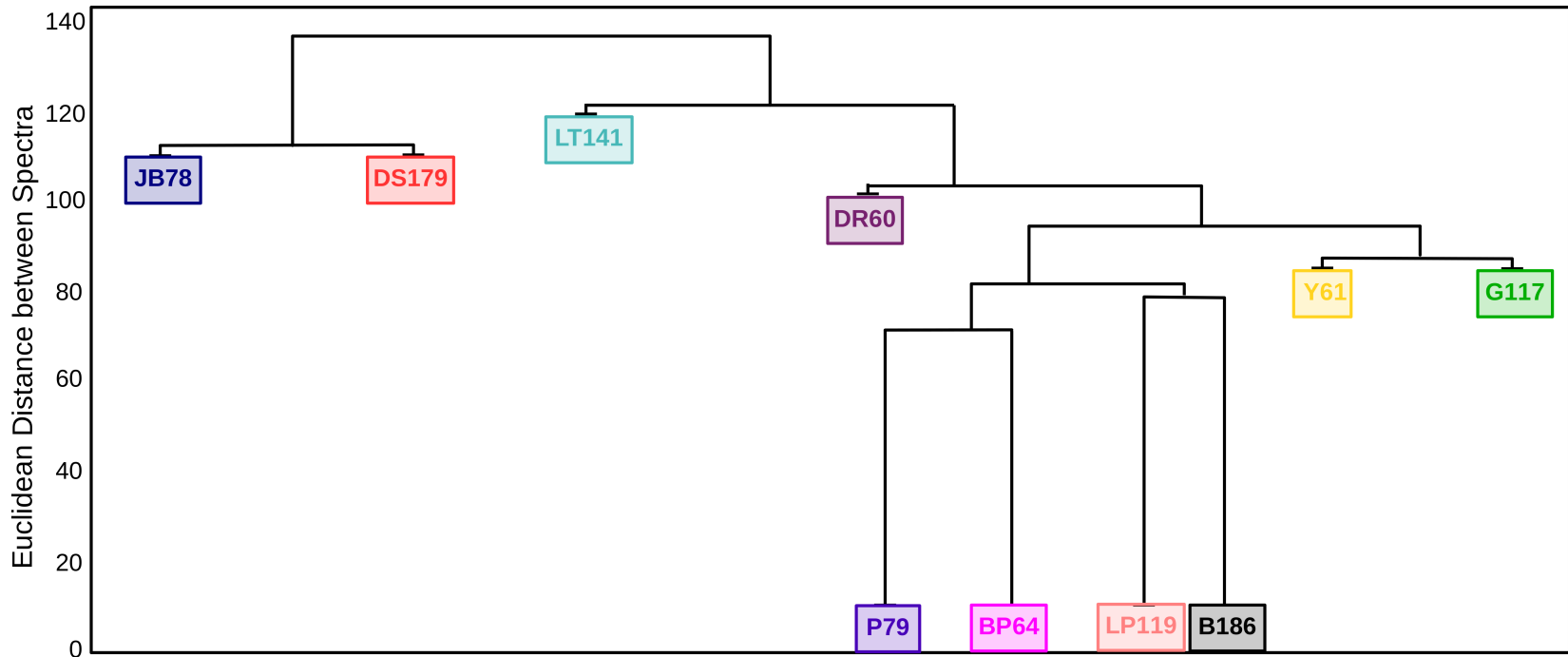


Figure 4.3: Dendrogram showing the 10 clusters formed after deciding which cluster divisions to keep/reject. Distance scale corresponds to that of a truncated 5-level dendrogram. Labels at the end of each branch use the key in Table 4.1. The **P79**, **BP64**, **LP119**, and **B186** clusters remain undivided at zero distance in this 5-level example.

ure 4.4), (2) shallow and extending to low terminal outflow velocities (Figure 4.6), and (3) deep and extending to any outflow velocity (Figure 4.8). Figures 4.5, 4.7, and 4.9 show the regions around the Al III, Si III], and C III] emission lines for the three categories. Figures 4.4 through 4.9 use the same colour code and refer to the same clusters as in Figures 4.1 and 4.2.

To interpret the results of these composites, we must understand the physical interpretation of the emission lines and how they relate to the broad absorption lines. Each of the emission lines has an associated *ionization potential*, which refers to the amount of energy needed to ionize an atom to a specific ionization state. For example, 18.8 eV is required to ionize the Al II ion to Al III [Lotz, 1967]. Table 4.2 summarizes the ionization potentials of the Si IV, C IV, He II, O III], Al III, Si III], and C III] ions in order of increasing ionization potential. Considering the values in Table 4.2, we refer to the Al III and Si III] emission lines as *low-ionization lines* and the Si IV, C IV, He II, O III], and C III] emission lines as *high-ionization lines*.

Table 4.2: Ionization potentials for emission lines around the C IV broad absorption line [Lotz, 1967]. Values reported are those required to ionize the atom in the previous ionization state to the state listed (i.e., C III \rightarrow C IV requires 47.9 eV).

Emission Line	Rest-Frame Wavelength (\AA)	Ionization Potential (eV)
Si III]	1892.03	16.3
Al III	1857.40	18.8
C III]	1908.73	24.4
He II	1640.42	24.6
Si IV	1396.76	33.5
O III]	1663.48	35.1
C IV	1549.06	47.9

The strength of an emission line (i.e., the integrated luminosity in the line) is dependent on several factors: the energetics of the ionizing continuum, the density of line-emitting gas, the probability of transition, and the temperature of the gas.

