DETECTION AND ANALYSIS OF QUASAR SPECTROSCOPIC ANOMALIES

A THESIS

submitted in partial fulfillment of the requirements

for the award of the dual degree of

Bachelor of Science - Master of Science

in

PHYSICS

by

ARIHANT TIWARI

(19050)

DEPARTMENT OF PHYSICS
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH BHOPAL
BHOPAL - 462066
April 2024

भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान भोपाल

Indian Institute of Science Education and Research Bhopal

(Estb. By MHRD, Govt. of India)

CERTIFICATE

This is to certify that **Arihant Tiwari**, BS-MS (Physics), has worked on the project entitled **'Detection and Analysis of Quasar Spectroscopic Anomalies'** under my supervision and guidance.

April 2024
IISER Bhopal

Dr. Vivek M
Indian Institute of Astrophysics (IIA), Bengaluru

Dr. Mayuresh Surnis
Indian Institute of Science Education and Research, Bhopal

Committee Member	Signature	Date

ACADEMIC INTEGRITY AND COPYRIGHT DISCLAIMER

I hereby declare that this project is my own work and, to the best of my knowledge, it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at IISER Bhopal or any other educational institution, except where due acknowledgement is made in the document.

I certify that all copyrighted material incorporated into this document is in compliance with the Indian Copyright Act (1957) and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless IISER Bhopal from any and all claims that may be asserted or that may arise from any copyright violation.

April 2024 **Arihant Tiwari**
IISER Bhopal

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my project advisor, Dr. Vivek M, for his invaluable guidance, encouragement, and support throughout the course of my research. His expertise, patience, and insightful feedback have been instrumental in shaping this work.
I am also deeply indebted to my thesis supervisor, Dr. Mayuresh Surnis, for his valuable input and encouragement during the development of this thesis. His expertise in the field of astronomical data handling and research methods has been a constant source of inspiration. Along with him, I am also indebted to Dr. Suhas Gangadharaiah and Dr. Snigdha Thakur for allowing me to pursue my project IIA.

I extend my heartfelt thanks to my parents Shri. Sandeep and Smt. Rekha Tiwari, grandparents Shri. Shiv Ram and Smt. Maina Tiwari, and my family for their unwavering love and support. Their sacrifices and encouragement have been the driving force behind my academic pursuits. Their constant understanding and stress removal have been a very important aspect of my life. Special thanks to my best friend Abhisek for his continuous support, encouragement, and curiosity throughout this journey, which helped me drive my scientific endeavors. I would also like to acknowledge my friends, Gourav and Yatharth, for sticking with me through thick and thin.

Lastly, I offer my sincerest gratitude in the lotus feet of Narayan Shree Lord Venkateshwara for his divine force made me this able. Hence I extend warm back pats to myself for believing in him and dragging myself out, no matter how deep I fell.

Arihant Tiwari

ABSTRACT

We perform anomaly detection on the spectral data of 26,818 quasars lying between $1.97 < z < 2.16$ from the SDSS DR16Q Quasar catalog. Two datasets were created, with and without BAL quasars, and the results were compared. We present 5 groups of peculiar quasars and a list of visually picked 8 truly bizarre objects of unknown nature. Our algorithm also yielded a serendipitous detection of 7 FeLoBAL and a concise selection of LoBAL quasars, a subgroup of BAL quasars known for being notoriously tough to detect. We also curate a list of corrupted spectra which require SDSS redaction. An exhaustive collection of incorrect BALnicity index and z-value quasars was also created, demanding a re-evaluation of the BAL_PROB label in DR16_v4 metadata.

We used PCA decomposition of the spectra and performed dual K-Means clustering in a 30-dimensional hyperspace to cluster the entire dataset into 3 groups. Anomalies were marked with a 5σ deviation from the cluster centroid and were then grouped into 5 groups i.e. Excess SiIV Emitters, Machine Error Anomalies, CIV Peakers, BALs, and True Anomalies. A completeness check was performed using CIV, CIII, and MgII flux ratios, yielding 94% accumulation. Anomalies presented enhanced and disproportionate spectral features, translating into physical phenomena, creating an isolation that helped us understand them in greater detail.

LIST OF SYMBOLS OR ABBREVIATIONS

z	Redshift
λ	Wavelength in (Unless specified otherwise)
AGN	Active Galactic Nuclei
PCA	Principal Component Analysis
QSO	Quasi Stellar Object
$SDSS$	Sloan Digital Sky Survey
$SMBH$	Super Massive Black Hole
RMS	Root Mean Squared
SSC	Synchrotron Self Compton

LIST OF FIGURES

List of Tables

Contents

Chapter 1

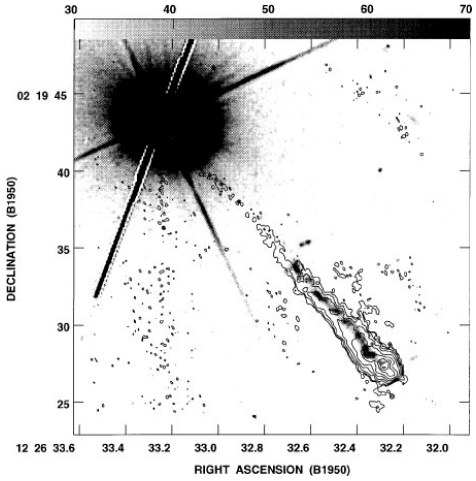
Introduction

1.1 A Brief Historical Introduction

The discovery of quasars in 1963 marked an important milestone in the field of astronomy and cosmology. The timeline began when the radars used in World War II were repurposed for astronomical observations, following the work of Karl Jansky in the 1930s. When these radio dishes and interferometers were pointed toward the sky, bright radio sources of cosmic origin began to pop up throughout the celestial sphere. These telescopes did not have a great resolution. Hence, the radio sources appeared to be blobs of contours, which could not ignite the interest of optical astronomers, as neither their shape nor the precise location could be determined.

This changed when a particularly bright radio source named 3C273 was occulted by the moon, leading to the disappearance of the radio blob it produced, hence projecting its precise location in the sky. When the optical telescopes were pointed to this location, a star-like object was discovered that apparently was emitting this massive radio emission, giving rise to the term "Quasi Stellar Radio Source" or "Quasar". The question arose: what could this strange object be?

To add to the conundrum, the spectral observation of 3C48 by Allan Sandage revealed a very unusual spectrum, with extremely strong emission lines (a property absent in the stellar spectra) and the light being variable.



(a) Quasar 3C 273

(b) Quasar 3C 48

Figure 1.1: First radio images of the earliest known quasars in the Third Cambridge Catalogue

The confusion piled up with the spectrum of 3C273, which exhibited four strong emission lines of oxygen and hydrogen, with their wavelength increased by a factor of 16%. This threw off the calculations that considered it a star in our galaxy!

> *Now, it had to be an extra-galactic object, a hundred times brighter than an entire galaxy to be seen placed at such a large redshift distance, yet compact enough to exhibit day-variability and to be confused with a star!*

With the advent of the resolution of radio telescopes, the 3C 273 turned out to be a two-component system with a 19.5 arcsec separation. As seen in Figure 1.1a, the optically identified component sat on the star-like object in the upper left corner of the image, while a jet-like object was seen as the second component of the system seen in the lower right region of the image. The presence of a jet indicated that the system was a rather violent one and not a star. During the 1960s, all these indications sparked a huge debate regarding all aspects of these objects because of their extreme luminosity, baffling redshifts (mind-boggling at that time without cosmology), extremely compact sizes, and uniform abundance.

1.1.1 The origin of redshift

The two quasars discussed above exhibited one of the highest concurrent redshifts ever observed. The astronomers in the 1960s were familiar with three kinds of redshifts:

1. **Doppler Shift** arises from the relative motion between the observer and source, with towards the observer meaning blueshift and away meaning red.

2. **Gravitational Redshift** of light that travels from a region of high gravitational field to a lower one.

3. **Cosmological redshift** arose from the universe's expansion due to dark energy.

All three of these explanations were tried by astronomers in the early days before the consensus settled with the latter most. The Doppler shift was not feasible as it would require a source pumping out these sources at a massive velocity outwards from the galaxy, which did not make statistical sense, as there was no reason for such bias in direction. Similarly, gravitational redshift required the invocation of complex astronomical systems, such as the presence of massive neutron star ensembles in the hearts of super-giant stars, to produce a gravitational field strong enough to account for the observed redshift.

The third alternative, the *Cosmological Hypothesis*, had always been the central theory but was not accepted fully until the discovery of CMB (Cosmological Microwave Background) in 1965. All galaxies, along with these peculiar objects, are placed in an expanding universe, and the redshift is explained by the time-dilation produced in the curved space-time of this universe. This was later named as the **Hubble Law**.

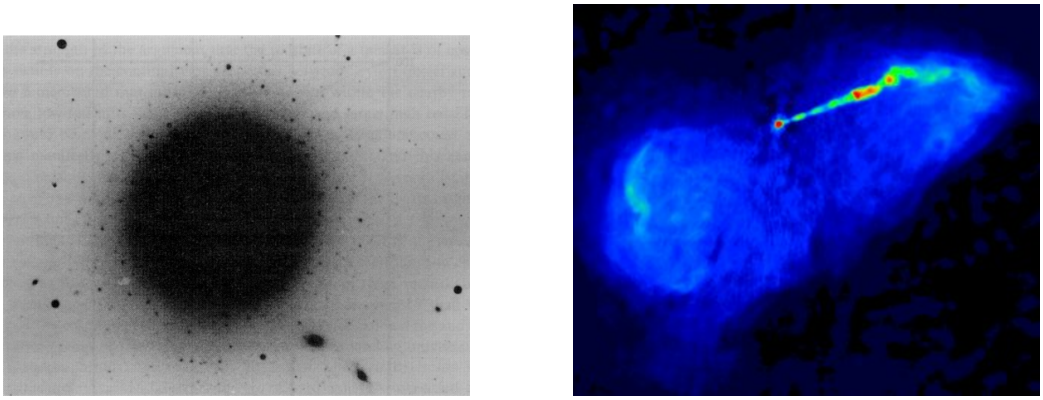
The cosmological hypothesis gained traction because it did not require complex physical scenarios to explain the great redshifts. The quasar properties were found to be similar to Seyfert galaxies and AGNs in general. Since the redshift of Seyferts was known to be cosmological, it was easily extended to that of quasars.

1.1.2 Physical Characteristics

Once the boiling debate about the origin of redshift for these objects settled down on the cosmological hypothesis, the astrophysics community faced a new challenge: *What the heck is happening over there?*
The observations held during this initial phase of quasar research helped the scientists identify the following general properties for them.

- It must be a star-like object identified with a radio source
- The light should be variable
- The ultraviolet flux should be large
- Broad emission lines in the spectra with absorption lines in some cases
- Large redshifts

Later on, it was seen that the radio emission property was not general and was rare, and it was seen only in 10 percent of the quasar population. However, this connection was obvious since most of the quasars discovered during that time were initially identified as strong radio sources.
In order to better understand these objects, observations will be made throughout the electromagnetic spectrum, from radio waves to gamma rays, using advanced telescopes in the coming decades. Strangely enough, by each of these methods, the quasars seemed to exhibit vastly different interconnected and disconnected characteristics, which also varied over time. This indicated a much more diverse phenomenon at play: An ensemble of high (due to the presence of Γ and X-rays) and low energy processes (Radio and Microwaves), which originated from different regions of the quasar, as seen in the images. These processes also affected each other as variability in one showed a corresponding response in some part of the spectrum.
For example Figure 1.2a shows the observation for the elliptical galaxy M87 (87^{th} object of the Messier Catalogue) in the optical regime. It appears as a dark circular blob with density falling with radius, which is the typical characteristic of an elliptical galaxy. But when the same galaxy was observed in



(a) Optical Image (b) Radio Image

Figure 1.2: Observations for M87 (Elliptical Galaxy with an AGN) in different wavelength domains

Radio frequency, it looked, unlike the optical image. As shown in Figure 1.2b, the object instead exhibited a jet, spewing warm hydrogen gas into a lobe that is being pushed back, diffusing into the galaxy, after a certain distance. When studied, a sharp rise in luminosity towards the center was observed. At the same time, the velocity of stars in the central region was measured to rise rapidly, indicating the presence of a compact and massive object in the center. The conclusion of these studies was that this central attractor was probably an SMBH of $\sim 5 \times 10^9\,\mathrm{M}_\odot$. The galaxy was also a strong X-ray source, requiring an energy production mechanism that could surpass the typical stellar emissions. Hence, with many such objects being discovered, the study of quasars and "Active Galactic Nuclei" started to emerge as one of the most fascinating fields in astronomy, studied extensively throughout the globe to this day.

Since the only information we receive from these objects is their light, the spectrum of quasars becomes the most important aspect in determining the exact taxonomy, characteristics, physical/emission processes, and chemical properties of quasars. The details about how each of these properties is derived from the spectrum itself and the motivation for using the quasar spectra to find anomalous objects will be discussed and made evident in the later sections.

1.2 Taxonomy of Active Galaxies

In the diverse zoo of galaxies of varying size, color, composition, shape, origin, and evolution, there is a special state in which a galaxy can exist

Definition 1.1. Active Galaxy refers to a galaxy that houses an energetic phenomenon, which cannot be attributed to stellar activity, in their central region or nucleus. The nuclei of such galaxies are collectively called **Active Galactic Nuclei** or **AGNs**.

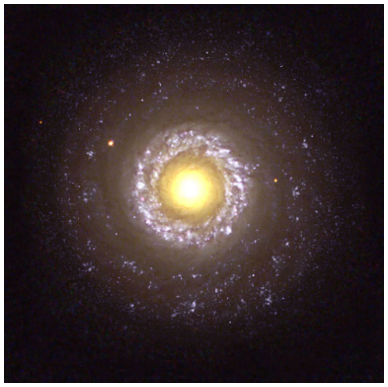
Semantically, the AGNs are broadly classified into two types

- **Seyfert Galaxies**: In this case, the total energy emitted by the nucleus in the visible regime is comparable to combined energy emitted by all the stars present in the galaxy, i.e., $\sim 10^{11} \odot$
- **Quasar**: Here, the energy output by the AGN is brighter than the entire galaxy by a factor of 100 or more.

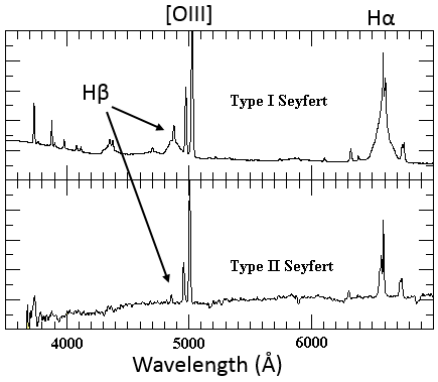
Since the probability of finding high-energy sources like quasars is rare, one is *statistically likely* to find them only at great distances. At such a large separation, because of its small angular size and feeble brightness as compared to the AGN, the light of the circumnuclear galaxy fades off, leaving only the star-like nucleus, giving it a "quasi-stellar" appearance.
A detailed taxonomy of active galaxies can be created considering various emission factors, such as Radio Loud Quasars, which exhibit extremely high radio emissions and are generally associated with the presence of jets, and Blazars, which are quasars with their jets pointed directly toward us (as an observer) giving rise to relativistic effects in their spectral profile.
There are also a few special types of quasars, which will be discussed in later chapters, as they become relevant to our project and start emerging in the SDSS data that we are using, such as BAL (Broad Absorption Line) Quasars, Heavily Reddened Quasars, etc.
Below, we briefly discuss the general taxonomy of active galaxies and their properties as they form the object set in the sample space for our anomaly detection project.

1.2.1 Seyfert Galaxy

Seyfert galaxies are a class of active galaxies characterized by their strong and broad emission lines in their spectra, indicative of highly ionized gas in their nuclei. They are named after the American astronomer Carl Seyfert, who first identified them in the 1940s. These galaxies typically have a compact core, which is believed to host a supermassive black hole accreting material from the surrounding disk. Seyfert galaxies are further classified into two main types: Type 1, which displays broad emission lines originating from the broad-line region (BLR), and Type 2, where only narrow emission lines are visible due to the obscuration of the BLR by dense gas and dust, characterized by strong infrared emission. Studies of Seyfert galaxies have contributed significantly to our understanding of the processes driving active galactic nuclei (AGN) and the role of supermassive black holes in galaxy evolution[23]. They frequently exhibit complex kinematics in their emission-line regions,



(a) Seyfert Galaxy (NGC 7742)



(b) Typical spectra of Seyfert galaxies

Figure 1.3: Optical image and spectral property of a typical Seyfert Galaxy

suggesting the presence of outflows and inflows of ionized gas driven by the central engine. Recent observations and simulations have highlighted the role of interactions and mergers in triggering AGN activity in Seyfert galaxies, contributing to our understanding of galaxy evolution processes [9]. Figure 1.3a shows the optical image of a Seyfert Galaxy, NGC 7742, and the typical spectra of Seyfert Type I and II are shown in figure 1.3b.

1.2.2 Quasars

Quasars, short for "quasi-stellar radio sources," are the intensely luminous centers of distant galaxies powered by supermassive black holes accreting matter at high rates. These celestial objects were first identified in the 1960s through their strong radio emissions, but they are also known for their high-energy radiation across the electromagnetic spectrum. Quasars exhibit extreme redshifts, indicating their immense distances from Earth, with some of the most distant quasars observed existing when the universe was only a fraction of its current age. They are crucial probes of cosmic evolution, shedding light on the early stages of galaxy formation and the growth of supermassive black holes. Quasar research has significantly contributed to our understanding of active galactic nuclei (AGN) and the role of black holes in shaping the universe's structure [21].

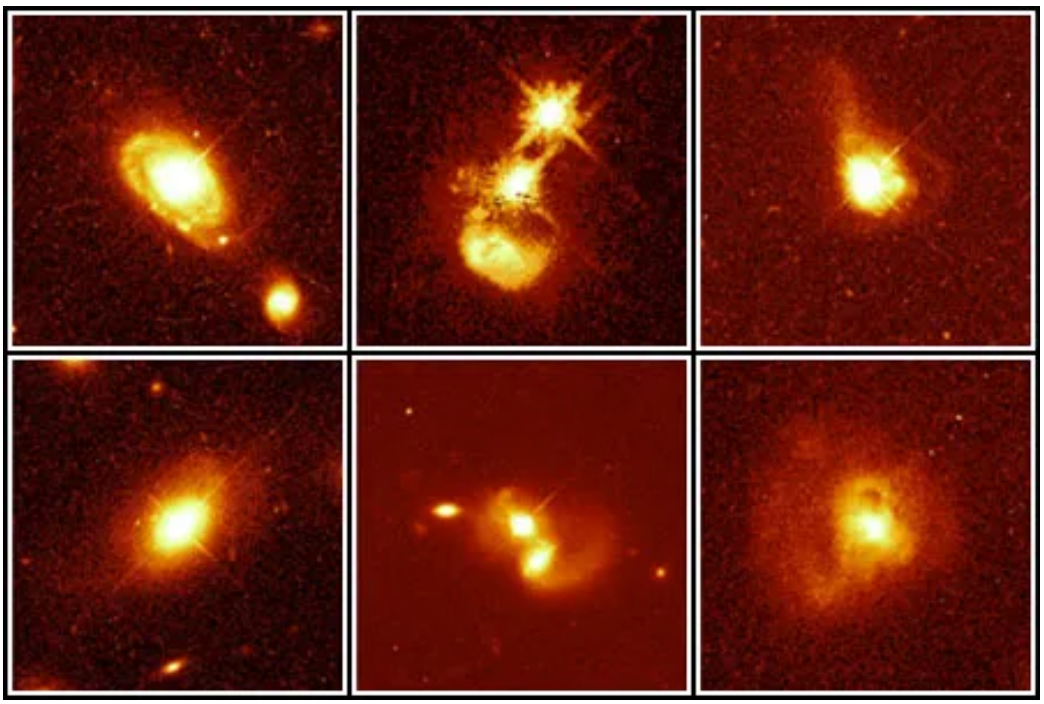


Figure 1.4: Quasar Host Galaxies, HST Nov 19, 1996. J Bahcall, Institute of Adv Studies, NASA

Quasars are the main objects we are interested in studying for this project and we will be looking at their UV-Optical emissions via SDSS Data.

1.2.3 Radio Galaxies

Radio galaxies are a class of active galaxies distinguished by their strong and extended radio emission, often originating from powerful jets expelled from their central regions. These jets, composed of relativistic particles, can extend over vast distances, interacting with the intergalactic medium and impacting their surrounding environments. Radio galaxies are known to exhibit a wide range of morphologies, including double-lobed, edge-darkened structures, and compact core-dominated sources. Studies of radio galaxies have revealed a connection between their radio emission and the presence of supermassive black holes at their centers, suggesting that accretion onto these black holes powers the jets. Furthermore, observations have shown that radio galaxies can influence the evolution of their host galaxies and the surrounding intergalactic medium through feedback processes driven by their energetic outflows. The study of radio galaxies provides valuable insights into the physics of active galactic nuclei and their role in the broader context of galaxy formation and evolution [6]. Figure 1.5 shows the radio image of

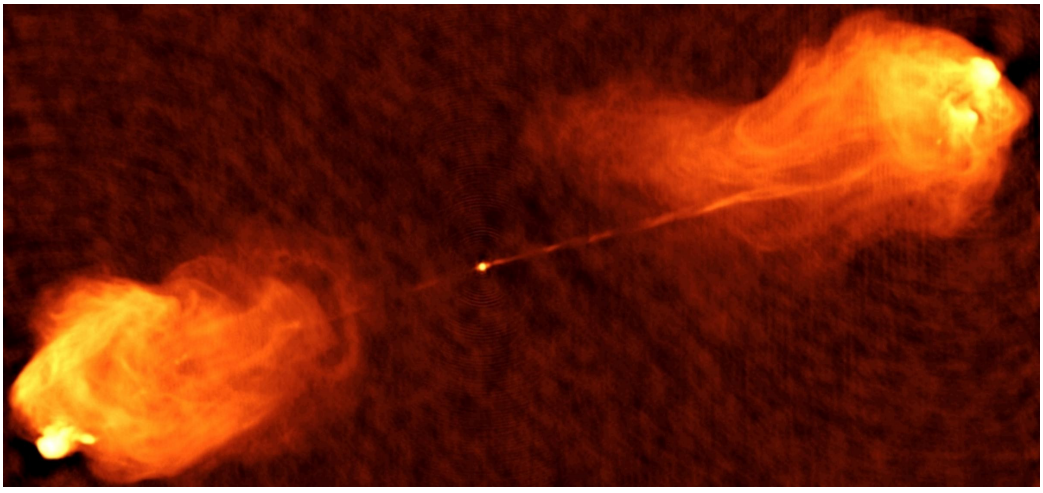


Figure 1.5: Radio Galaxy Cygnus A (3C 405)

the galaxy Cygnus A, the first radio galaxy to be discovered and identified optically. At a distance of 232 MPc (Mega Parsecs), it is one of the brightest radio sources in the sky. The central bright dot is the galaxy shooting two jets on both sides, which diffuse into radio lobes surpassing the galactic scale.

There are a few minor but rare classes of AGNs, such as

- **LINERs** (Low-Ionization Nuclear Emission-line Regions) represent a diverse class of galaxies with weak emission lines originating from their central regions. These galaxies often exhibit low-ionization states of gas, suggesting various underlying mechanisms for their activity. LINERs are found in both spiral and elliptical galaxies, hinting at a wide range of possible origins, including processes such as star formation, shocks, and the presence of low-luminosity active galactic nuclei (AGN)[12].

- **BL Lac Objects** are a subset of active galactic nuclei (AGN) known for their featureless optical spectra and high degree of polarization. They are characterized by their rapid and unpredictable variability across the electromagnetic spectrum. BL Lac Objects are thought to be powered by relativistic jets emanating from supermassive black holes at their centers, with these jets oriented toward Earth, resulting in intense emission from the core region. These objects provide valuable insights into the physics of AGN and the nature of compact, energetic phenomena in the universe [18].

- **Blazars** are a subclass of AGN characterized by their extreme variability and intense emission across all wavelengths, from radio to gamma rays. They are believed to be powered by relativistic jets pointed directly towards Earth, resulting in extreme Doppler boosting effects and enhancing their observed luminosity. Blazars are divided into two main sub-classes: BL Lac Objects, which exhibit weak or no emission lines, and Flat-Spectrum Radio Quasars (FSRQs), which display prominent emission lines in their spectra. The study of blazars provides crucial insights into the physics of AGN jets, particle acceleration mechanisms, and the interplay between black holes and their host galaxies[28].

With a general idea of the types of active galaxies and quasars that can be found in the cosmos, we move on to understand and discuss the various radiative processes and their physics that constitute the spectrum of a quasar.

1.3 Radiative Processes

The radiation from a quasar covers the entire electromagnetic spectrum, from radio-emitting jets to high energy gamma gays by annihilation. The major part of this radiation is very different from the simple black body radiation of stellar sources, giving rise to the name *Non-Stellar emissions* or *Non-Thermal Emissions*. A few of the most important and prominent radiative processes are discussed below, which will become relevant when we try to explain the anomalous spectra with their help.

1.3.1 Basic Radiative Transfer

Radiative transfer deals with the interaction of radiation with matter and the subsequent phenomena. In order to understand it, we need to describe three quantities, i.e., Specific Intensity (I_ν), Monochromatic absorption cross-section, ($\kappa_\nu(cm^{-1})$) and Volume emission coefficient (j_ν), described as:

$$I = \frac{F}{\sigma A \nu s} \qquad\qquad j_\nu = \frac{F}{\sigma \nu V s}$$

Where, F is locally emitted flux, σ is the solid angle, A/V is Area/Volume, and s is time in seconds.

These three combined give us the equation of radiative transfer,

$$\frac{dI_\nu}{ds} = -\kappa_\nu I_\nu + j_\nu \tag{1.1}$$

The first term on the right describes the radiation loss due to absorption, while the second shows the radiation gain due to the local emission process. When Equation 1.1 is divided by κ_ν on both sides, we obtain,

$$\frac{dI_\nu}{d\tau_\nu} = -I_\nu + S_\nu \tag{1.2}$$

Where ($d\tau_\nu = \kappa_\nu ds$) is the optical depth and $S_\nu = j_\nu/k_\nu$ is the *source function.* Considering the AGN vicinity the radiating matter behaves like an opaque source in full thermodynamic equilibrium (TE), and the optical depth

is large, hence both I_ν and S_ν approach the Planck Function,

$$\boxed{B_\nu(T) = \frac{2h\nu^3/c^3}{exp(h\nu/kT) - 1}} \tag{1.3}$$

1.3.2 Synchrotron Radiation

Synchrotron radiation is emitted when a relativistic electron is accelerated by a magnetic field; hence it also goes by the name *Magneto Bremsstrahlung*. It is believed to majorly constitute the radio emission from quasars, sometimes modified by intervening processes such as absorption and re-emission. Synchrotron polarization and absorption is also an important aspect of quasar radiation.

Considering an electron with energy E being accelerated in a magnetic field B of energy density $u = B^2/8\pi$. The power emitted by this electron (P), i.e., $-dE/dt$, is given by

$$P = 2\sigma_T c\gamma^2\beta^2 u_B \sin^2\alpha \tag{1.4}$$

Where σ_T is the Thomson Cross section, γ is the Lorentz Factor, $\beta = v/c$ and the angular terms represent the direction of motion with α as the pitch angle.

When Equation 1.4 is averaged over isotropic pitch angles, it gives

$$\bar{P} = \frac{4}{3}\sigma_T c\gamma^2\beta^2 u_B \tag{1.5}$$

The radiation emitted by such a single electron is beamed in the direction of motion. The final emission is calculated by considering an ensemble of electrons with an energy distribution, which emits radiation at a characteristic peak frequency $\gamma^2\nu_L$ where ν_L is the Larmor Frequency. Hence, the luminosity is calculating the integral over the optically thing source, which is given by

$$L_\nu = \frac{1}{4\pi}\int_V \int_1^\infty [\bar{P}(\gamma)n(\gamma)d\gamma]\,\nu^\alpha dV \tag{1.6}$$

Where $n(\gamma)d\gamma$ gives the energy distribution. This formulation brilliantly explains the slopes of many AGNs at radio, UV-optical, and X-ray energies.

Synchrotron Self-Absorption

This situation arises when the source of these electrons is optically thick to its own radiation. The opacity is maximum at the lowest frequency, which results in significant modification of the emergent spectrum. Here, with p being the power index, it is seen

$$\kappa_\nu \propto \nu^{-\frac{p+4}{2}} \tag{1.7}$$

that the largest absorption is at the lowest frequencies. Using Equation 1.1 for a homogeneous medium, we get $I_\nu \propto \nu^5/2$ for large optical depths.

Polarization

Synchrotron radiation exhibits high linear polarization, with an intrinsic level of up to about $\sim 70\%$. However, the observed radiation has a much smaller polarization extent $(3 - 15\%)$, which indicates the presence of a strong non-polarized source mixed with this highly polarized source. This percentage drops at lower wavelengths, which is contrary to what is expected from a pure synchrotron source. This is explained by the presence of an additional thermal non-polarized source. Synchrotron radiation with both these secondary processes, is thought to make up most of the AGN's non-thermal emission.

1.3.3 Compton Scattering

Compton scattering refers to the interaction between an electron and a beam of radiation. The energy and momentum conservation relations are used to calculate the relationship between the frequencies of incoming and outgoing photos, for slow or stationary electrons. If $\vec{n_\nu}$ and $\vec{n_{\nu'}}$ are unit vectors in the

direction of incoming and outgoing photons,

$$\nu = \frac{m_e c^2 \nu'}{m_e c^2 + h\nu'(1 - \cos\theta)} \tag{1.8}$$

For non-relativistic electrons, the cross section for this is

$$\frac{d\sigma}{d\Omega} = \frac{1}{2} r_e^2 [1 + \cos^2\theta] \tag{1.9}$$

Where $r_e = e^2/m_e c^2$ is the classical electron radius. Integrating over angles gives the Thomson cross section, σ_T.

Comptonization

It refers to the process of photons and electrons reaching an equilibrium. The fraction of energy lost by the photon after each scattering is,

$$\frac{\Delta\nu}{\nu} = -\frac{h\nu}{m_e c^2} = -\epsilon \tag{1.10}$$

Here, the loss of energy, hence a cooling of the gas is the direct result of *Inverse Compton Scattering.*

Definition 1.2. Inverse Compton Scattering refers to the phenomenon which involves the scattering of a photon by an electron, where the outgoing photon as more energy than the incoming one.

Considering the fraction x of the electron energy kT being transferred to the photon, the cooling term for the electron gas with temperature T_e can be written as

$$C_{CS} = \int \frac{L_\nu}{4\pi r^2} N_e \sigma_T \left[\frac{x k T_e}{m_e c^2}\right] d\nu \tag{1.11}$$

If we consider the Compton Heating and cooling to be the only source-sink process, and as previously discussed, if the radiation field is a Planck function the equilibrium requirement would give $x = 4$ in the above equation. In AGNs, the energy density of photons exceeds that of the electron, giving rise

to the Compton equilibrium temperature of

$$T_C = \frac{h\bar{\nu}}{4k} \tag{1.12}$$

Where the mean frequency is given by integrating over the SED of the source. For a typical quasar, such calculations reveal a temperature of $T = 10^8$ K. Hence, Compton and Inverse Compton scattering, along with some other processes, form the basis of the primary continuum emission source from the denser parts of the AGN, which we will see inscribed in the quasar spectra as discussed in Chapter 4.

Synchrotron Self-Compton

Definition 1.3. Synchrotron Self-Compton (SSC) refers to the phenomenon where, in a compact synchrotron source, the emitted photons get inverse Compton scattered by the relativistic electrons that give rise to the synchrotron radiation, which results in a big boost of energy for the photons.

The emergent flux from this process is calculated by integrating the electron velocity distribution and the synchrotron radiation spectrum. The medium is usually so dense that this SSC process repeats several times before an emergent photon is released. The natural limit of this process occurs when the scattered photon energy becomes γ-radiation and the condition $h\nu_\gamma \ll m_e c^2$ no longer holds, which results in a dramatic radiation density decrement. [4]

1.3.4 Thomson Scattering

The classical scattering of photons with energy $h\nu \ll mc^2$, incident on an electron is called Thomson Scattering. The photons render the electron oscillating, radiating as per the Larmor formula. Here, there is no change in frequency of the incident photon i.e., the scattering is elastic, which is the main difference between Thomson and Compton scattering. For a linearly polarized incident radiation, the differential cross section for scattering is given by

$$\left(\frac{\sigma_T}{\Omega_T}\right)_{pol} = \left(\frac{e^2}{mc^2}\right)^2 sin^2\Theta \tag{1.13}$$

Whereas for an un-polarized source, the cross-section is given by Equation 1.9. Summed over all the angles, the cross-section comes out to be about 6.65×10^{-25} cm^2. When un-polarized radiation is Thomson scattered, it becomes partially polarized, and the degree is given by

$$\Pi = \frac{1 - \cos^2\theta}{1 + \cos^2\theta} \tag{1.14}$$

It accounts for a few low-energy exchanges taking place in the outermost regions of the AGN where the photon & electron energy exchange is elastic.

1.3.5 Annihilation and Pair Production

Many AGNs are known to exhibit γ-ray jets, which indicates that in certain parts of the central engine, the conditions are just right such that the density of high-energy photons becomes large enough for pair production. This results in the formation of concentrated electrons and positrons pockets in those parts.

Pair production is a fundamental process in particle physics where a high-energy photon converts into a particle-antiparticle pair, typically an electron and a positron, in the vicinity of a heavy nucleus or another energetic photon. This phenomenon requires the presence of a target particle, usually a nucleus, to conserve both energy and momentum. The produced particles carry away the excess energy in kinetic form, leading to the creation of an electron-positron pair.

$$e^+ + e^- \rightleftharpoons \gamma + \gamma$$

Pair annihilation, on the other hand, is the reverse process where a particle (electron) and its antiparticle (positron) collide and annihilate each other, resulting in the production of two photons with energies equivalent to the rest mass energies of the initial particles. Pair production and annihilation are

crucial concepts in understanding particle interactions and play significant roles in various astrophysical phenomena, such as gamma-ray emission from compact objects and the formation of electron-positron plasma in high-energy environments [13, 30]. These processes are likely to occur in the corona of the accretion disk or inside the gamma-ray jets.

Considering the interaction between a gamma ray photon with frequency ν_γ with an X-ray photon of frequency ν_X, the threshold frequency for pair production is given by

$$\nu_\gamma = \left(\frac{m_e c^2}{h}\right)^2 \frac{2}{\nu_X(1 - \vec{n}_\gamma.\vec{n}_X)} \tag{1.15}$$

The size of the optical source plays an important role in determining its optical depth and hence the probability of pair production. The dependence is called the *compactness parameter*.

1.3.6 Bremsstrahlung Radiation

Free-free emission, also known as bremsstrahlung or free-bound continuum emission, is a process in which a free electron is accelerated or decelerated by the Coulomb force of a positively charged ion, resulting in the emission of electromagnetic radiation. This radiation covers a broad spectrum, from radio to X-rays, depending on the energy of the electrons involved. Free-free emission is prevalent in hot and ionized astrophysical environments, such as HII regions, stellar atmospheres, and accretion disks around compact objects. It contributes significantly to the observed continuum emission and can be used to infer physical properties such as temperature, density, and ionization state in these environments [16].

Free-bound emission occurs when a free electron is captured by an ion, transitioning to a bound state within the ion's atomic structure. During this transition, the electron releases energy in the form of a photon. The emitted photon's energy corresponds to the difference in energy levels between the initial and final electron states. Free-bound transitions are responsible for

producing discrete spectral lines observed in various astrophysical environments, such as nebulae, stellar atmospheres, and interstellar medium. These emission lines provide valuable information about the chemical composition, temperature, and density of the emitting regions [17].