Assuming all of the above factors are held constant, the energetics of the ionizing continuum can be understood using the shape of the SED for a quasar. Recall that the SED contains information about the energy output of a source across a wide range of wavelengths (Figure 1.1 shows radio to X-ray). The continuum photon source (i.e., the accretion disk) is said to be “hard” if there is more flux density at X-ray wavelengths when compared to the UV. Conversely, the ionizing continuum is said to be “soft” when the SED shows more energy in the UV when compared to the X-ray. The energy output of the source, or the shape of the SED, affects which emission lines are present, and to what degree, in a spectrum. Therefore, we can use the relative strengths of high- to low-ionization lines to constrain the properties of the ionizing continuum.

If there is more relative strength in high-ionization lines, the continuum source carries more high energy photons (hard) and there is sufficient energy to ionize atoms to higher ionization states. Conversely, if there is more relative strength in low-ionization lines when compared to high-ionization lines, this tells us that the continuum source carries fewer high energy photons (soft). In this case, there is insufficient energy to ionize atoms to higher ionization states. If the ions of high-ionization lines are not present due to insufficient energy from the continuum, they cannot be excited electronically and therefore cannot contribute to the strength of the line.

With this in mind, we can consider the following as coarse diagnostics of the strength of the ionizing continuum: (1) the height of the He II and O III] emission line peaks above the local continuum, and (2) the ratio, Al III/C III], of the emission line peaks above the local continuum for the low- and high-ionization lines Al III and C III], respectively. If the peaks of the He II and O III] lines are high or if the Al III/C III] ratio is low, then the ionizing continuum is more energetic (hard).

The properties of the outflowing wind are sensitive to the properties of the ionizing continuum. In addition, the terminal velocity of the outflow is closely related to the location (radius) within the accretion disk from which the wind is launched [Gallagher et al., 2006, Chartas et al., 2003]. If winds are launched from smaller radii within the disk, they must have higher velocities in order to escape the black hole.

With respect to the radiation from the accretion disk, a less energetic continuum allows for outflows to be launched to higher velocities. This is because the higher the energy of the ionizing photons, the more likely the launched particles will be overionized. If, however, the continuum is less energetic (yet still energetic enough to contain photons at the resonant frequency of C IV), the ions in the wind will stay intact and will be able to be launched to higher velocities than they would be if the wind were overionized. We would thus expect to see C IV broad absorption lines with troughs extending to high outflow velocities to have a lower peak heights in the He II and O III] lines and higher Al III/C III] line ratios (as a result of a softer ionizing continuum). Conversely, C IV broad absorption lines that do not extend to high outflow velocities should have higher He II and O III] emission line peak heights and lower Al III/C III] line ratios (as a result of a harder ionizing continuum).

4.2.1 First Category

Using the above interpretations, we can examine the relative shapes of the composites in the 10 clusters. The first category contains the **DR60**, **LP119**, and **B186** clusters which contain 60, 119, and 186 objects, respectively (see Figures 4.4 and 4.5). Objects in these clusters have C IV troughs that are broad, relatively shallow, and extending to high outflow velocities.

Both the C IV and Si IV emission lines are very similar for all composites in the first category. The C IV broad absorption line in the **DR60** cluster extends to high outflow velocities and is moderately deep. Its He II and O III] emission is not very strong and its Al III/C III] line ratio is larger than that of the **LP119** or **B186** clusters. These results are consistent with our discussion above on the effects of the ionizing continuum on the broad absorption line profile.

The C IV broad absorption lines in the **LP119** and **B186** clusters have similar initial and final outflow velocities, He II and O III] emission line peaks, and Al III/C III] line ratios suggesting similar continuum source and broad line region properties. The only difference between the **LP119** and **B186** composites is the depth of the C IV broad absorption line. This may imply that the objects in these

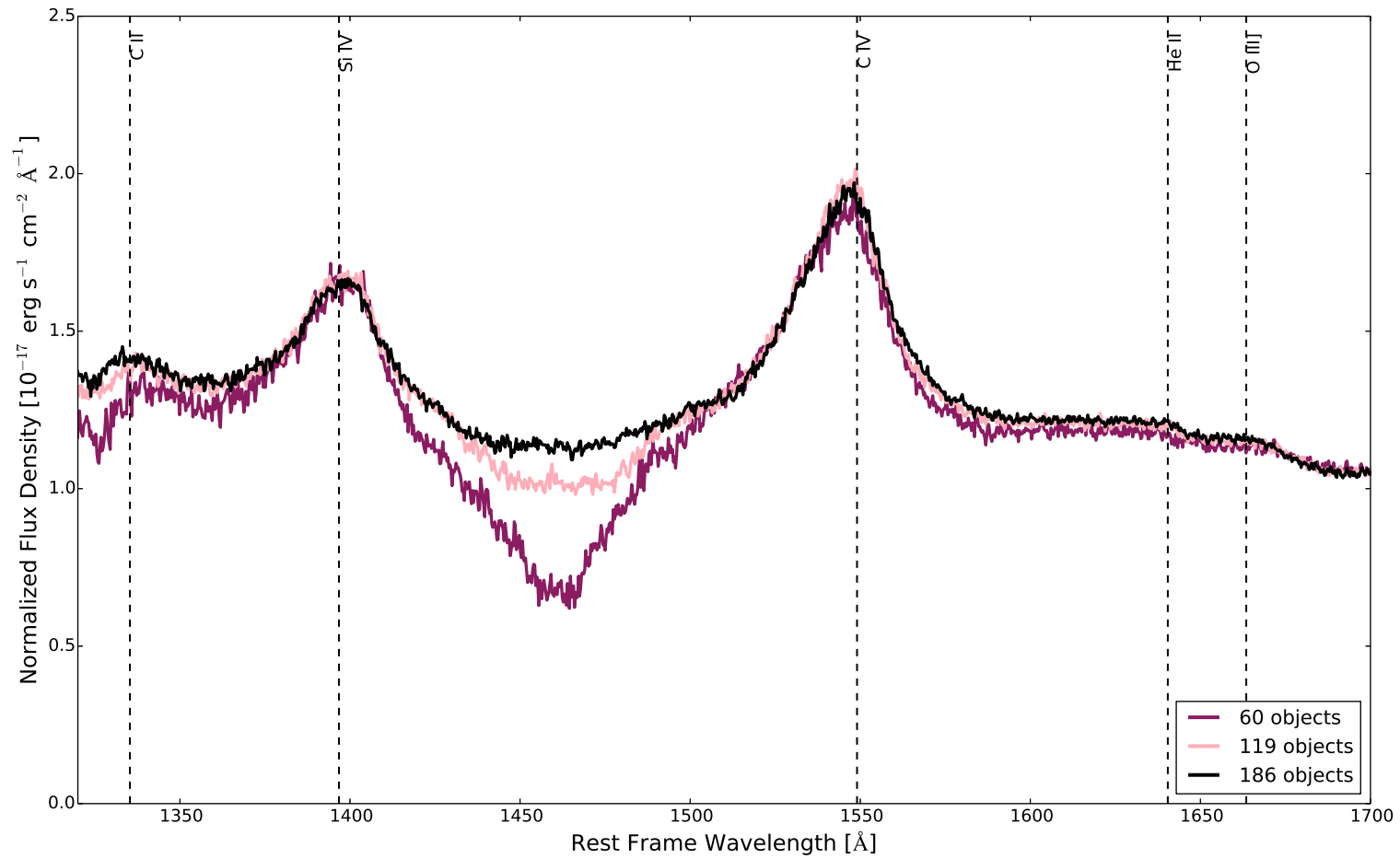


Figure 4.4: Median composite spectra in the wavelength range 1320–1700 \AA for the three clusters with broad, high-velocity C IV profiles (**DR60**, **LP119**, and **B186**) discussed in Section 4.2.1.

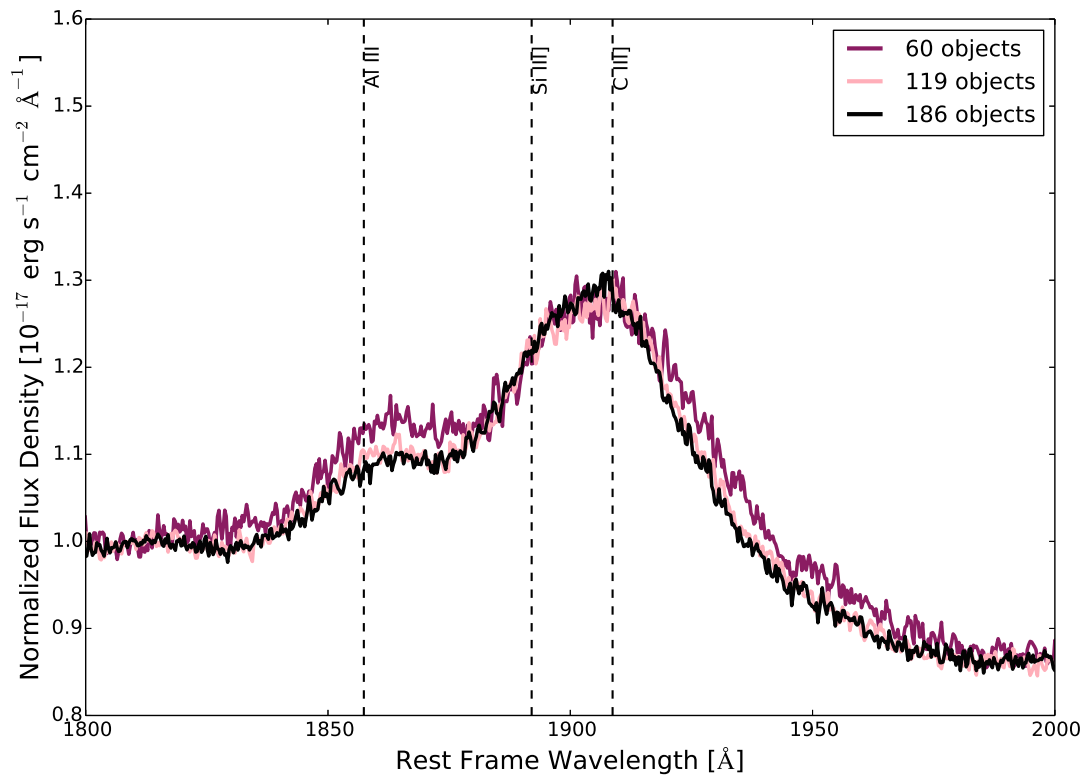


Figure 4.5: Median composite spectra in the wavelength range 1800–2000 Å for the three clusters with broad, high-velocity C IV profiles (**DR60**, **LP119**, and **B186**) discussed in Section 4.2.1.

two clusters come from the same parent class of objects, but contain different amounts of material along the line of sight (i.e., the wind is thicker in the **LP119** case).

4.2.2 Second Category

The second category contains the **Y61**, **BP64**, and **P79** clusters which contain 61, 64, and 79 objects, respectively (see Figures 4.6 and 4.7). Objects in these clusters have C IV broad absorption line troughs that are relatively narrow and shallow, and extend to lower outflow velocities.

The C IV and Si IV emission lines in the **Y61** and **BP64** clusters are similar, but not exactly the same. Furthermore, their He II and O III] emission line peaks and Al III/C III] line ratios are almost identical. This implies that the source of their continuum emission may be quite similar. As for the broad absorption lines in the **Y61** and **BP64** clusters, the **Y61** trough does not extend to as high or as low outflow velocities as the **BP64** trough. The blue wing of the **Y61** C IV emission line is less absorbed than that of the **BP64**.