The spectral shape for these emissions differs significantly from a black body radiation. The Free-Free emissivity due to ion i of an element Z with a number density N_i is given by

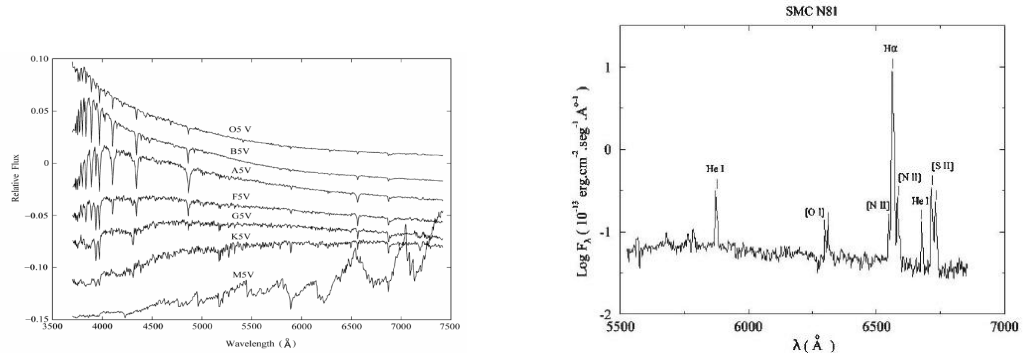
$$4\pi j_\nu = 6.8 \times 10^{-38} Z^2 T_e^{-1/2} N_e N_i \bar{g}_{ff}(\nu, T_e, Z) e^{-h\nu/kT_e} \tag{1.16}$$

where $\bar{g}_{ff}$ is the energy averaged Gaunt Factor.
The Bremsstrahlung radiation extends over a large range of energies and looks like a very flat powerlaw. The total energy per unit volume per second (which is also the cooling rate) can be found by integrating Equation 1.16 over all the frequencies.
This concludes the introduction of the various emissions present in a quasar. An anomaly would be an unusual amalgamation of many of these processes.

1.4 The Spectra of Quasars

1.4.1 What forms the spectra?

A quasar or galaxy in general has four major constituents: gas, dust, stars and dark matter. Although dark matter is the main component of a galaxy, it does not interact with normal matter or photons. Since we are talking about spectra, we will be ignoring the dark matter component, leaving us three bulk sources of radiation. As seen in Figure 1.6a, a typical stellar spectrum contains absorption lines cut into a thermal continuum. Using these absorption lines one can learn many things about the star, such as the chemical composition, surface temperature, and luminosity are encoded in the strength and width of these absorption lines. Doppler shifts in these lines can help measure the radial velocity, and a periodicity in the Doppler shift



(a) Stellar Spectra for various stars (b) Typical HII Region Spectrum

Figure 1.6: Major components of a galactic spectrum

is indicative of a binary star system.

The primary emissions and major contribution to the galactic spectrum by gas in a galaxy are observed from hot clouds known as **HII regions.** They are usually seen around active star-forming zones and hence are prominently present in spiral and irregular galaxies. Their spectrum consists of few emission lines as seen in Figure 1.6b.

The dust in a galaxy performs the cooling task and hence does not have any optical emission signature. Its main effect is absorbing the optical and high energy radiation (hence cooling) and emitting them in lower energy wavelengths (Infrared and beyond).

In the case of active galaxies, the spectrum extends to a wider range of wavelengths hence becoming a *broadband spectrum.* It has features in addition to the combination of the three features mentioned above that make a typical galactic spectrum. We will shortly see, how the features in a broadband spectrum reveal phenomenal physics of extreme nature.

1.4.2 Optical Spectra

The optical spectrum of a galaxy is obtained by adding up the spectra of its individual components i.e. the stars, gas, and dust. When adding up spectra, the following common scenarios show up,

- Different types (by age, composition, mass, etc) of stars have different absorption lines. When added, these lines get diluted as one line might not be present in the other spectra and vice versa.
- All the lines share the galaxy's redshift, but additional Doppler shifts can also arise because of the individual motion of objects within the galaxy, resulting in the broadening of the lines making them wider and shallower.
- When the spectrum of HII regions is added to stellar spectra, their steep emission lines tend to be the prominent feature, unless it coincides with an absorption line and fades out. These emission lines also face Doppler broadening to different extents.

Doppler Broadening

Light from a quasar is emitted by individual atoms in motion, causing red and blue shifts of the emitted radiation which broadens the overall received spectrum. The motion of these atoms is a consequence of their thermal energy, hence hotter gas tends to show higher broadening effects than colder gas. The velocity dispersion of a gas with atoms of mass m, at a temperature T is given by

$$\Delta v = \sqrt{\left(\frac{2kT}{m}\right)} \tag{1.17}$$

The broadening caused by this velocity dispersion is given as

$$\frac{\Delta\lambda}{\lambda} \approx \frac{\Delta v}{c} \tag{1.18}$$

and it has become a standard practice to express the broadening in terms of velocity. Thermal motion is not the only source of Doppler broadening. The

bulk motion of particles can also cause velocity dispersion. The bulk motion must impart different velocities to different atoms along the line of sight to cause a broadening effect. These motions are visualized in Figure 1.7 The

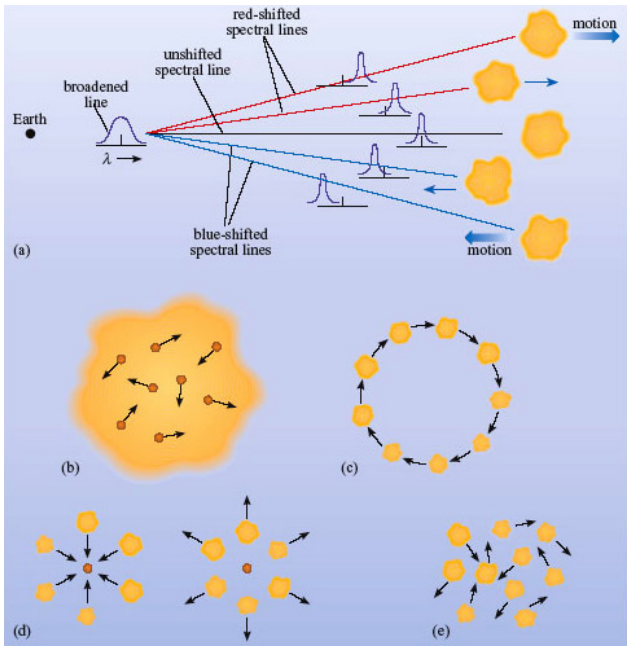


Figure 1.7: Doppler Broadening is caused by particles having different velocities along the line of sight (a). This can be due to (b) Thermal motion, (c) galactic rotation, (d) gas inflow or outflow (e) chaotic gas motion.

thermal broadening depends on the mass of individual atoms that emit the radiation as described in Equation 1.17; Hence lines of different elements will have different extents of broadening. Whereas, the bulk motion of particles will affect every line equally. This fact is used to differentiate between thermal and bulk broadening.

Figure 1.8 shows the spectrum for *elliptical galaxies*(First row) which primarily contain absorption lines due to lack of HII regions and constitute mostly of old stars, and *spiral galaxies* (bottom row) which show a mixture of both emission and absorption because of star formation, dust and HII regions around the stars. Notice, the narrow sharply peaked emission lines in the spiral galaxies. The Doppler broadening due to galactic rotation and other processes described above has caused them to become sharp narrow

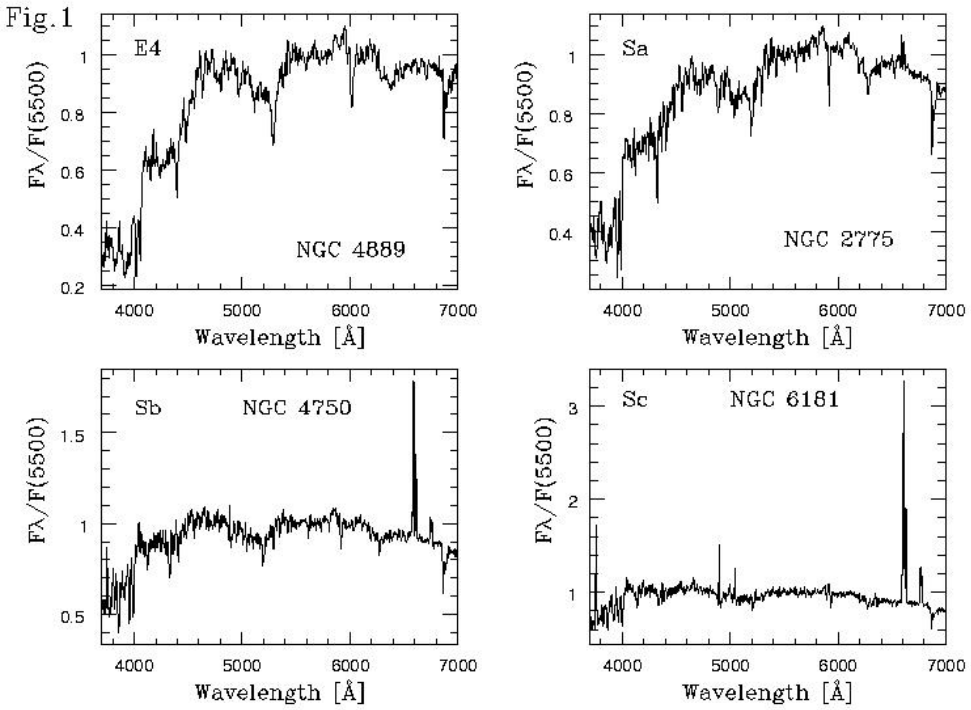


Figure 1.8: Typical spectra of galaxies

peaks with $\Delta v \sim 100 - 300$ Km/s instead of a single line (resembling a Dirac Delta Function).

Active Galactic Spectra

Looking at the spectrum of a typical quasar in Figure 1.9, it immediately becomes apparent that the emission lines here are much stronger and broader than that of normal or even starburst galaxies. The strong lines point towards excessive amounts of hot gas which also needs to be extremely hot or in rapid motion to account for the broadening. When we try to account for this broadening thermally, the temperature needed sores up to 10^9 K!, which is higher than the core temperatures of most massive stars and is enough to rip atoms apart, hence unfeasible. This suggests that the hot gas is in rapid motion with a velocity dispersion of several thousand Km/s. To calculate

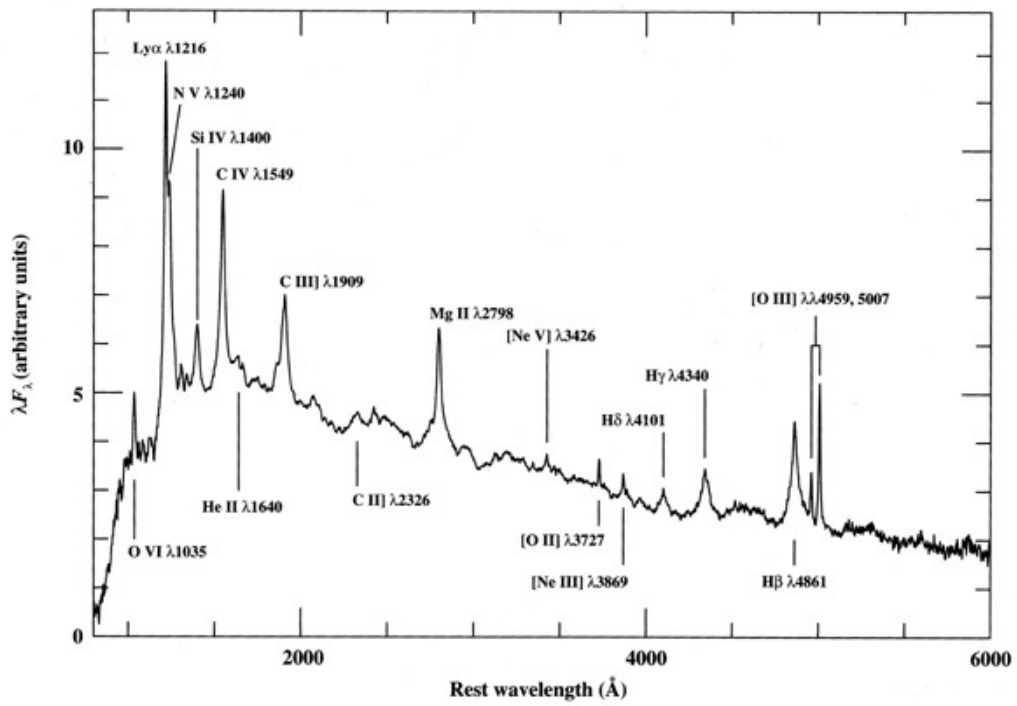


Figure 1.9: A mean QSO spectrum formed by averaging spectra of over 700 QSOs from the Large Bright Quasar Survey [8]. Prominent emission lines are indicated. Data courtesy of P. J. Francis and C. B. Foltz.

the actual temperature of the emitting gas, we use the relative strengths of all the emission lines, which turns out to be about 10^4 K.

It is also noted that the active galaxies exhibit several times the total energy output of a normal galaxy. The exact physical process behind it will be discussed in Section 1.5. Two important features of the AGN spectra are the **Big Blue Bump** which refers to a prominent feature in the UV to soft X-ray portion of the AGN spectrum, characterized by a broad, smooth continuum emission. This emission is believed to originate from thermal radiation emitted by the accretion disk surrounding the supermassive black hole at the center of the AGN. The term ”blue” is used because this component of the spectrum peaks at relatively short wavelengths, typically in the UV region. It provides crucial insights into the physical properties of the accretion disk, such as its temperature profile and luminosity, and plays a significant role in

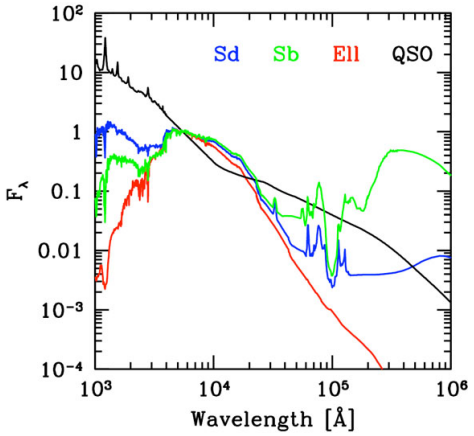


Figure 1.10: Broadband SEDs for galaxies and QSO

When we compare the broadband Spectra Energy Distributions (SED) of active and normal galaxies, we find that the active galaxies have a much flatter SED. As seen in the figure to the left, there is strong emission in Gamma, X-rays and UV light for the QSOs indicating that there is relatively much more emission at high energy wavelengths stringed to a high energy phenomenon taking place.

understanding the accretion processes powering AGN. [25]

In contrast, the **Small Blue Bump** refers to a secondary emission component observed in the optical to UV portion of the AGN spectrum, typically peaking at shorter wavelengths than the Big Blue Bump. The origin of the Small Blue Bump is less well-understood compared to its larger counterpart. It is thought to arise from a combination of contributions, including emission from the outer regions of the accretion disk, as well as other processes such as reprocessing of radiation by optically thick material or contributions from the host galaxy. It plays a significant role in constraining models of AGN energetics and the geometry of the accretion flow. [31]

1.5 Unified Model: Structure of Quasar

In the previous sections, we discussed the different types of radiation processes that we were able to decipher after studying the spectrum. We were also able to come up with size, temperature, density, and composition by studying and comparing various properties present in the quasar spectra. Now, we finally are in the position to discuss the ”possible” and most accepted structure that a quasar must have in order to account for all the observed characteristics.

The size of AGNs

AGNs appear as point sources in optical images of even the most advanced telescopes such as the Hubble Space Telescope. Using their resolution parameters, we have calculated an upper constraint on the size of a typical AGN to be about $\sim 1pc$. One parsec is an extremely small distance on a galactic scale given our galaxy is 30,000 parsecs in diameter, making AGNs much smaller than the galaxies.
A second size constraint is placed using variability. The spectral variation time scales at X-ray frequency for AGNs can be from a few hours to over a year. In order to observe a brightness variability with time scale Δt, the size (R) of the source must be no larger than $R = c\Delta t$. If the source is bigger than this, the luminosity variations will be smeared out and the observer will see a continuum instead. When we input the day variability of a few quasars ($\Delta t \sim 10^4 s$), we obtain a size of about 0.0001 pc. This result is staggeringly small as compared to the optical image constraint discussed above, which means the X-rays are being emitted from a very smaller region of the AGN. Using **Very Long Baseline Interferometry**, radio astronomers also place a size constraint about 100 times smaller than the optical image counterpart.

The Luminosity of AGNs

It is observed that a typical quasar emits thrice as much as optical at UV and IR frequencies, making it at least four times as bright as a typical galaxy. Most of the quasar's emissions are attributed to the AGN. Calculations suggest that in general AGNs have a luminosity of more than $2 \times 10^{10} L_\odot$ or $8 \times 10^{36} W$ which is produced in a minuscule volume.

1.5.1 The Supermassive Black Hole

The gravitational field of am SMBH acts as the key component that powers the central engine of a quasar. The radius of the **event horizon** (the so-called surface of a black hole, beyond which escape velocity exceeds c) for a

black hole that powers a quasar is given by

$$R_S = \frac{2GM}{c^2} \tag{1.19}$$

Where R_S is the Schwarzschild Radius and M is the black hole mass. When R_S is calculated from the variability constraints discussed above, we get a mass equivalent to $\sim 10^9 M_\odot$. This calculation ensures that it is in fact possible to fit an SMBH inside the AGN, but does not mandate it to be this massive.

1.5.2 The Accretion Disk

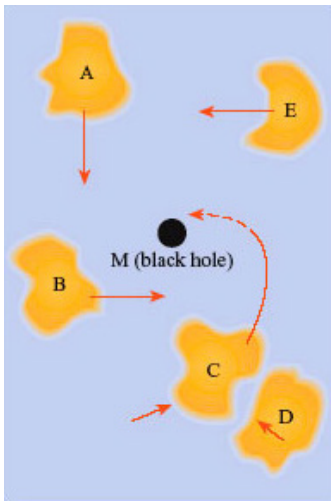


Figure 1.11: Gas cloud infall around the black hole

When a number of gas clouds come in the vicinity of the SMBH, they start to accelerate towards it under gravity and begin orbiting it. At the closest approach, they collide with each other, losing kinetic energy, hence retreating to a lesser distance with each collision. Subsequent collisions make their orbits circular and also heat up the gas. Since these clouds are in a Keplerian orbit, the inner clouds orbit faster than the outer ones. A form of friction (viscosity) starts to act between neighboring clouds accelerating their energy loss. As a result, the innermost regions of the clouds fall to even smaller orbits an **accretion disk** is created.