Interestingly, the **BP64** cluster shows both a narrow, shallow, low-velocity component and a broad, shallow, high-velocity component. The **Y61** cluster also shows a very shallow high-velocity component in addition to its low-velocity component. The presence of these high-velocity components are supported by low peak heights in He II and O III] emission and lower Al III/C III] ratios when compared to the **P79** cluster.

The C IV broad absorption line in the **P79** cluster does not show a broad, high-velocity component as supported by its emission line properties. In addition, the **P79** C IV emission line is not heavily absorbed on its blue side.

4.2.3 Third Category

The third category contains the **JB78**, **G117**, **LT141**, and **DS179** clusters which contain 78, 117, 141, and 179 objects, respectively (see Figures 4.8 and 4.9). Objects in these clusters have relatively deep C IV broad absorption line profiles extending to

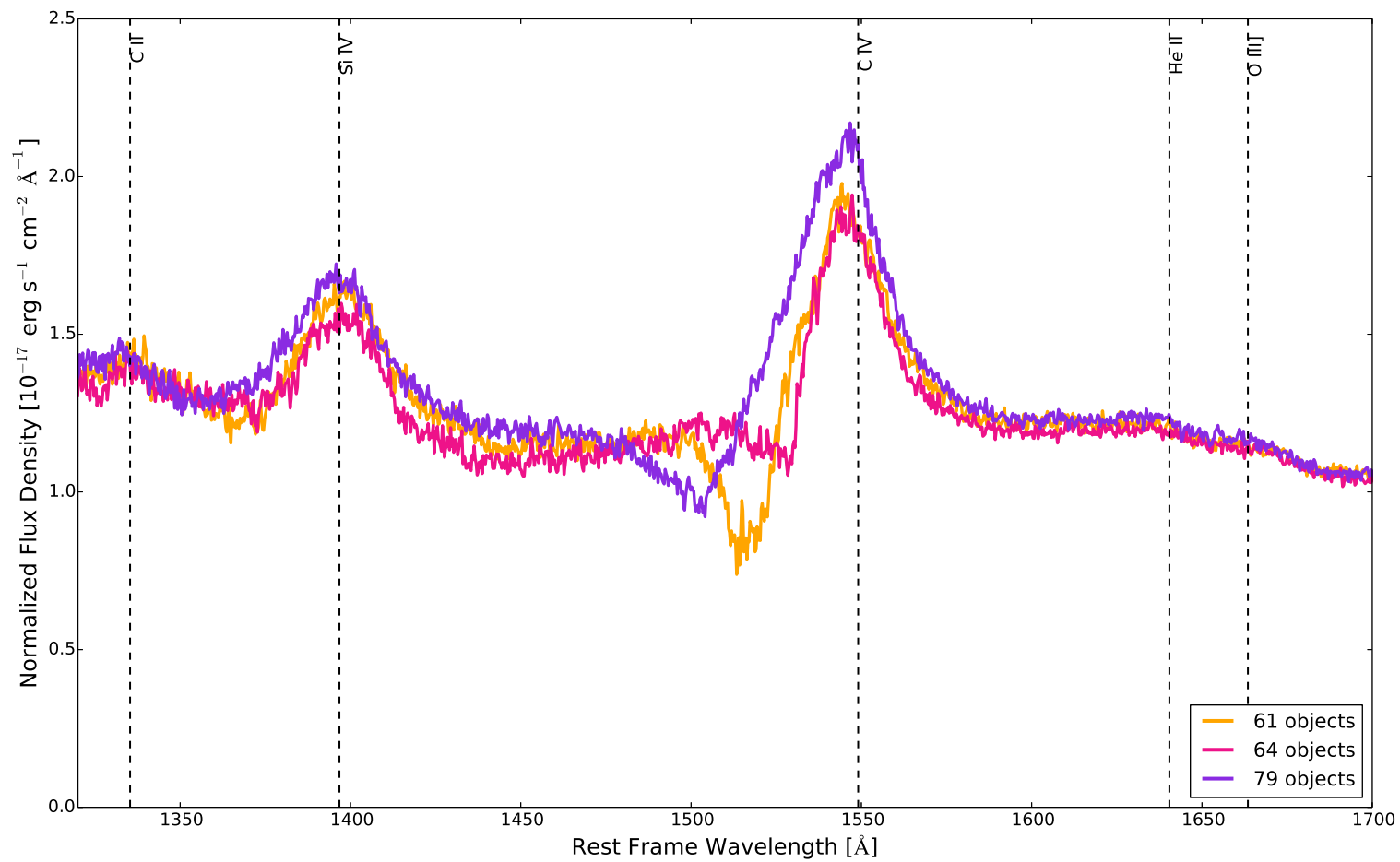


Figure 4.6: Same as Figure 4.4, but for the three clusters with shallow, low-velocity C IV profiles (Y61, BP64, and P79) discussed in Section 4.2.2.

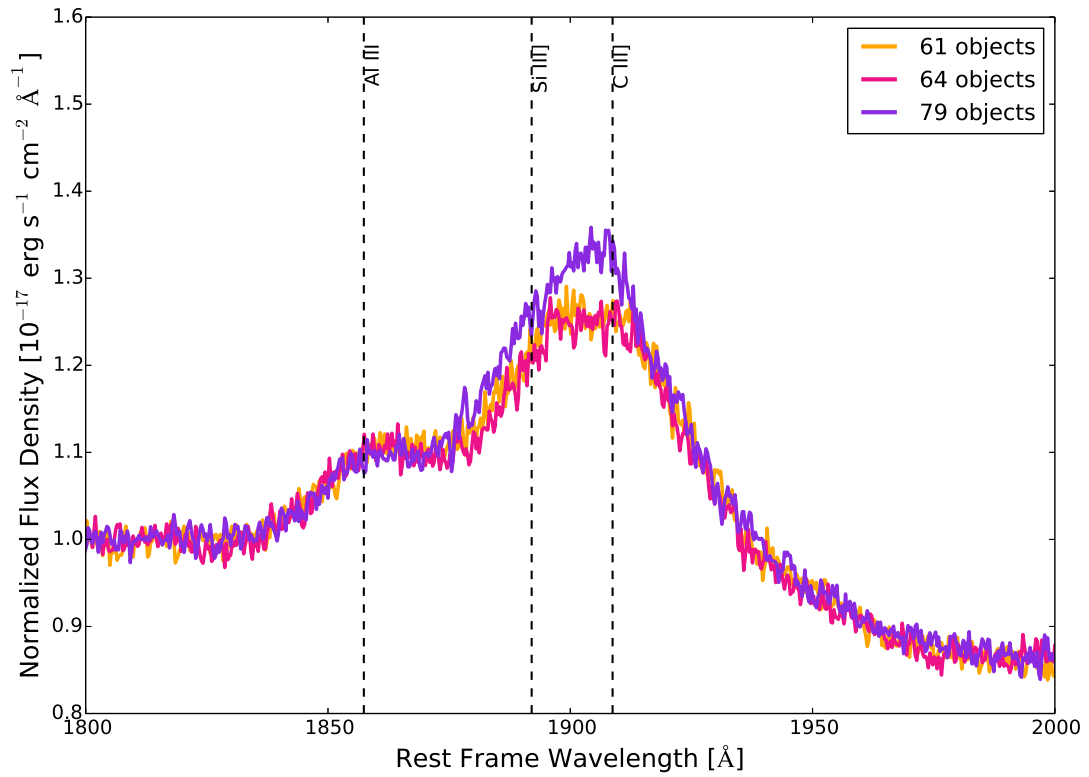


Figure 4.7: Same as Figure 4.5, but for the three clusters with shallow, low-velocity C IV profiles (**Y61**, **BP64**, and **P79**) discussed in Section 4.2.2.

any outflow velocity.

The C IV, Si IV, He II, and O III] emission lines are the strongest and the Al III/C III] ratio is the lowest, out of all 10 clusters, in the **JB78** cluster. The high-velocity end of the C IV broad absorption line in the **JB78** cluster is the least absorbed out of all 10 clusters. The above observations imply a highly energetic ionizing continuum. In addition, the onset of initial outflow velocity in the **JB78** cluster is gradual. This means we are seeing a range of initial outflow velocities, which could imply a range of different launching radii for the winds.

In the **G117** cluster, there appear to be two components to the C IV broad absorption line: a narrow, deep, low-velocity component and a broad, shallow, high-velocity component. In addition, the low-velocity absorption is embedded in the C IV emission line and completely separated from the broad component since the blue wing of the C IV emission line is clearly visible in the composite. When comparing this profile to the emission line properties, the He II and O III] emission is weak (implying a softer continuum) and the Al III/C III] ratio is low when compared to most other clusters (implying a harder continuum). So where, then, does the narrow, low-velocity absorption come from and why is there a contradiction in the continuum energetics?

Since broad absorption line winds are not smooth, but rather clumpy, the two components could arise as a result of different parts of the wind moving at different velocities. It could also be the case that we are observing different zones within the same wind- one being launched from close in and the other from further out in the accretion disk. Alternatively, it could be that the continuum of the **G117** cluster is reddened by dust along the line of sight, but exterior to the wind. In this case, the high-velocity component would only be an artifact of the redder continuum and there would likely only be a low-velocity component. This could imply that since the continuum of the **G117** cluster would be in fact quite strong (comparable to that of the **DS179** cluster, see Figure 4.9), the particles in the wind would be destroyed before they could be launched and accelerated. Alternatively, the possibly redder continuum of the **G117** cluster may imply a different viewing angle. When the wind is launched, the velocities of the particles at the base of the wind are lower and as