The amount of power that can be produced by this accretion disk has an upper limit called the **Eddington Luminosity**. As the accretion becomes faster, so does the radiation pressure of the energy being emitted by the disk. This outward-pushing radiation pressure balances the gravitational force pulling the matter inside hence placing a limit to the rate of accretion.

The inward spiralling ends abruptly at $\sim 5R_S$ from the black hole's center. Hereon, the infalling material begins to fall rapidly and passes through the event horizon, into the black hole. The accretion disk is situated outside this event horizon and radiates a vast amount of energy which is believed to power the quasar. Based on calculations, if a mass m falls into the black hole, it can radiate energy $\approx 0.1\ mc^2$ before disappearing. This makes accretion the most effective mass to energy conversion mechanism after matter-antimatter annihilation.

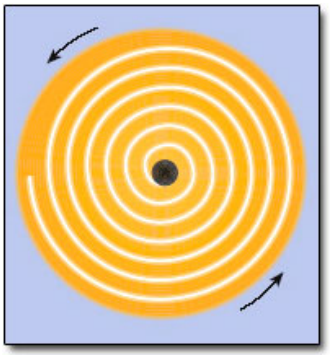


Figure 1.12: Formation of accretion disk spiral

Mathematically,

$$L_E = \frac{4\pi G M m_p c}{\sigma_T} \cong 1.3 \times 10^{31} \frac{M}{M_\odot} W \tag{1.20}$$

Where m_p is the mass of the proton, σ_T is the Thomson cross section, M is the mass of SMBH and $M_\odot$ is the solar mass. Hence to account for the observed luminosity, a supermassive black hole of the order of billion solar masses is needed to sit in the center of the AGN.

1.5.3 The Jets

Jets are narrow, highly collimated streams of particles and electromagnetic radiation emitted from the vicinity of compact astronomical objects, such as black holes, neutron stars, and young stellar objects. These structures can extend over vast distances, sometimes spanning thousands of light-years. Jets are observed across the entire electromagnetic spectrum, from radio waves to gamma rays, and exhibit complex morphologies, including knots, shocks, and collimated flows. They are believed to originate from accretion processes onto the central object or from the interaction of the object's magnetic fields with surrounding matter. Jets play a crucial role in various astrophysical phenomena, including the formation and evolution of galaxies, the launch-

ing and propagation of relativistic outflows, and the feedback mechanisms regulating star formation and black hole growth [2, 15]

1.5.4 The Dust Torus

The obscuring torus, often referred to simply as the "dusty torus," is a key component of the unified model for active galactic nuclei (AGN). This toroidal structure consists of a dense distribution of dust and gas surrounding the central supermassive black hole in AGN. The torus is thought to play a crucial role in shaping the observed properties of AGN and their classification. In the unified model, the orientation of the torus with respect to the observer's line of sight determines the observed appearance of the AGN. When viewed edge-on, the torus obscures the central engine and broad-line region, resulting in the classification of Type 2 AGN, characterized by narrow emission lines and strong infrared emission. In contrast, when viewed face-on, the torus allows a direct view of the central engine and broad-line region, leading to the classification of Type 1 AGN, which exhibits broad emission lines in their spectra. Understanding the structure and properties of the obscuring torus is essential for elucidating the diversity of AGN and its role in galaxy evolution. [1, 5]

1.5.5 Broad and Narrow Line Regions

The Broad-Line Region (BLR) and Narrow-Line Region (NLR) are distinct components of the emission-line regions observed in active galactic nuclei (AGN) spectra. The BLR is located close to the central supermassive black hole and is characterized by broad emission lines with widths ranging from thousands to tens of thousands of kilometers per second. These broad lines arise from gas clouds moving at high velocities in the gravitational field of the black hole. The BLR is thought to be ionized primarily by the intense radiation from the accretion disk surrounding the black hole. In contrast, the NLR is located further from the central engine and exhibits narrow emission lines with widths typically less than a few hundred kilometers per second. The NLR is believed to be ionized by a combination of processes, including

photo-ionization by the AGN's central engine and shocks from outflows or interactions with the interstellar medium. Understanding the properties and dynamics of the BLR and NLR is crucial for unraveling the mechanisms driving the observed spectral features in AGN and their role in galaxy evolution [18, 17].

The anatomy presented in the sections above is known as the **Unified Model** or **Standard Model of AGNs**, as it includes the final product or structure of what a quasar needs to be in order to fit in with every observation to date. There had been no direct observations of a supermassive black hole until the recent radio images of the supermassive black hole in the center of M87 active galaxy, and later Sagittarius A^* at our galaxy's center.

1.6 Motivation

So far, we have discussed the methods and novel ways in which the astrophysics community, throughout the decades, was able to extract bits and pieces of information about quasars, using just the light received from them. These quanta of information fitted together like the pieces of a puzzle to finally reveal one of the most mysterious objects in the cosmos, a quasar; An object so bizarre that its existence was denied for years until they had to accept the unfathomable realities that the cosmos hides.

The important thing to notice here is that, all this was accomplished by analyzing the spectra of these quasars. All that we know, from their structure to size, composition to distance, age to evolution, everything has been encoded in the spectrum. This makes the spectrum, an invaluable asset when studying these objects.

Despite the rigorous ongoing research, the field still has many unanswered questions and unexplored phenomena such as the formation of jets, co-evolution of an AGN and its host galaxy, accretion disk dynamics and its effects on the outflows, etc. The answer to such questions can be answered in two ways, either by **inductive reasoning**, which uses the statistical properties of the bulk sample space to reach a conclusion or by **deductive reasoning**, which involves finding unique instances of the sample space and studying them to know more about the bulk.

Hence in this project, we find spectroscopic anomalies using machine learning. An anomaly would be caused by a disbalance in one or more of the "radiation processes" discussed in Section 1.3, which will translate to an odd behavior being exhibited by one of the parts of the quasar as discussed in Section 1.5. This odd behavior will highlight that particular section of the quasar in the anomalous spectrum and hence will help us study it in isolation giving a deeper insight into the physics at play.

Therefore, we wish to find anomalies in quasar datasets in the hope and attempt to understand and answer a few unsettled questions in the field of quasar astrophysics.

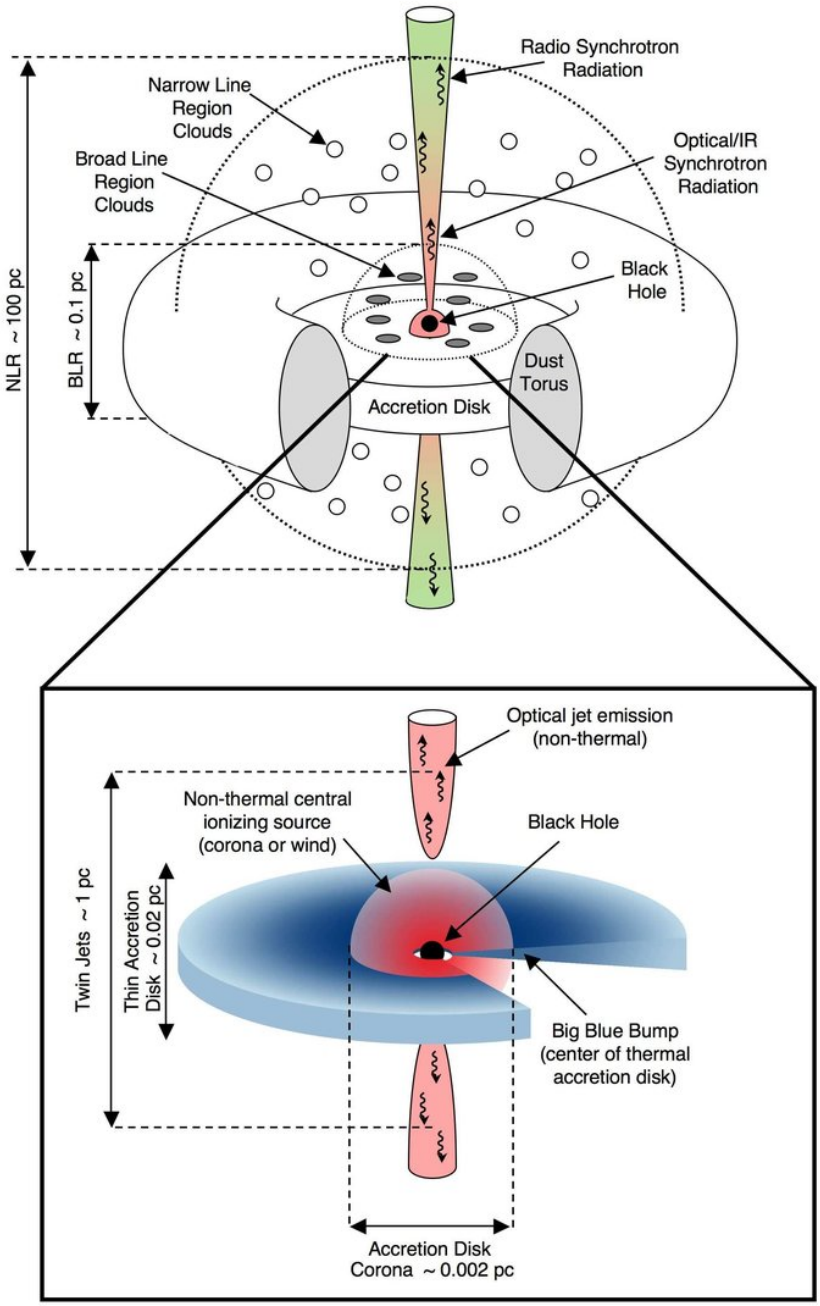


Figure 1.13: Structure of a Quasar and Active Galactic Nucleus

Chapter 2

Methodology

The spectral realm of electromagnetic radiation is a temporal, topological, compositional, and phenomenological signature of the physics at play in the source. Hence, the various lines, their intensities, and widths in the spectrum become the prime messengers in the quest to understand the workings and dynamics of quasars. As we have a standard model for AGNs [29], all quasars are thought to consist of a similar overall structure (Section **??**) emitting a "typical spectrum" as shown in Figure 1.9. Therefore it is straightforward that an anomalously behaving quasar will leave an "uncommon" imprint on the spectrum. This anomalous spectrum can be considered as the "odd-one-out" in a group of similar spectra.

2.1 Data

The spectral data for this project is obtained from SDSS DR16Q Quasar Catalog [14], which contains 750,414 quasars in the redshift range from $0 \leq z \leq 7.1$. This massive redshift window allows the measurement of quasar spectra throughout the high energy electromagnetic spectrum, as the rest wavelength gets shorter with increasing redshift value corresponding to the formula $\lambda_0 \equiv \frac{\lambda_{obs}}{z+1}$. For anomaly detection, we needed a "similar looking" sample space so the odd ones could be detected and labeled as outliers. To achieve

this, we restricted our wavelength window from 1250 to 3000 which captured 4 prominent emission lines (Si[IV](λ1400), C[IV](λ1549), C[III](λ1909), Mg[II](λ2798) of a typical quasar spectrum. With the SDSS Telescope's wavelength coverage of $[3800, 9200]$, our wavelength window translates to a redshift range of $z \in [1.920, 2.167]$ as shown in Fig 2.1.

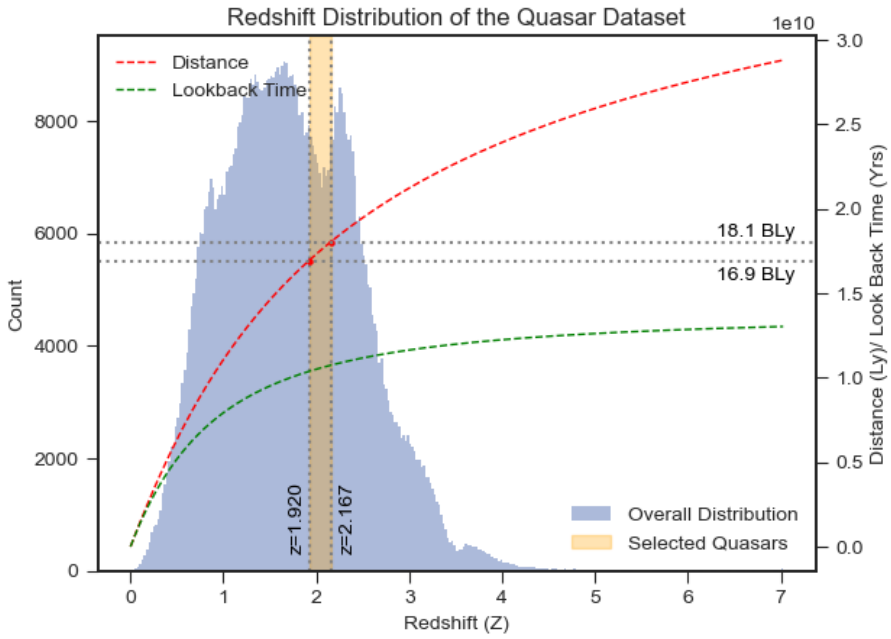


Figure 2.1: Selected quasar sample from the SDSS DR16Q Catalog

In our chosen redshift window, we carry our analysis with two datasets: a) `Control Dataset` containing all the 26830 confirmed quasars in the sample and b) `No BAL Dataset` which contains 17683 Non-BAL quasars which were identified using the 'BAL PROB' keyword ([10]) being > 0.5. All the spectra also have a Signal To Noise Ratio (SNR) > 5. This allows us to extract only the high-quality spectra, as the structure and information retained in the spectrum are of crucial importance in this project due to the lack of human aid and the involvement of machine learning, which is very susceptible to the quality of the input data.

2.2 Spectral Pre-Processing

Each spectrum used in the project undergoes a four-step pre-processing pipeline to be included in the anomaly detection dataset. The pipeline can be understood in the following sections:

2.2.1 Redshift Correction

The quasar spectra from SDSS are read as measured flux ($ergs/cm^2/sec/$Å) for a given observed wavelength (Å). The quasars are spread over a large redshift range, which leads to the Doppler shifting of emission and absorption lines. Hence, we began by applying redshift correction to the spectra using:

$$\lambda_0 = \frac{\lambda_{obs}}{z+1} \tag{2.1}$$

Where λ_0 is the rest wavelength, λ_{obs} is the observed wavelength, and z is the quasar redshift. The redshift value for each quasar is obtained from the "Z" keyword in the SDSS DR16Q_v4 documentation. This brings all the spectra in the singular wavelength window of $[1250, 3000]$.

2.2.2 Flux Correction

Once the spectra are brought to the rest frame, we perform a series of smoothing methods to remove the noise and system-introduced artifacts from the spectra, which otherwise would reduce the efficiency of the anomaly detection algorithms or lead to false outlier detection. The steps followed for this in chronological order are:

1. **Re-Sampling:** The spectra are resampled using `Spectres` [3] (A Python-based spectral resampling module). Resampling means binning the flux values and taking the average flux of each bin as the new value, which reduces the size of the data array. The spectra also contain system-generated Gaussian Noise [22], which, on average, smooths out the fluctuations. The spectra are sampled at 1Å by the telescope, resulting in a typical flux array of length 4000, which, when resampled

with binning of 2 in the desired wavelength window, reduces the dimensions to around 875, which makes it computationally less expensive. This process leads to a neglectable information loss and enhances the quality of the information retained by the spectra by noise removal.

2. **Normalization:** The luminosity of quasars exhibits a wide range of values as it is a sensitive function of the quasar's distance, the mass of its SMBH, composition, orientation, and many more parameters. Despite this variability in luminosity flux values, the overall shape remains more or less the same; there is just a translation of the spectrum on the flux axis. Hence, the resampled spectra are normalized using their maximum flux value, which brings the flux value range for all spectra to $[-1, 1]$. This helps us reduce the chances of the algorithm labeling a spectrum as an outlier just because of its extreme luminosity, which can merely be the effect of it being close as compared to the other quasars in the sample space.

$$(F_{normalized})_\lambda = \frac{F_\lambda}{\mathbf{max}(F_i : i \in n)} \tag{2.2}$$

3. **Smoothing**: The normalized-resampled spectra are passed through a `Savitzky Golay Filter`[20]. This filter works by fitting an n^{th} degree polynomial to a short section (called "window length") of the entire spectrum. The window begins for the start and shifts a "window length" amount after each fitting until it reaches the end. Finally, the fitted polynomials from each window are joined to form a smooth new spectrum. In our project, we used a window length of five and fitted cubic polynomials. This smoothing method removes most of the noise in the spectra without altering the length and shape of the spectral array but leaves out artifacts like extremely narrow spikes in the spectra caused either by cosmic ray encounters or system-induced errors.

4. **Padding:** The spectra recorded by the SDSS survey telescope have

varying dimensions. Hence, after all this pre-processing and extracting the flux values only from 1250 to 3000 , there are still some spectra that are shorter than the required grid. This is resolved by placing each spectrum in this wavelength grid and padding any remaining space with zeros on either side. This step is crucial as the anomaly detection algorithms need all of their samples of the same dimensions for clustering in the same hyperspace.

The final result of all the pre-processing above can be seen in Figure 2.2. The raw spectrum extracted from the SDSS FITS file is represented by blue, while the black shows the processed spectrum, which will be used hereafter.