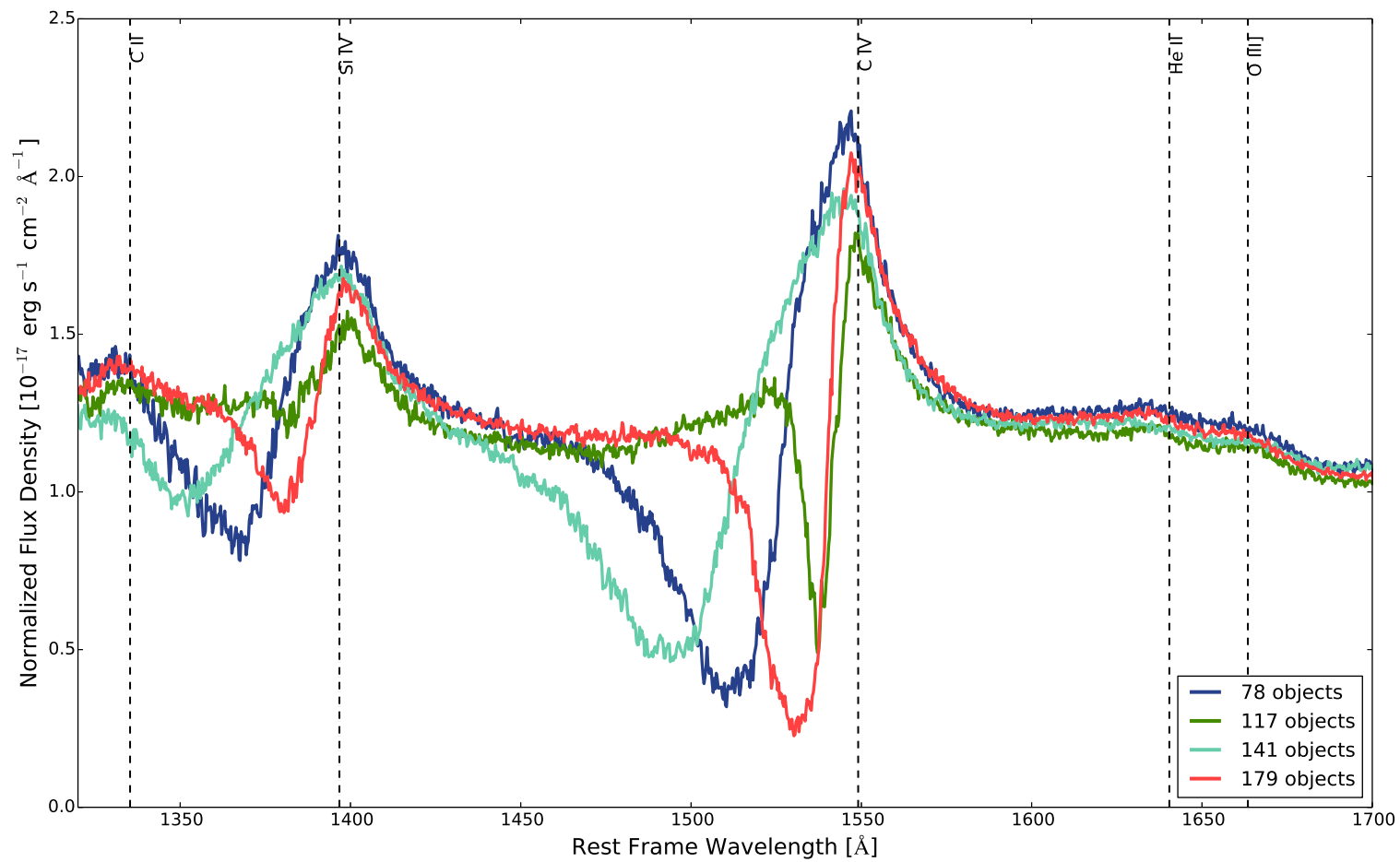


Figure 4.8: Same as Figure 4.4, but for the four clusters with deep C IV profiles (**JB78**, **G117**, **LT141**, and **DS179**) discussed in Section 4.2.3.

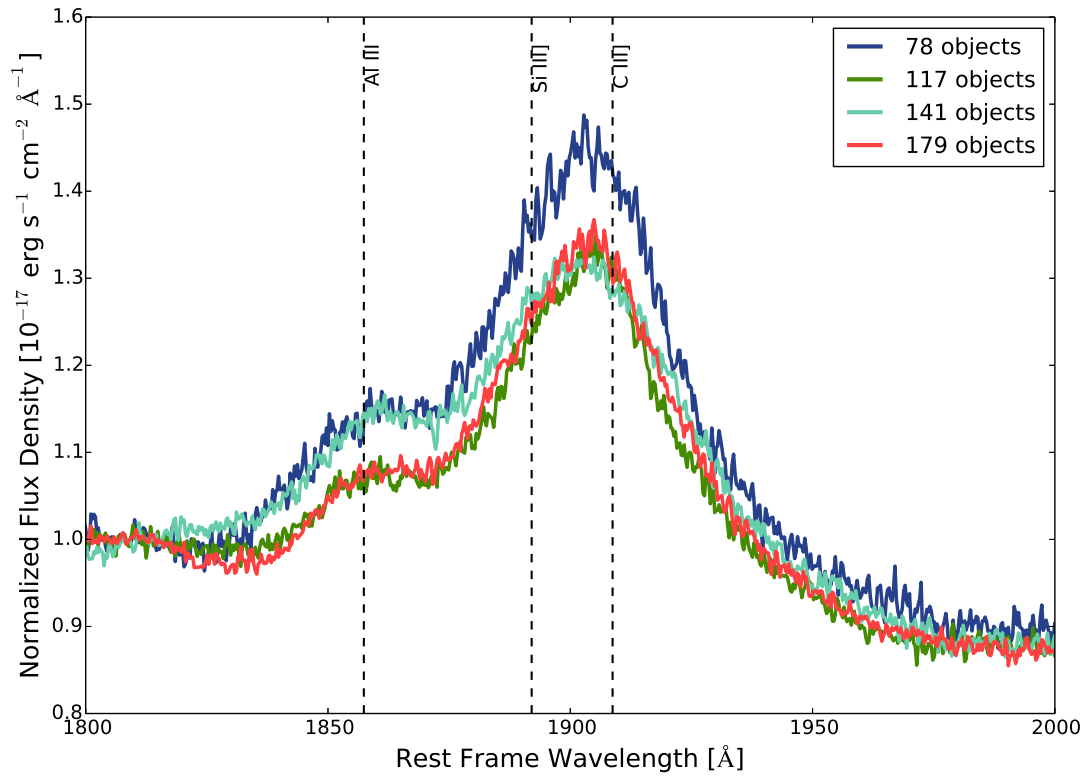


Figure 4.9: Same as Figure 4.5, but for the four clusters with deep C IV profiles (JB78, G117, LT141, and DS179) discussed in Section 4.2.3.

the wind is driven up and out of the disk, it is accelerated to higher velocities. If the viewing angle were less inclined, we would be looking through more dust (causing a redder continuum) and toward the base of the wind (accounting for the low-velocity component).

The onset of initial outflow velocity in the low-velocity component of the **G117** cluster shows a sharp drop. This sharp drop implies that there is a lot of gas along the line of sight at low *radial* (i.e., line of sight) velocities. In addition, we could be looking through the wind in a direction transverse to the velocity vector of the wind (i.e., the vector starting from the base and pointing in the direction of motion of the wind). If we were looking transverse to this vector, we would see a lot of gas at a single velocity, resulting in a sharp drop in the broad absorption line profile.

The **LT141** cluster shows a broad, deep C IV broad absorption line profile with a gradual initial onset in outflow velocities, similar to the **JB78** cluster, but more gradual, and reaching a trough minimum at a higher velocity than that of the **JB78** cluster. The C IV emission line, however, is shifted to slightly shorter wavelengths than the C IV emission in most of the other clusters.

The **DS179** cluster has a C IV broad absorption line with a narrow, deep, low-velocity component. The initial onset of the outflow is sharp, similar to that of the **G117** cluster. All of the emission lines in the **DS179** cluster are similarly shaped to those in the **JB78** cluster, but are not as strong. This could mean that the ionizing continuum in both the **JB78** and **DS179** clusters is similar and that in the **JB78** case the covering fraction of the broad line region gas is higher, allowing it to absorb and re-emit more of the light from the accretion disk.

Finally, the **G117** cluster shows similar emission line properties (Al III, Si III], and C III]) to the **DS179** and **JB78** clusters. This could mean that the quasars in these three clusters (**JB78**, **DS179**, and **G117**) are similar in their physical nature, but are at different levels of brightness. This claim could be further validated using other properties of the quasar such as the black hole mass. If true, this would be an interesting result since although the emission line properties are similar (i.e., the broad line region properties are similar), the wind shows differences. In this case,

the brightest quasars (in the **JB78** cluster) are able to launch their winds to higher outflow velocities than the faintest ones (in the **G117** cluster).

4.2.4 Summary

The above discussion can be used to further separate our clusters into at least 5 classes of physically distinct objects. If we merge the **BP64** and **Y61** clusters, the **B186** and **LP119** clusters, and the **P79**, **JB78**, **DS179**, and **G117** clusters, then we have 5 classes of objects total. We have merged the **JB78**, **P79**, **DS179**, and **G117** clusters because their emission line properties are similar (assuming that the continuum of the **G117** cluster is reddened). This gives us the following 5 classes of objects:

1. Broad, shallow, medium- to high-velocity C IV troughs with a less energetic ionizing continuum. This category, which combines the **B186** and **LP119** clusters, accounts for about 28% of our sample.
2. Broad, moderately deep, highest-velocity C IV troughs with a less energetic ionizing continuum. This category, which includes the objects in the **DR60** cluster, accounts for about 5.5% of our sample.
3. Two-component C IV troughs (broad, shallow, high-velocity and narrow, shallow, low-velocity) with a less energetic ionizing continuum. This category, which combines the **BP64** and **Y61** clusters, accounts for about 11.5% of our sample.
4. Broad, deep, medium- to high-velocity C IV troughs with a less energetic ionizing continuum. This category, which includes the objects in the **LT141** cluster, accounts for about 13% of our sample.
5. Narrow, low-velocity C IV troughs with a more energetic ionizing continuum. This category, which combines the **JB78**, **P79**, **DS179**, and **G117** clusters, accounts for about 42% of our sample.

The above results can be compared with recent results in the literature. For example, Tammour et al. (in prep.) applies a K -means clustering algorithm to a sample of broad absorption line quasars in the Gibson et al. [2009] catalog. They cluster their spectra based on the properties (equivalent width, v_{min} , and v_{max}) of the C IV broad absorption line. The results of their clustering show similar composite spectra, in that objects with more energetic ionizing continua preferentially show broad absorption lines with lower outflow velocities.

Chapter 5

Conclusions

We have applied a hierarchical clustering algorithm to a sample of 1,084 broad absorption line quasar spectra from the SDSS DR5. We identify at least 5-10 subclasses of broad absorption line quasars based on the shapes of their C IV broad absorption line profiles. We use median combined spectra of the objects in each cluster to validate the clustering technique. By comparing the shapes of the C IV profiles within clusters of objects to other features in the ultraviolet spectrum, we can examine how the properties of the underlying continuum are intimately connected with the properties of the broad absorption line wind. The results of this work show similarities to the results found in the literature using both different clustering techniques and different parameters over which to perform the clustering [Tammour et al., in prep.].

We emphasize that the classification put forth in this work is one of many possible interpretations of the results of the clustering. Nevertheless, physical insight can still be gained by comparing the properties of the spectra in each of the clusters. Broad absorption line shapes are governed by a variety of contributing effects such as the geometry of the outflow, the launching radius of the wind, and even the strength of the continuum radiation from the accretion disk. Thus, we can constrain the structure and dynamics of the outflowing wind by understanding how they respond to changes in factors such as the strength of the ionizing continuum and the geometry of the outflow.

Bibliography

- J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, et al. The Fifth Data Release of the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 172:634–644, October 2007. doi: 10.1086/518864.
- J. A. Baldwin, G. J. Ferland, K. T. Korista, et al. Very High Density Clumps and Outflowing Winds in QSO Broad-Line Regions. *The Astrophysical Journal*, 461:664, April 1996. doi: 10.1086/177093.
- A. Baskin, A. Laor, and F. Hamann. The average absorption properties of broad absorption line quasars at $800 < \lambda_{rest} < 3000 \text{ \AA}$, and the underlying physical parameters. *Monthly Notices of the Royal Astronomical Society*, 432:1525–1543, June 2013. doi: 10.1093/mnras/stt582.
- T. Bovermann, J. Bohrhuber, and H. Ritter. In *Proceedings of the 14th International Conference on Auditory Display*.
- G. Chartas, W. N. Brandt, and S. C. Gallagher. XMM-Newton Reveals the Quasar Outflow in PG 1115+080. *The Astrophysical Journal*, 595:85–93, September 2003. doi: 10.1086/377299.
- M. Elitzur and I. Shlosman. The AGN-obscuring Torus: The End of the “Doughnut” Paradigm? *The Astrophysical Journal Letters*, 648:L101–L104, September 2006. doi: 10.1086/508158.
- M. Elvis, B. J. Wilkes, J. C. McDowell, et al. Atlas of quasar energy distributions.