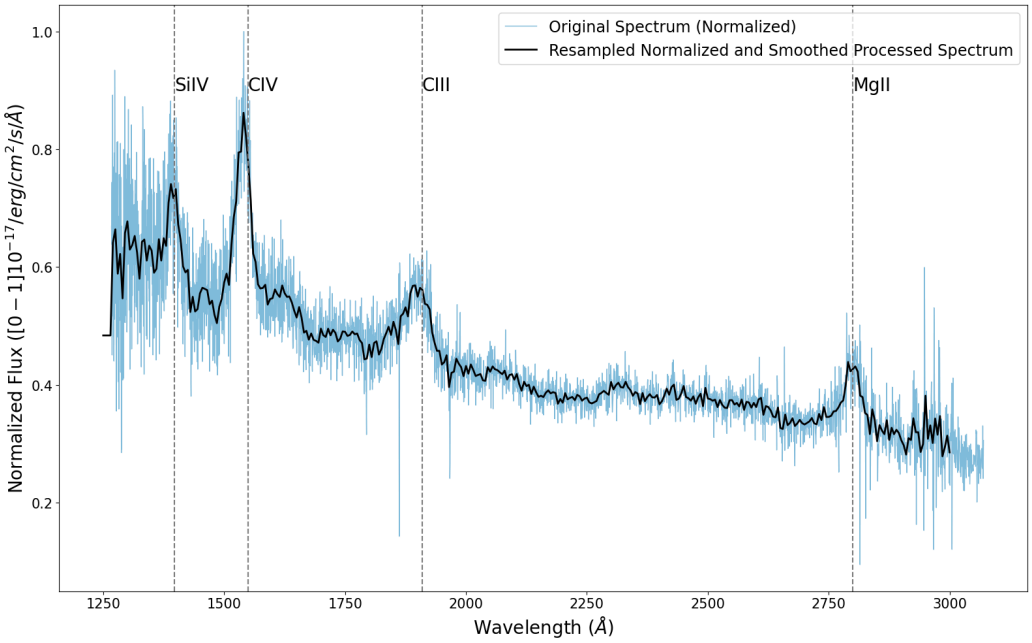


Figure 2.2: Example of a spectrum before and after pre-processing.

2.3 Principal Component Analysis

Definition 2.1. Principal component analysis (PCA) is a dimensionality reduction method that engages eigenvector decomposition to simplify a large

data set into smaller sets while retaining its significant or "Principal" patterns and trends.

Our quasar spectra can be considered as $\sim$ 800-dimensional vectors, with each wavelength corresponding to a unit vector and the flux value at that particular wavelength as its coefficient. This allows us to visualize the entire spectral dataset as points in an $\sim$ 800-dimensional hyperspace, which can then be clustered for anomaly detection.

PCA helps us to reduce the number of these dimensions by constructing new variables called Eigenspectra, which are linear combinations of the initial variables. It is done so that the majority of information stored in the initial variables is squeezed into the initial few components of the PCA components, with each subsequent component containing lesser information than the previous. As we can see in Figure 2.3 the first component PCA 1 lies

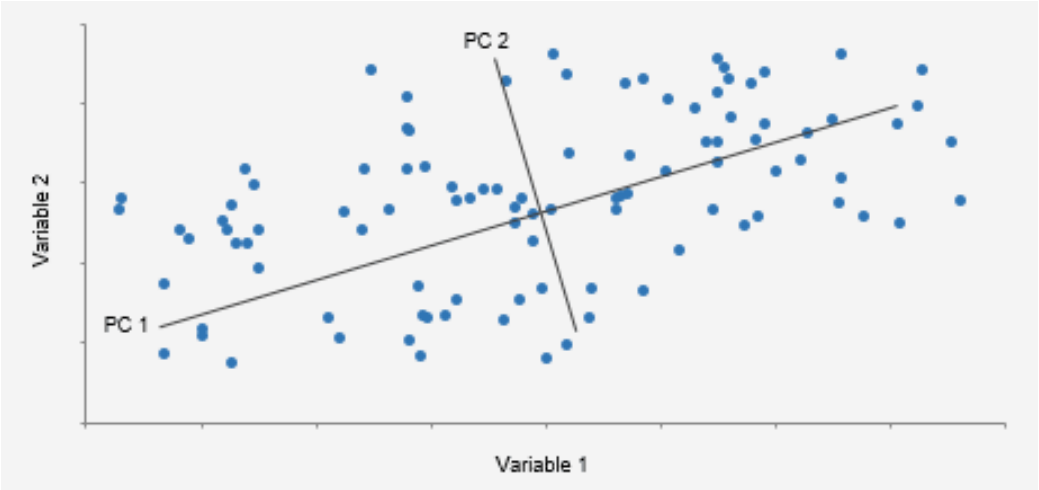


Figure 2.3: Visualization of typical principal components

in the direction of maximum variance of the data, while the second component is orthogonal to it and explains the second highest variance, and so on. This allows us to use only the first few components to retain most of the information, drastically reducing the dimensions while discarding those components with minimal contribution to information retention. The information retained by the components is measured in terms of **Explained**

Variance. It refers to the amount of variability or total variance in the original dataset that is accounted for or "explained" by each component. The higher the explained variance the more important the particular component as the explained variance directly translates to the amount of information retained by the eigenspectrum.

In our project, we use 30 Principal components for both of the datasets which is able to have a cumulative explained variance of 95.2% for the `Control Dataset` and 95.3% for the `No BAL Dataset`.

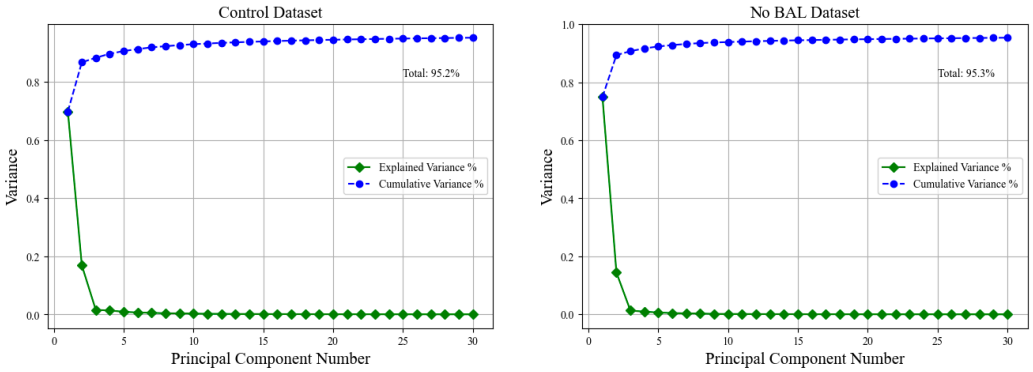


Figure 2.4: Individual (green) and Cumulative (blue) Explained Variance per Principal Component for both datasets.

We agreed on a total explained variance of $\sim 95\%$ for our project. Figure 2.4 shows the contribution in the total explained variance rapidly diminishes with each subsequent principal component and the cumulative sum reaches our required value with 30 components, with more than 90% of the variance being explained by just the first three components.
Application of PCA reduces the 875 dimensional vectors to 30 dimensions and at the same time drastically reduces the dimension of the spectral hyperspace too. This makes it much more computationally efficient. The clustering is done by calculating the Euclidean distance between data points, and grouping the close by points, a reduction in dimensions also helps in better clustering since the sense of cluster structure tends to become hazier with increasing dimensions.

2.3.1 PCA Reconstruction

Since our PCA has a variance explanation of 95%, which describes the total fraction of spectra that can be described by the linear combination of PCA eigenvectors; there are some spectra that are not being accounted for by the PCA. If we plan to use the PCA Eigenvector coefficients for our clustering, we need to remove the 5% unresolved spectra for better results. We achieve this by reconstructing each spectrum using the PCA Eigenvectors. We then calculate the RMS error for each spectrum, which is calculated as:

$$\text{RMS Error} = \frac{\Sigma_i^N \sqrt{(F_O|\lambda_i - F_R|\lambda_i)^2}}{N} \tag{2.3}$$

Where, N is the total number of spectra, $F_O|\lambda_i$ is the flux value in the original spectrum at a given wavelength λ_i and $F_R|\lambda_i$ is the flux value of reconstructed spectra at the same wavelength.

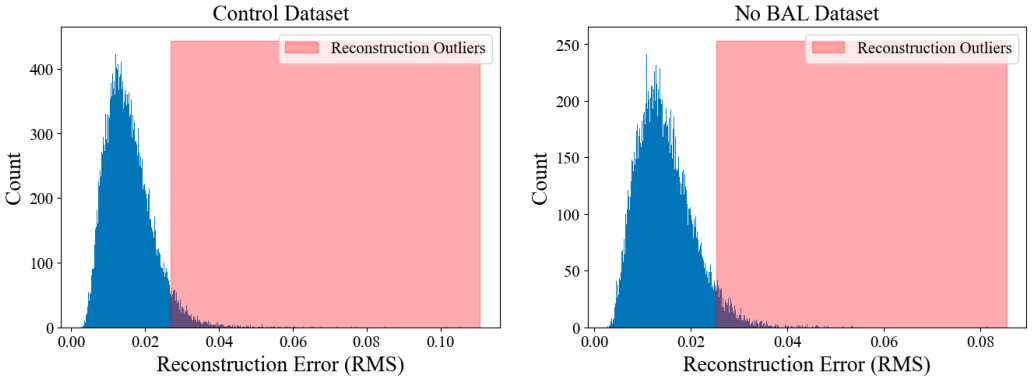


Figure 2.5: RMS Error Distribution for PCA Spectral Reconstruction. The red-shaded region denotes removed spectra.

We then remove the top 5% of spectra with maximum error, which is marked by the red-shaded region in the figure. These top 5 percent of spectra account for "Un-Explained" Variance by PCA.

2.4 K-Means Clustering

Definition 2.2. Clustering is a technique used in data mining and machine learning that includes grouping similar data points based on their properties, either human conceivable or hidden.

K-Means Clustering is one of the most widely used types of clustering algorithms. It partitions the data points into "K" disjoint subsets usually denoted as "C_K" containing "N_K" points each, in such a way that the sum-of-squares objective function is minimized:

$$\sum_{K=1}^{K} \sum_{i \in C_K} ||x_i - \mu_K||^2 \tag{2.4}$$

Where, $\mu_K = \frac{1}{N} \sum_{i \in C_K} x_i$ is the mean of the points in set C_K, and $C(x_i) = C_K$ denotes that the class of x_i is C_K.[24]

In layman's terms, K-Means clustering works by assigning "K" (user-defined) data points as nucleation sites. It then calculates the distance of each point to all K nuclei and assigns the point to the nuclear closest to it, hence forming a group. With the addition of each point in the group, the centroid of all the points in each group is recalculated and new data points are added until all of the data points belong to any one of the clusters. This entire process is repeated a user-defined times and each time optimization matrices calculate the efficiency of the clustering. The centroids that provide the most compact and distant groups are chosen as the best.

2.4.1 Optimum Number of Clusters

There are several types of AGNs present in the SDSS catalog. On the basis of different properties they can be grouped in a large number of different disjoint groups. Since we were looking for anomalous quasars, where the anomalous behavior could have emerged from any section of the spectrum, we did not use domain-specific knowledge to sort the quasars into groups

before looking for anomalies. Instead we two used standard techniques to find the optimum number of clusters for our datasets. These are:

1. **Silhouette Coefficient:** The silhouette coefficient is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters[19]. The Silhouette Coefficient for a singular data point is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.5}$$

   Where:

   - $s(i)$ is the silhouette coefficient for data point i,
   - $a(i)$ is the average distance from i to other data points in the same cluster,
   - $b(i)$ is the smallest average distance from i to data points in a different cluster.

   The silhouette coefficient ranges from -1 to 1, where:

   - $s(i) = 1$ indicates that the data point is very well matched to its own cluster,
   - $s(i) = 0$ indicates that the data point is on the boundary of two clusters,
   - $s(i) = -1$ indicates that the data point is probably assigned to the wrong cluster.

2. **SSE Score:** The Sum of Squared Errors (SSE) score, also known as the Within-Cluster Sum of Squares (WCSS), is a measure commonly used to evaluate the quality of clustering algorithms. It represents the sum of the squared distances between each data point and its assigned centroid within a cluster.[7]

The formula for calculating the SSE score is as follows:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{2.6}$$

Where:

- C_i represents the data points assigned to cluster i,
- μ_i is the centroid of cluster i,
- $||x - \mu_i||^2$ denotes the squared Euclidean distance between data point x and the centroid μ_i of its assigned cluster.

We use these two metrics to determine the optimum number of clusters in our datasets. This is done using a technique called the "**Elbow Method**".

Definition 2.3. The Knee Method, also known as the "elbow method," is a heuristic used to determine the optimal number of clusters in a dataset for a clustering algorithm. It involves plotting the SSE (Sum of Squared Errors) against the number of clusters and identifying the "knee" or the point where the rate of decrease in SSE slows down, indicating the optimal number of clusters. [27]

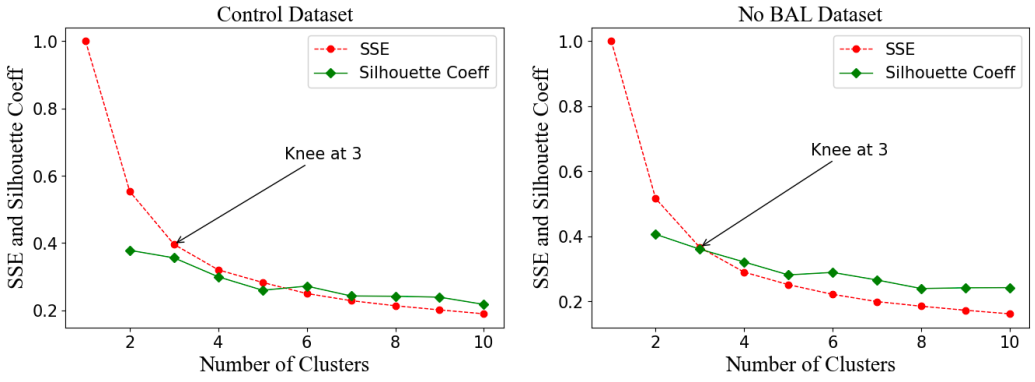


Figure 2.6: The optimum number of clusters as concluded by Elbow Method for both datasets

As per Figure 2.6, since Knee is at 3, the value of K=3 for both datasets.

2.4.2 Cluster Visualization

Once the data points are clustered, we can visualize them by plotting the principal component coefficients. The clustering would be best visible in the initial few principal components as those are the ones that have the maximum variance explanation. Figure 2.7 shows the clustering for the control dataset, where we can see that the clustering becomes less and less evident with the higher order of the principal component. The X and Y axis of the

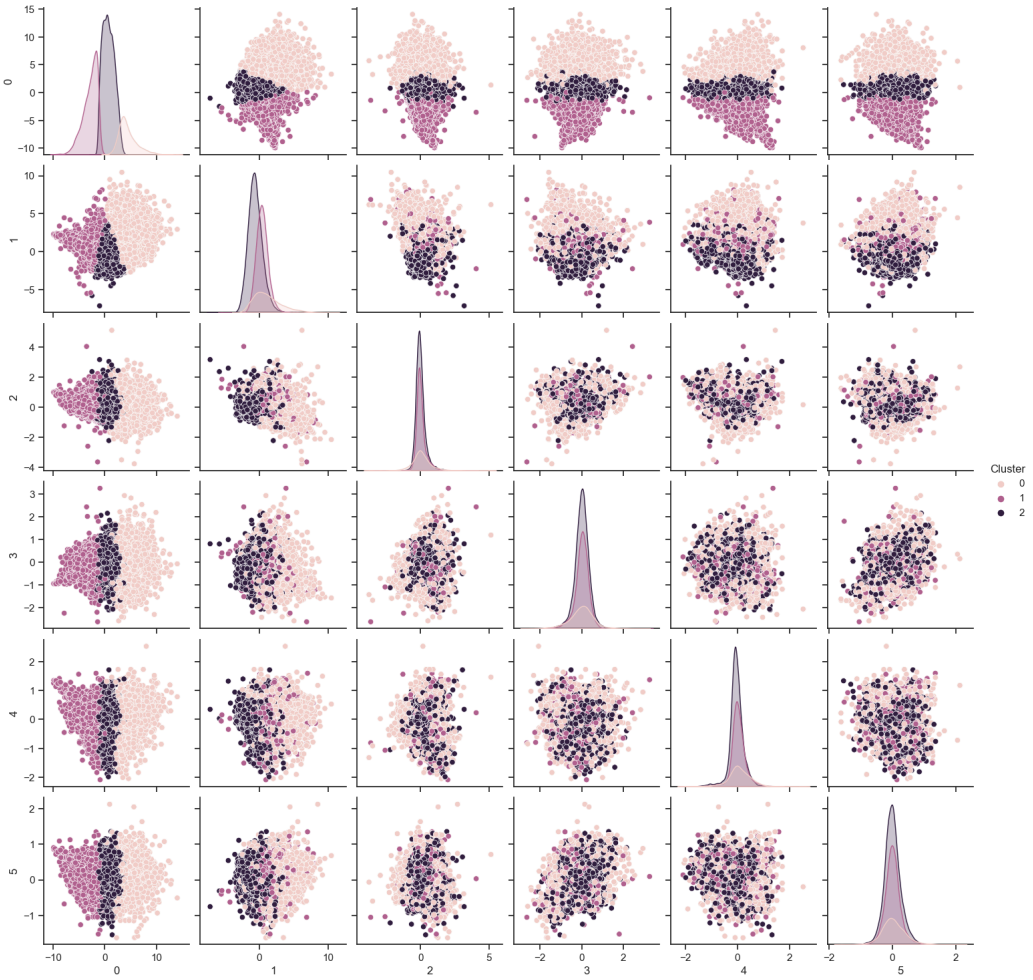


Figure 2.7: Principal Component clusters for Control Dataset

plot depicts the first six principal component coefficients scattered with respect to each other, while the diagonal contains the KDE (Kernel Density

Estimate) histogram for each PC coefficient. Similarly, Figure 2.8 shows the cluster visualizations for the No BAL Dataset. Since the first two principal components account for the maximum cumulative variance, the cluster separation while plotting them is evident. Hence, from here on we will be using PCA_0 and PCA_1 for visualizing the clusters. It is also observed that

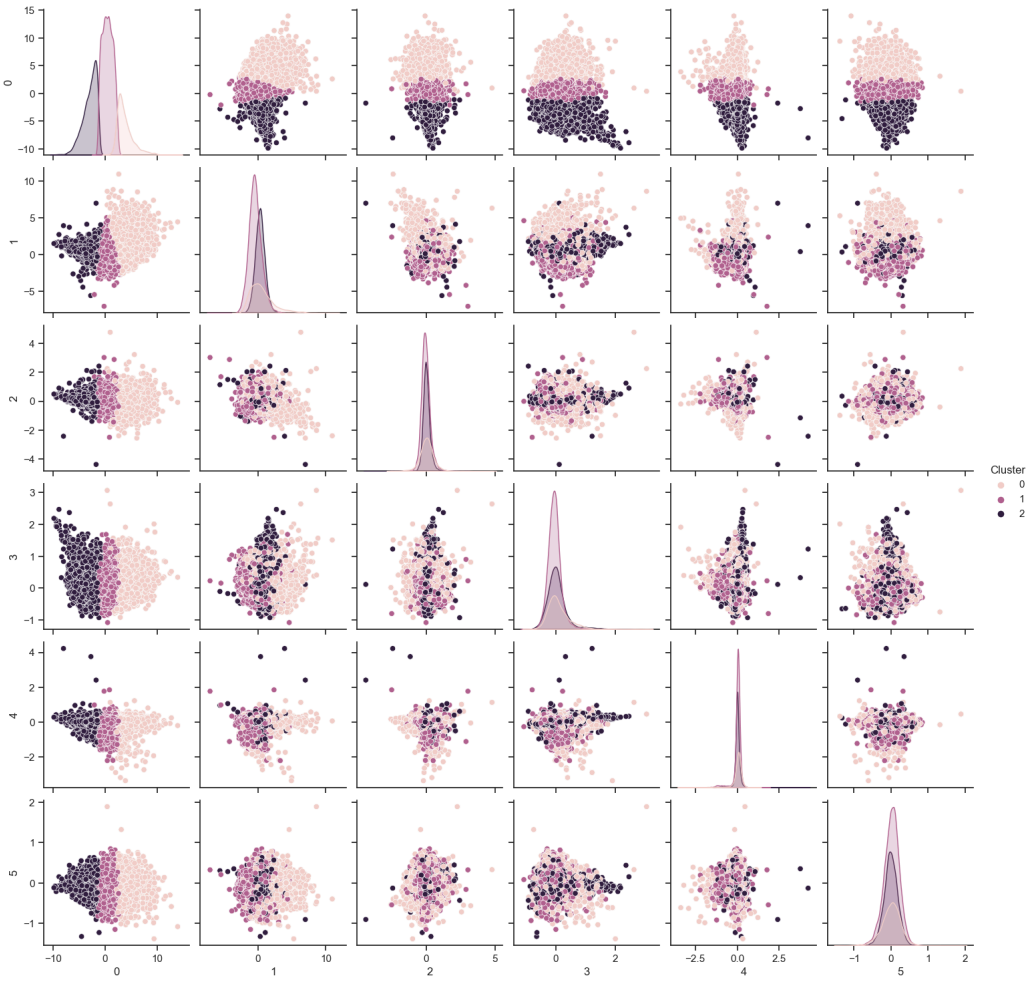


Figure 2.8: Principal Component Clusters for No BAL Dataset

the clusters do not seem to have an evident separation. This is because they exist in a 30-dimensional hyperspace while the visualization is a mere two-dimensional projection, which might not capture the entire separation of the clusters. One would need access to 30 dimensions to see the exact cluster, which is beyond human comprehension. Figure 2.9 shows the color-coded

clusters as a 2D projection with the help of a scatter plot between PCA_0 and PCA_1 coefficients for both datasets. Now that we have three different clusters, any data point residing very far from the cluster centroid will be called an outlier and would have a high probability of being an anomalous quasar.

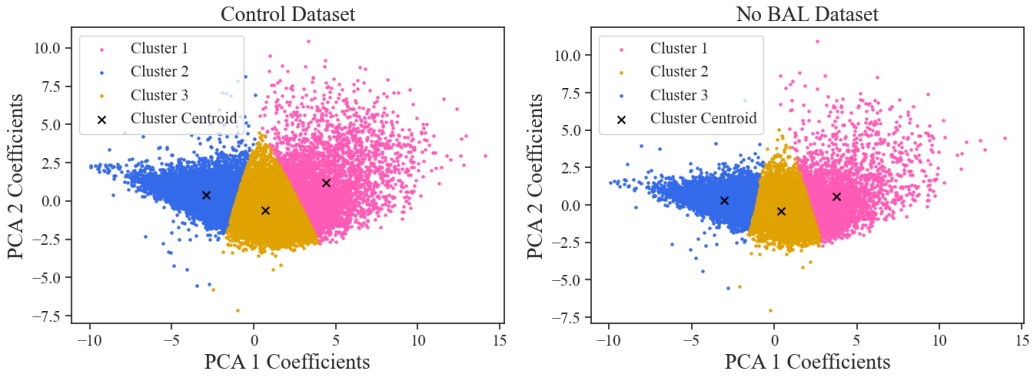


Figure 2.9: Quasar spectroscopic clusters with centroids marked

2.4.3 Anomaly Detection

For anomaly marking, we calculate the Euclidean distance of each data point from its respective cluster centroid and plot the histogram for `Control Dataset` in Figure 2.10a and the `No BAL Dataset` in Figure 2.10b. Any point residing beyond 5σ distance was labeled as an outlier. Where σ is the standard deviation of the distribution, given as:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \tag{2.7}$$

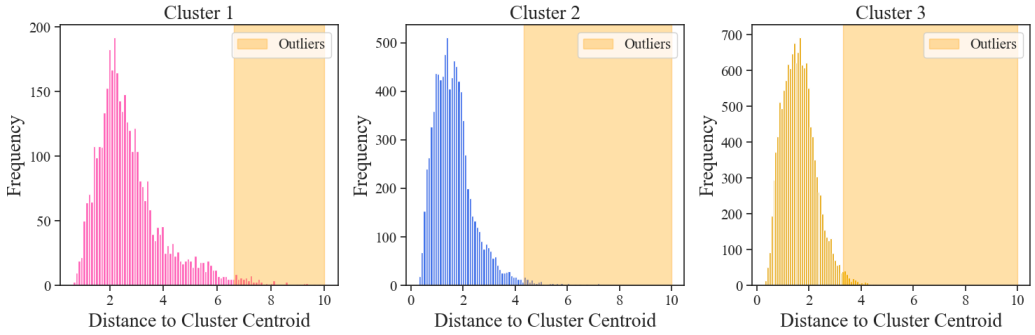
In the figures below, all the quasars inside the yellow-shaded region are considered outliers or anomalous as they are too far from the cluster centers hence having a high probability of containing a "weird" feature in their spectra. When the PCA tried to account for this weirdness, it generated unusual coefficients for the eigenvectors which caused these points to reside farther than a typical point from the cluster center.

Once this distance threshold is applied to the clusters in both datasets, we

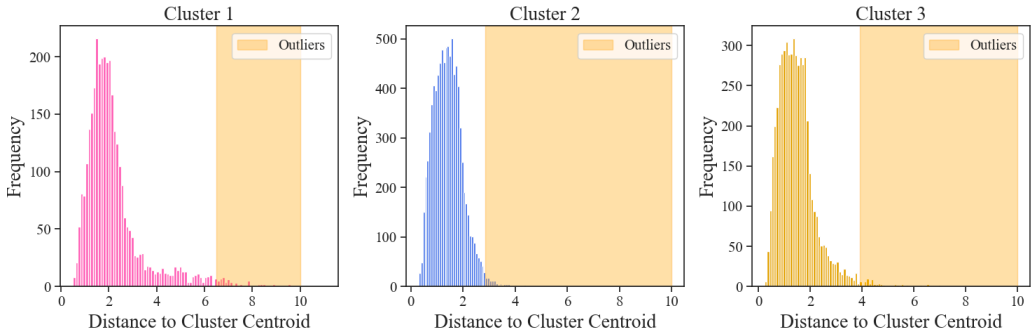
get a total of 472 anomalies in `Control Dataset` and 277 anomalies in `No BAL Dataset`. The exact distribution per cluster is given in Table 2.1.

Dataset	Cluster 1	Cluster 2	Cluster 3
`Control`	81	146	245
`No BAL`	63	82	132

Table 2.1: Anomalies in each cluster for both datasets



(a) Control Dataset



(b) No BAL Dataset

Figure 2.10: Histogram for the Euclidean distance of each point from its respective cluster centroid

Once the anomalies are detected, we plot the coefficients of their first and second principal components on top of Figure 2.9. The anomalies are marked with black dots concentrated on the peripherals of each cluster as seen in Fig 2.11. Again, some anomalies might seem "inside" the clusters; This is because the clusters visualization is just a 2D projection of the 30-D

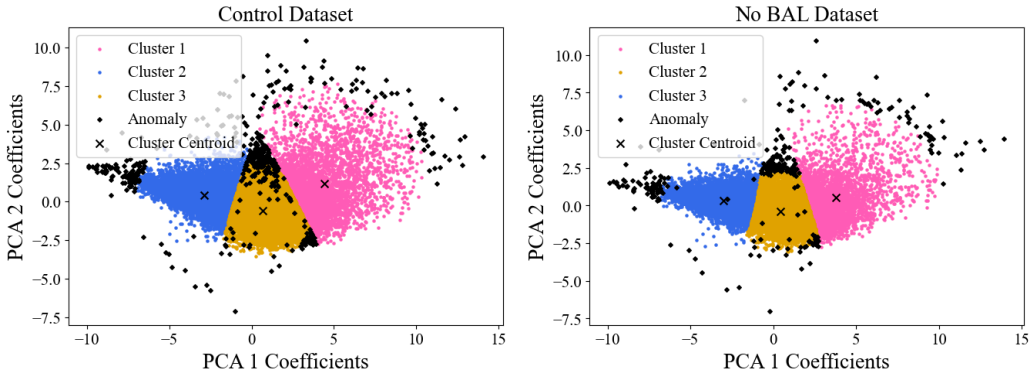


Figure 2.11: Quasar spectroscopic clusters with centroids marked and Anomalies overlayed

hyperspace, and the point seemingly inside the cluster might very well be far off in one of the dimensions making it an outlier.

2.4.4 Composite Cluster Spectra

Since the clustering is done in the hyperspace that contains coefficients of the PCA, we do not know for sure, what exactly causes these quasars to split up into three different groups, because PCA coefficients do not translate into any physical property of the quasar. Hence in order to have a crude understanding of the "cause" of this clustering, we will have to look at the mean spectrum for each of the clusters.

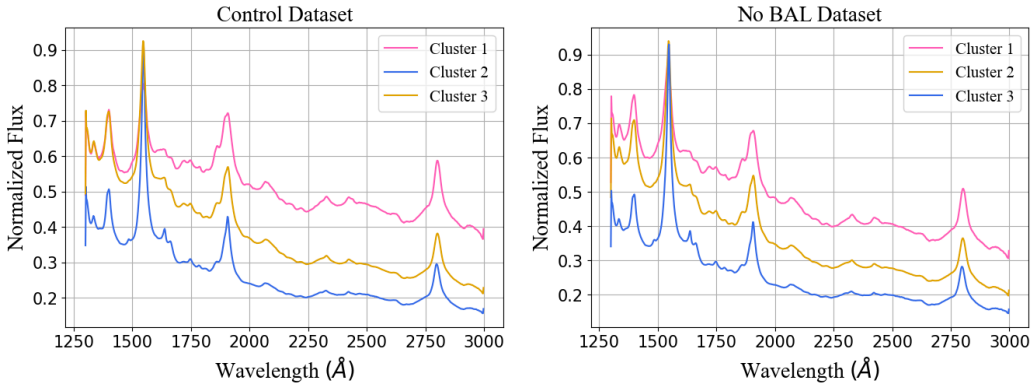


Figure 2.12: Composite Spectrum for each cluster

The mean spectrum helps us to understand the general trend observed in the spectra of the cluster members, which translates to a common physical property as seen in Figure 2.12. A detailed discussion about the implications and the mean spectra as a whole can be found in section **??**

2.5 Anomaly Grouping

As stated in Table 2.1 the number of anomalies is very large, and these anomalies are spread throughout the PCA Coefficient hyperspace as seen in Figure 2.11. When a visual inspection of the anomalous spectra was done, they were found to be exhibiting repetitive properties such as extremely sharp C_{IV} peaks, excessive Si_{IV} emissions, faulty spectrum (Machine Error) etc. Hence, in order to make a sensible inference and to study these anomalies better, we decided to group these anomalies again using K-Means clustering following the same steps we did for the core dataset.

2.5.1 Principal Component Analysis

This time, a PCA (with 30 components) was run only on the anomalous spectra detected in Section 2.4.3. The total explained variance is about ∼98% for

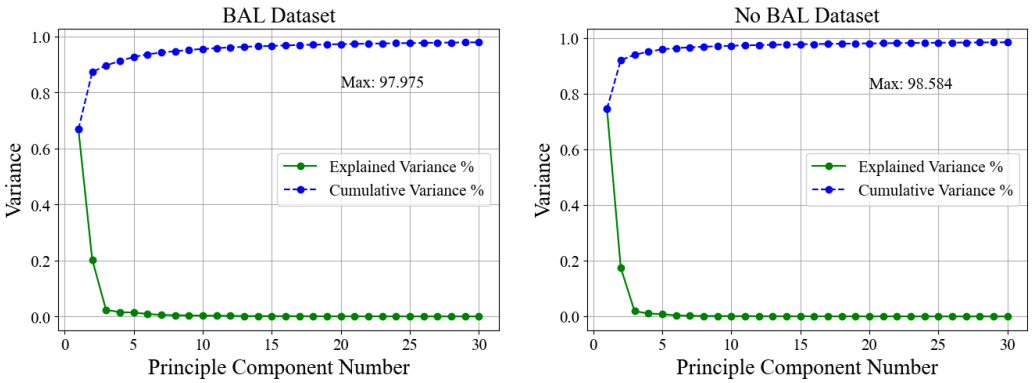


Figure 2.13: Individual (green) and Cumulative (blue) Explained Variance per principal component for anomalies

the anomalies of both the datasets. Since the explanation is very high in this

case, we did not feel the need to perform the removal of unexplained spectra by the means of PCA reconstruction RMS error, as was done previously in Section 2.3.1.

2.5.2 Optimum Number of Clusters

Once the spectral outliers are captured in a 30-Dimensional hyperspace by PCA eigenvectors, we progress forward by finding the optimum number of clusters for the datasets. This is done exactly following the paradigm as discussed in Section 2.4.1. This process was much faster for this case as compared to the entire dataset because of the staggeringly less sample space of anomalies as compared to the complete dataset. As seen in Figure 2.14 the optimum number of clusters varies this time, with K=3 for `No BAL Dataset` while K=4 for `Control Dataset`. This difference in K value seems obvi-

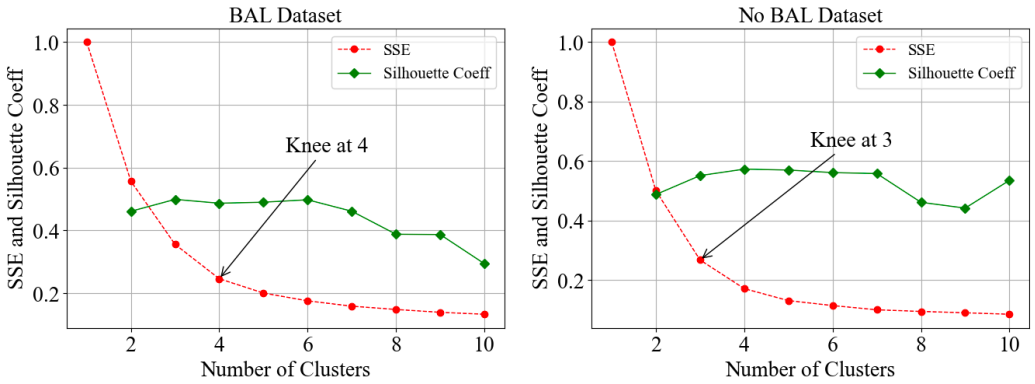


Figure 2.14: Optimum Number of Clusters for Anomalies of Both datasets

ous because the `Control Dataset` contains the additional "BAL Quasars" which are deliberately removed from the `No BAL Dataset`. The weird trend observed in the Silhouette Coefficient (sudden drop towards the higher side of cluster number) points towards a chunky cluster i.e. the clusters have sub-clusters or regions of high and low density, which can be considered as smaller clusters within the bigger clusters. These clusters are too small to be considered as an efficient clustering, similar to the concept of over-fitting by a model. Hence we stick to the number of optimum cluster suggested by the Elbow Method (Def 2.3).

2.5.3 Cluster Visualization

The anomaly clusters are much more prominent and fairly separated as compared to the clustering in the original dataset as seen in Figure 2.7 and 2.8.