- The Astrophysical Journal Supplement Series*, 95:1–68, November 1994. doi: 10.1086/192093.
- P. J. Francis, P. C. Hewett, C. B. Foltz, et al. A high signal-to-noise ratio composite quasar spectrum. *The Astrophysical Journal*, 373:465–470, June 1991. doi: 10.1086/170066.
- S. C. Gallagher and J. E. Everett. Stratified Quasar Winds: Integrating X-ray and Infrared Views of Broad Absorption-line Quasars. In L. C. Ho and J.-W. Wang, editors, *The Central Engine of Active Galactic Nuclei*, volume 373 of *Astronomical Society of the Pacific Conference Series*, page 305, October 2007.
- S. C. Gallagher, W. N. Brandt, G. Chartas, et al. An Exploratory Chandra Survey of a Well-defined Sample of 35 Large Bright Quasar Survey Broad Absorption Line Quasars. *The Astrophysical Journal*, 644:709–724, June 2006. doi: 10.1086/503762.
- R. R. Gibson, L. Jiang, W. N. Brandt, et al. A Catalog of Broad Absorption Line Quasars in Sloan Digital Sky Survey Data Release 5. *The Astrophysical Journal*, 692:758–777, February 2009. doi: 10.1088/0004-637X/692/1/758.
- Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley series in probability and mathematical statistics. Wiley, New York, 1990. ISBN 0-471-87876-6. A Wiley-Interscience publication.
- Ajit K. Kembhavi and J.V. Narlikar. *Quasars and Active Galactic Nuclei: An Introduction*. Cambridge University Press, 1999. ISBN 9780521479899.
- Wolfgang Lotz. Ionization potentials of atoms and ions from hydrogen to zinc*. *J. Opt. Soc. Am.*, 57(7):873–878, Jul 1967. doi: 10.1364/JOSA.57.000873.
- C. B. Markwardt. Non-linear Least-squares Fitting in IDL with MPFIT. In D. A. Bohlender, D. Durand, and P. Dowler, editors, *Astronomical Data Analysis Software and Systems XVIII*, volume 411 of *Astronomical Society of the Pacific Conference Series*, page 251, September 2009.

- Louis L. McQuitty. Hierarchical linkage analysis for the isolation of types. 20(1): 55–67, 1960. doi: 10.1177/001316446002000106.
- S. L. Morris. The covering factor of quasar broad absorption line clouds. *The Astrophysical Journal Letters*, 330:L83–L86, July 1988. doi: 10.1086/185210.
- N. Murray, J. Chiang, S. A. Grossman, et al. Accretion Disk Winds from Active Galactic Nuclei. *The Astrophysical Journal*, 451:498, October 1995. doi: 10.1086/176238.
- M. Nenkova, Ž. Ivezić, and M. Elitzur. Dust Emission from Active Galactic Nuclei. *The Astrophysical Journal Letters*, 570:L9–L12, May 2002. doi: 10.1086/340857.
- H. Netzer. Active Galactic Nuclei: Basic Physics and Main Components. In D. Alloin, editor, *Physics of Active Galactic Nuclei at all Scales*, volume 693 of *Lecture Notes in Physics*, Berlin Springer Verlag, page 1, 2006. doi: 10.1007/3-540-34621-X_1.
- B. M. Peterson. *An Introduction to Active Galactic Nuclei*. February 1997.
- T. A. Reichard, G. T. Richards, P. B. Hall, et al. Continuum and Emission-Line Properties of Broad Absorption Line Quasars. *The Astronomical Journal*, 126:2594–2607, December 2003. doi: 10.1086/379293.
- G. T. Richards, N. E. Kruczek, S. C. Gallagher, et al. Unification of Luminous Type 1 Quasars through C IV Emission. *The Astronomical Journal*, 141:167, May 2011. doi: 10.1088/0004-6256/141/5/167.
- J. Surdej and D. Hutsemekers. Geometry of the mass-outflows around broad absorption line QSOs and formation of the complex Ly-alpha + N V line profile. *Astronomy and Astrophysics*, 177:42–50, May 1987.
- A. Tammour, S. C. Gallagher, N. Filiz Ak, et al. Constraining the Diversity of Broad-Absorption Troughs in BAL Quasars with Unsupervised Clustering Analysis.

- D. A. Turnshek, C. J. Grillmair, C. B. Foltz, et al. QSOs with PHL 5200-like broad absorption line profiles. *The Astrophysical Journal*, 325:651–670, February 1988. doi: 10.1086/166036.
- D. E. Vanden Berk, G. T. Richards, A. Bauer, et al. Composite Quasar Spectra from the Sloan Digital Sky Survey. *The Astronomical Journal*, 122:549–564, August 2001. doi: 10.1086/321167.
- R. J. Weymann, S. L. Morris, C. B. Foltz, et al. Comparisons of the emission-line and continuum properties of broad absorption line and normal quasi-stellar objects. *The Astrophysical Journal*, 373:23–53, May 1991. doi: 10.1086/170020.
- R. Xu and D.C. Wunsch. *Clustering*. IEEE series on computational intelligence. ISBN 9780470276808.