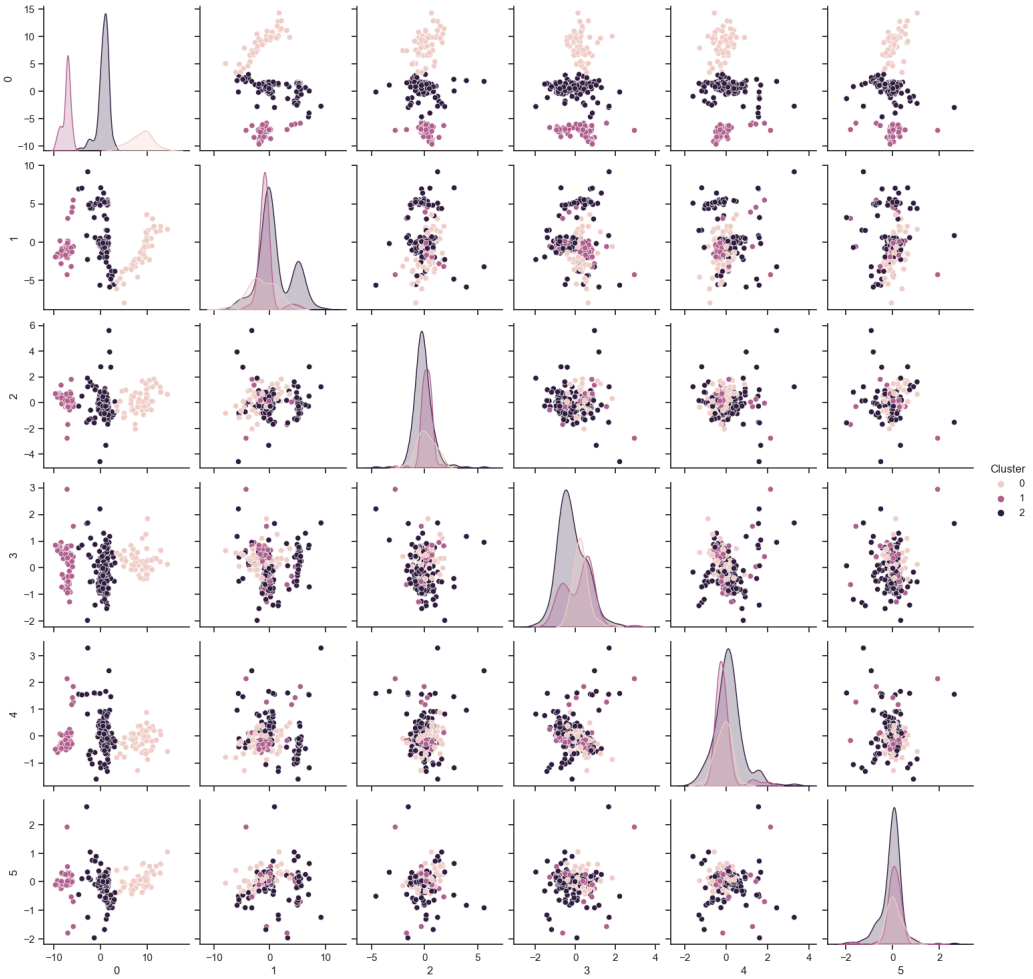


Figure 2.15: Principal Component Coefficient Clusters for No BAL Dataset Anomalies

This is primarily because of two reasons,

- There are much less data points in the sample space making the distribution less crowded providing the aesthetic "negative space" which makes the clusters much more distinguished and dense.

- The anomalies in a group have much more in common than the cluster members of the initial clustering. This is a direct implication of removing most of the objects and grouping only those are are far away from the cluster center. This property of being far away is common among all the anomalies which makes clustering much simpler and efficient.

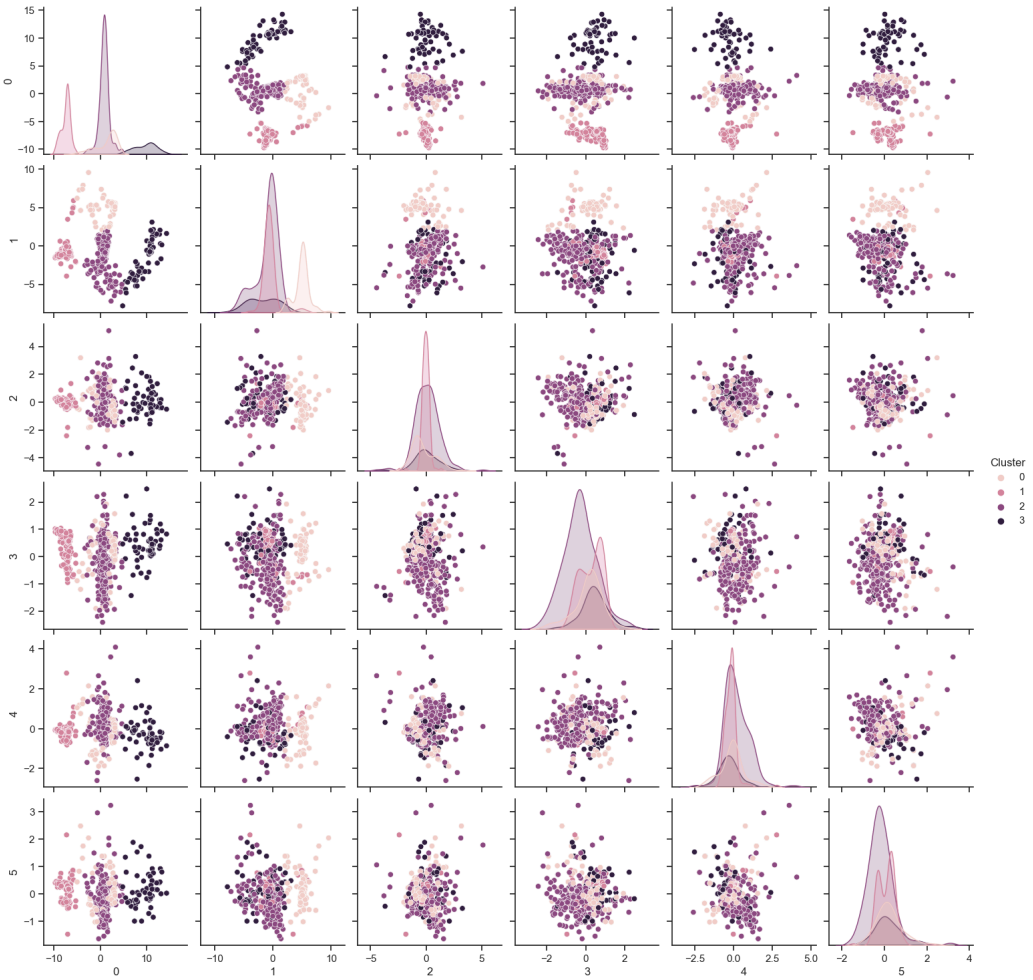


Figure 2.16: Principal Component Coefficient Clusters for Control Dataset Anomalies

As seen in Figure 2.15 there are only three visible clusters which matches with the elbow method calculations, where as Figure 2.16 contains a fourth (bit non-obvious) cluster which is supposed to be that of BAL Quasars. Figure

2.17 shows the color coded Anomaly groups with their centroids marked. Most of the members of the green group in the control dataset anomalies are absent in the No BAL anomaly diagram, which suggests that this group must contain the BAL Quasar anomalies. For the exact details of the general properties for the members of the group, we will have to plot the mean spectrum for each. This will give us a crude idea of the general trend in their spectra which would translate to a common physical process happening in those quasars that set them apart from rest of the cluster members hence making them an outlier.

> **Note:** *Here on, we will use the word* ***Cluster*** *for the initial clustering of the entire dataset into three clusters, and the word* ***Group*** *for the grouping of the anomalies into three and four groups for the* `No BAL Dataset` *and the* `Control Dataset` *respectively.*

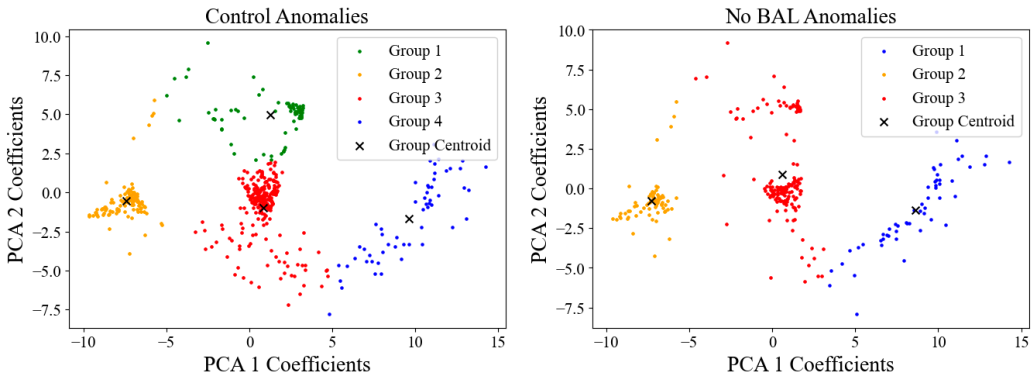


Figure 2.17: Color coded Anomaly Cluster with centroid marked

The exact number of anomalies in each group is given in Table 2.2. In both

Dataset	Group 1	Group 2	Group 3	Group 4
`Control`	77	110	230	55
`No BAL`	54	76	147	-

Table 2.2: Anomalies in each group for both datasets

cases the group 3 (Shown in "red" in Figure 2.17) contains the most data points.

2.5.4 Composite Anomaly Spectra

The mean spectrum of the anomaly groups helps us to identify the general properties of the quasars common to the group. As seen in Figure 2.18, the mean spectra for each of the group are color coded in the same color scheme as in their PCA coefficient distribution plot in Figure 2.17. The spectra vary greatly in terms of luminosity, spectral index, shape, absorption and emission lines and relative equivalent line widths. In depth analysis of these anomaly

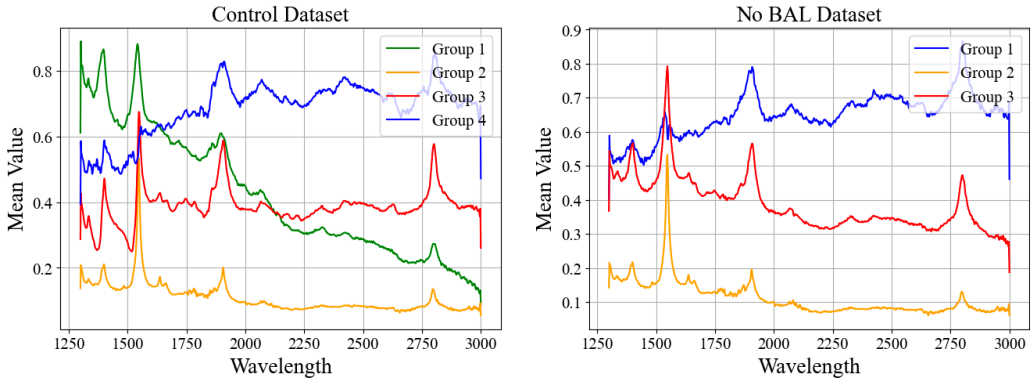


Figure 2.18: Composite Spectra for Anomaly Groups

groups can be found in Section 3.1, where we discuss the physical properties of each of the group members, their possible origin, population statistics, and manually hand pick some of the most interesting members of each group by the means of detailed visual inspection.

2.6 Science Products

The analysis discussed in the above sections from the spectral pre-processing up till the anomaly grouping provided us with a few science products, that we will be using for our further physical analysis and discussion about the quasars. The science products include,

- **Plot PDFs** These PDFs contain a spectrum plot on each page, where each plot contains
  - Anomaly Spectrum

  - Mean Group (of the anomaly) Spectrum
  - Mean Cluster (of the anomaly) Spectrum
  - SDSS Skyserver link hyperlink
  - PLATE, MJD, FIBERID of the object
  - RA, DEC and Redshift values

  The PDF product is for each anomaly group for both the dataset, making a total of seven PDFs.

- **Catalogs:** We created three catalogs based on the most prominent property of the anomaly group, which are:
  - Sharp C_{IV} Peaking Quasars
  - Heavily Reddened Quasars
  - Unaccounted BAL Quasars

- A list of unique objects, that were placed in one of the groups by the algorithm but revealed much weirder features upon visual inspection. These object either contained additional anomalous properties in addition to the mean property of the group or were totally different of unexplainable origin, but were placed in that particular group because of sheer resemblance of some sections of the spectrum to the group mean plot.

Chapter 3

Results

The methodology presented in Chapter 2 is a refined and well organised version of the endeavours and branching efforts that we took in the past year. Several amendments and minor adjustments were made in the algorithms throughout the project in order to make the outcome robust and replicable. In this chapter, we present the major findings of our analysis and their implications.

3.1 Anomaly Groups

As discussed in Section 2.5, we used K-Means clustering to group the anomalous quasars of into three and four groups for the `No BAL Dataset` and the `Control Dataset` respectively. In order to understand what each of these groups correspond to, we created their composite spectra as shown in Figure 2.18. The shape and features present in the mean spectrum of any group betrays the common feature of that group, helping us identify the quasars present in the group. We also created plots for individual anomalies, and compared them to their mean group and their mean cluster spectra, which helped us understand what exact feature of these anomalies makes it stand out from the rest of the cases. This also helps us identify the exact emission process and hence the physical location and process that might be causing the anomalous behavior. The Figure 3.1 below, shows the anomaly groups

marked in different colors, overlayed on top of the clustered "normal" quasars, and labelled on the basis of their mean spectrum.

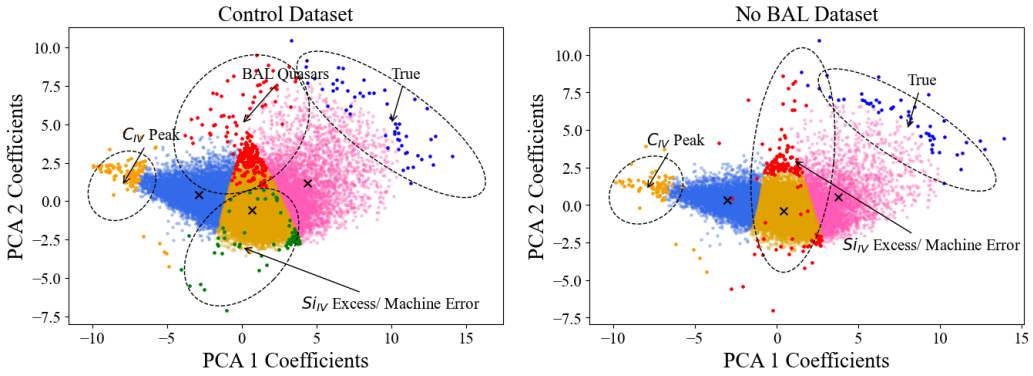


Figure 3.1: Anomaly groups

The four groups (three common in both) of anomalies are:

1. **Sharp CIV Peakers:** These are the quasars marked in yellow on the left most part of both the distributions. These quasars exhibit an excessively high and disproportionate **CIV** emission line (λ1549). The peaks are extremely narrow and and are the overall maxima of the plot.

2. **Excess SIV/Machine Error**: This group contains two types of quasars. First, those that have a corrupted spectrum caused by missing data points due to a fault in the capturing device(SDSS Telescope). We do not analyse them, they are just removed as bad data. Second, quasars with excessively high and broadened emission at **SIV** (λ1400) and immediately higher wavelengths. These have a physical implication, which will be discussed in the following sections.

3. **BAL Quasars**: This group is only present in the control dataset, where we allowed the presence of BAL Quasars. The anomalies are unusual or rare BAL quasars, typically being Lo-BAL and FeLo-BAL Quasars. The exact details about these subgroups and their distribution will be discussed shortly.

4. **True Anomalies:** This group, on the rightmost edge of the distribution is named true anomalies, because of the lack of a general trend in

the group members. The anomalies present here seem to have no correlation and most of them have an "unexplainable" feature. These contain a mixture of properties, from heavily reddened quasars, to quasars having weird absorption lines and spectral indices. We will pick out a few peculiar objects from this group, which we consider are interested and need detailed studies on individual basis in order to elaborate upon the physical processes causing them.

3.1.1 Machine Error Anomalies

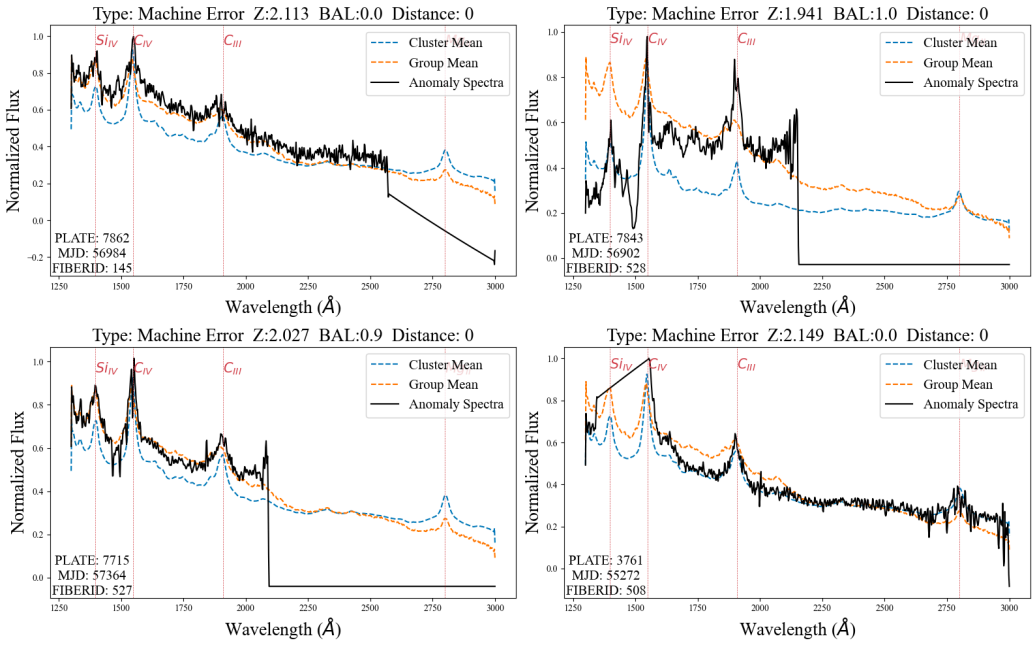


Figure 3.2: Machine Error Anomalies

The plots shown in the figure above depict typical anomalies caused by a corrupted spectrum (observe the blank spots in the spectra filled in by a linear line connecting the disjoint ends). The SDSS telescope utilizes an optical fiber with CCDs to capture these spectra which is subject to failure because of several external factors such as cosmic ray hitting the CCD hence overexposing the pixels, which are then removed by the algorithm or the object of interest moves out of view during raster scan. There can also be

internal factors, such as lapse of data during storage and transfer etc. These are instances of bad data points and show up as anomalies because of these flat features in them. In some sense, these can be called "pseudo anomalies", because their weirdness is merely an instrumental mistake, hence, we do not utilize or analyse them, but rather store them into a dataset which would then be sent back to the SDSS team for rectification or removal from the database.

3.1.2 Excess SiIV Emitters

These quasars exhibit an unusually high and broadened SiIV emission line, which, in some cases starts to blend in with the Ly-α forest just blueward to it. In the canonical quasar spectrum seen in Figure 1.9 we observe that the SiIV emission line is much shorter and narrower than the CIV emission line (λ1549). These emissions are in UV-Optical range which originates in the outer edges of the accretion disk [18]. The excessive and broad emission of SiIV is always paired with shallow and narrow absorption. SiIV is known

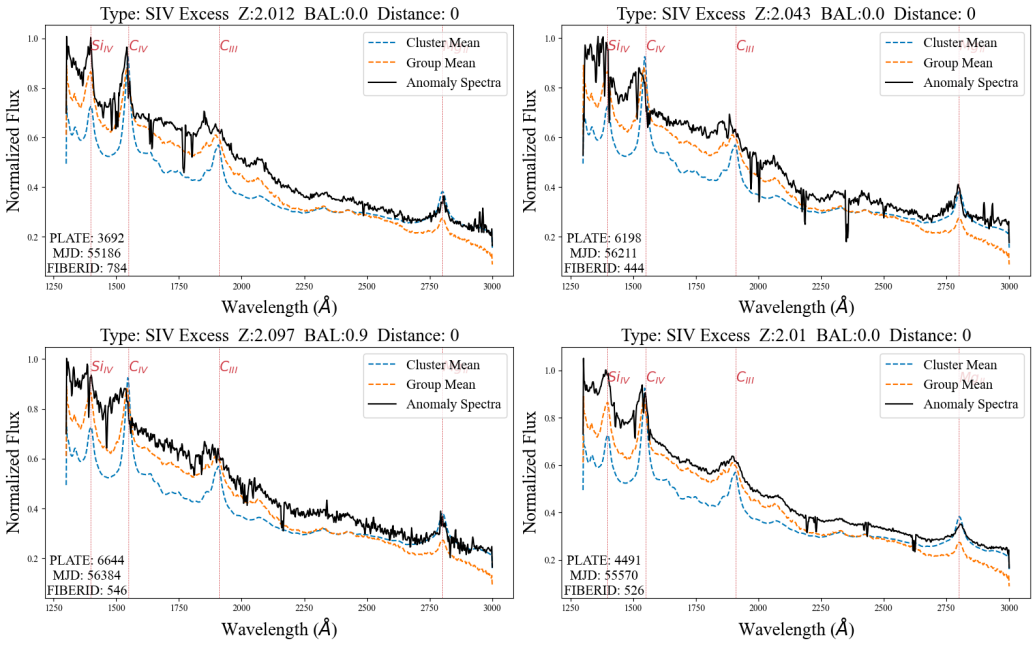


Figure 3.3: SiIV Excess Anomalies

to form in the penultimate stage of a high or intermediate mass star. Since the emission lines are broad, we know that they are moving at high velocities as thermal motion is insufficient to account for such massive velocity spread. For this to happen, the Si atoms sitting in the core of the stars need to be pulled out and fall in the vicinity of the accretion disk which them heats them up to cause the emission and rotates them at high velocities causing broadening. The exact mechanism of this happening is a subject of extensive studies and would be pursued in future work.

3.1.3 Sharp CIV Peaking Quasars

The CIV emission line is the most prominent feature of the chosen wavelength window for nearly all the quasars (except a few BALs). This emission lies in the higher end of UV regime which mostly originates in the accretion disk [18]. The emitting particles present in the accretion disk travel at extremely high virially induced velocities, which translates into a heavy broadening of the emission line. The CIV emission lines seen in Figure 3.4 seem to be extremely narrow and have significantly higher flux as compared to a normal quasar. The high-ionization broad emission lines are usually related to the metallicity of the emitting gas, given all the other factors remain constant [11], some even considering metallicity as the only factor affecting it. The metallicity is noted to increase with an increase in a) quasar luminosity b) black hole mass c) accretion rate d) emission line outflow signatures . With the current observed luminosities, the object would need about five times the solar metallicity, indicating to a rapid chemical enrichment process in the AGN withing the first 500 Myr of the first stellar formation in the galaxy, which is then to be sustained.[26]. The rapid chemical enrichment is usually accomplished by frequent tidal disruption events (TDE) of the circumnuclear stars. The narrow spread of the CIV emission like could also be an illusion created because of its extreme luminosity, which needs to be carefully calculated and verified in future work. We also obverse, that the spread in the emission peaks are asymmetric, pointing towards a biased flow of materials. This can be understood by the scenario of a hot gas cloud

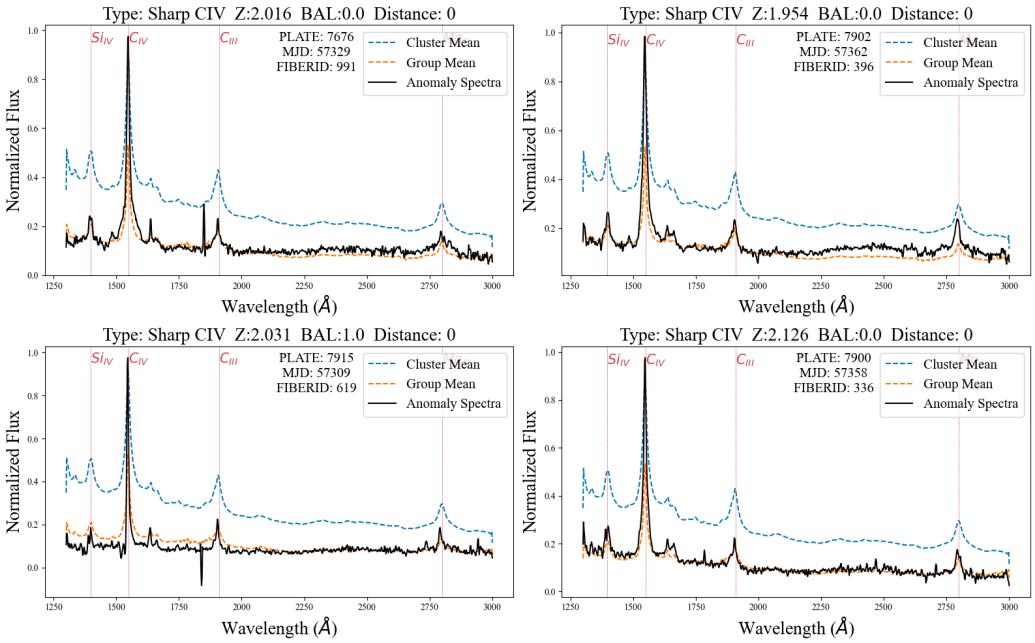


Figure 3.4: Sharp CIV Peaking Anomalies

moving towards us individually while being dragged away by the galactic motion and cosmological redshift, hence causing the slightly lower spread in the redder portion than the bluer one.

3.1.4 BAL Quasars

This anomaly group belongs only to the `Control Dataset` as we allowed the presence of about 9000 BAL quasars in this dataset. BAL quasars are identified by the presence of one or more broad and deep absorption lines in the spectrum, and constitute about 15% of the total quasar population. This feature can be observed in all the spectra in Figure 3.5, where two broad absorption lines can clearly be observed around 1500. On a closer inspection, we find additional small and shallow absorption lines at Al[III](λ1857) and MgII (λ2798), which identify these objects as **LoBAL** quasars, which account for about 2% of the total quasar population (2.4% in our sample space). The general interpretation of BAL quasars is that we are looking at the bright AGN through a dense outflowing dust and gas cloud, which results in a redder

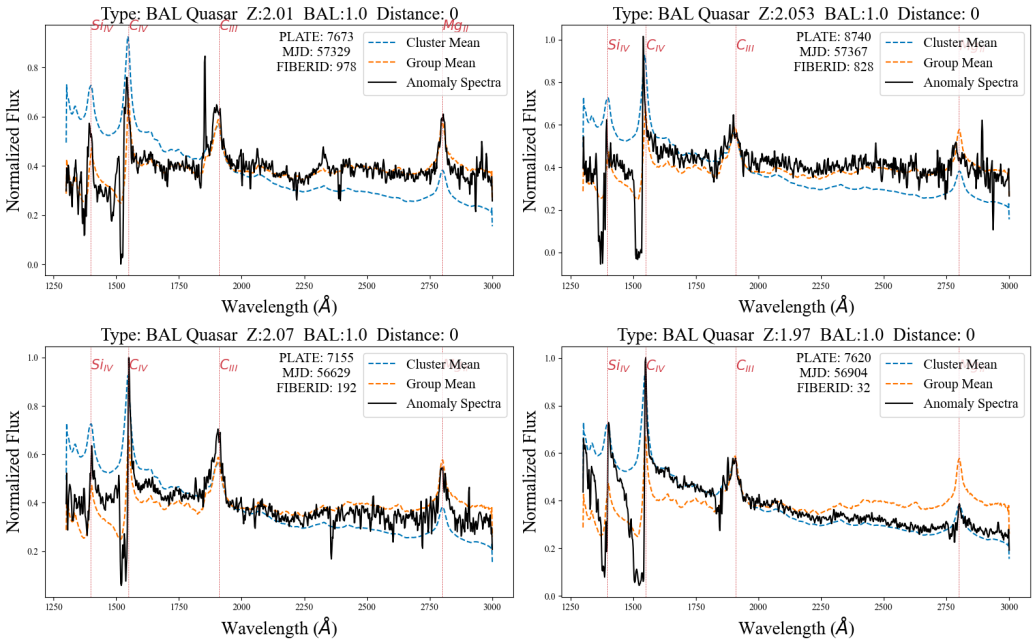


Figure 3.5: BAL Quasar Anomalies

continuum (as can be seen in the figure, the spectrum has a higher slope than the group and cluster mean), and absorption lines caused by the intervening dust and gas clouds present in the dust torus as described in section 1.5.4.

3.1.5 True Anomalies

The true anomaly group contains a mixture of quasars with varying properties with degrees of weirdness. Contrary to the other groups, they do not have a common feature that sets them apart from the normal quasars. Hence, in order to understand these, we need to look at each of them individually. The discussion for all 77 of the true anomalies is beyond the scope of this text, hence we will focus on four quasars shown in Figure 3.6. The first quasar on the top left looks nothing like a typical quasar spectral plot. There are no familiar emission lines, but only an extremely broadened MgII emission line. The spectral index is also reversed for this case being nearly opposite to that of a normal quasar. This spectrum also exhibits absorption lines and a significantly shallow FeII absorption basin. All these factors contribute to

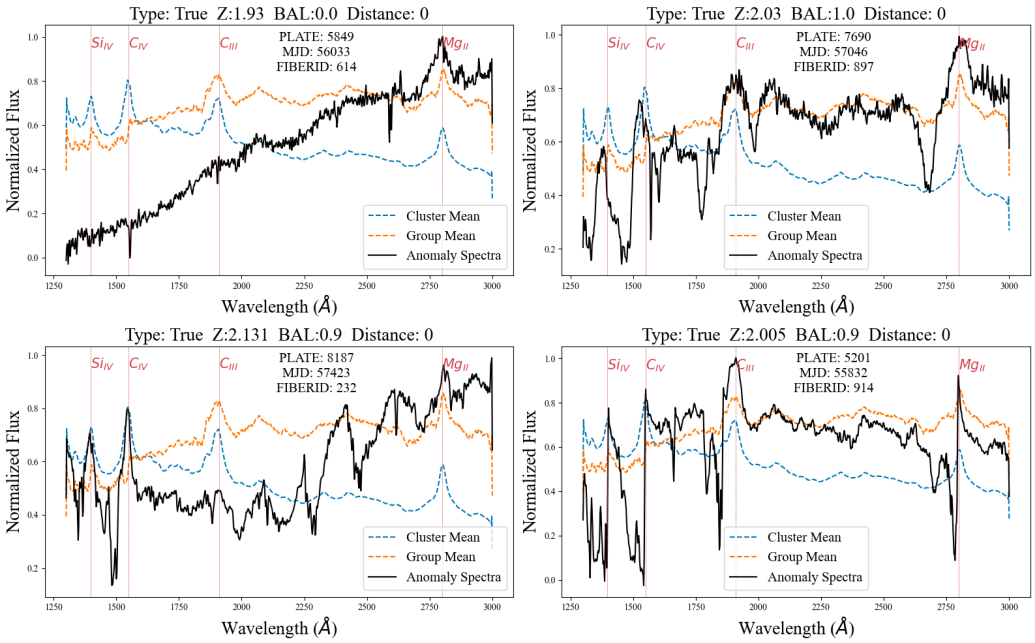


Figure 3.6: True Anomalies

identify this particular quasar as a **Heavily Reddened Quasar**. These are quasars, where most of the high energy radiations (left edge) are absorbed by the intervening dust and then re-emitted at longer wavelengths hence increasing the flux at those places leading to a positively sloping curve. MgII being extremely broadened points to a origin in the BLR clouds of the quasar which have extremely high virial velocities which impart huge Doppler broadening to the emission lines.

The third quasar at the lower left of the figure exhibits a truly unique feature not seen in any other sample we analysed. The spectral profile for this spectrum is parabolic in nature, featuring absorption in the intermediate wavelengths and higher emission in the wavelength extremes with a bias towards the higher wavelength region. This, feature is supposed to be caused by the eating away of the continuum flux by the closely spaced energy levels of the iron which mimic dust absorption and hence impart a characteristic spectra profile inversion to the quasar.

The third type of quasar present in the figure are the two images in the second column. These contain broad absorption lines near the CIV region as well

as prominent absorption at MgII emission line. They also exhibit absorption features and continuum absorption throughout the spectrum caused by FeII and FeIII. These features together make them an even rare type of BAL quasars, known as **FeLoBAL** quasars. They have been known to be notoriously tough to catch using statistical analysis because of the widespread feature profile in doing which the normal methods such as PCA and Variational Auto-encoders (VAEs) fail.
Thus a major difference between the BAL Anomaly group and the True Anomaly group is that the BAL group contains all the LoBAL quasars where as the true group is selective for the FeLoBAL. This can be attributed to specialized eigenvectors that capture high end and low end ionization in the spectrum. This difference was apparent in the PCA coefficient space hence when clustered, these two different kinds of quasars separated.

This concludes the basic understanding of the results obtained from out analysis and brief discussion on the possible causes of these anomalies. This analysis in no way complete, and would take several hours of analysis and additional literature review and data verification to conclude. The current conclusion of the study so far and possible future scope is discussed in the next chapter.

Chapter 4

Conclusion

The studies and analysis performed on the SDSS DR16 Quasar Dataset as described and discussed in the previous chapters can be condensed into the following conclusions.

4.1 Data Products

- **Data Selection:** As seen throughout the project, the emission lines in the spectrum play a major role in helping us determine the physical processes that distinguish the different quasars. Hence, we can conclude that the selection of a small redshift window was necessary as it allowed us to selectively place four of the most prominent emission lines present in the quasars for our observation, which served as the basis of differentiation. The dual clustering performed by conducting the K-means clustering twice played a major role in creating the final groups of anomalies that are presented in the results section.

- **Anomaly Grouping:** When the K-Means clustering was performed initially, it created clusters in the quasar population on the basis of their relative luminosity differences of various emission lines, as was evident in the composite spectra. The difference between these clusters was not very apparent to our human eyes as the clusters were made in a 30 dimensional PCA coefficient hyperspace. But, when the clustering

was done for the second time for the outliers of each individual cluster, we observed that the new groups created this time were much more compact, and clearly distinguishable by human eyes. This was also evident in the mean group spectra, which exhibited extremely different characteristics. Hence we can say that our algorithm is able to group the quasars based on their spectrum observing particularly significant properties that then translate into physical processes happening inside the quasar.

- **Anomaly Groups:** As seen in the result sections, there are five prominent groups of anomalies namely a) Machine error anomalies b) Excessive SiIV Emitting anomalies c) The Sharp CIV peaking anomalies d) The two classes of BAL quasars (LoBAL and FeLoBALs) and e) The true anomalies, which contains an amalgamation of disconnected bizarre properties. We have discussed the probable cause for the observed prominent features that sets the quasars apart from rest of the normal ones, such as excessive emission, concurrent absorption etc. But the exact physical reasoning is yet to be determined. This process requires rigorous data analysis, literature survey as well as multi-messenger astronomy through several epochs of observations. Only after this paradigm, we will be able to find ourselves in a position to conclusively determine the exact physical process that led to the anomalous spectrum. This right now is beyond the scope of this project in the stipulated timeline.

- **Rare BAL Quasars:** We also observed that algorithm is able to distinguish between the types of BAL quasars and is also brilliantly, capturing few of the rarest types of quasars i.e. are the LoBAL and FeLoBAL quasars. Several attempts have been made to identify and classify these quasars statistically in the SDSS catalogue, but because of the widespread or broadband features of the spectra of these, the success rate has been low. [10]. This came out to be an additional perk of our algorithm as it is able to place two of the subs features of the BAL quasars into two separate separate groups which became

apparent when we plotted their spectra.

4.2 Completeness Check

In order to determine if our algorithm and methodology is able to capture all the anomalous spectra belonging to a group or not, we perform a completeness check. For this we choose the CIV Peaker anomalies and calculate the equivalent line width ratios, CIV/CIII and CIV/MgII for the complete dataset and the detected anomalies. Since the special property of these anomalies is the extremely luminous CIV emission lines, both the ratios mentioned above must have the highest values in the entire dataset. Hence,

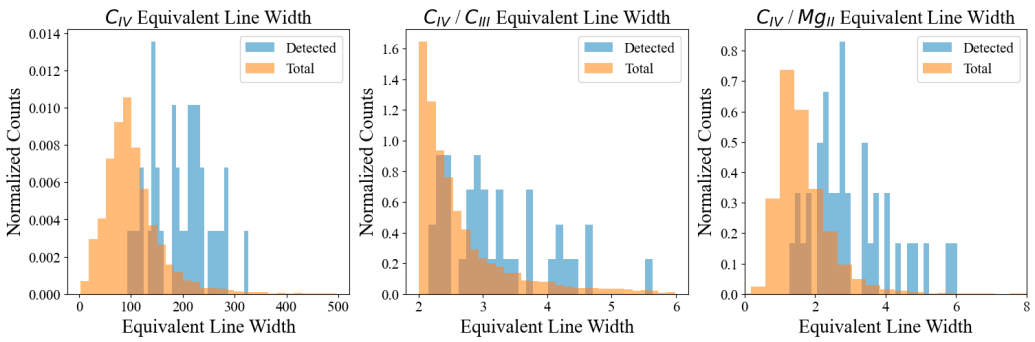


Figure 4.1: Equivalent Line Width Ratios

as seen in Figure 4.1, when plotted against the histogram for the complete dataset, the ratio values of the anomalies lies in the higher tail regions. Out 51 total CIV peakers, as determined by the tail of the histogram, we were able to successfully capture 48 anomalies rendering $\sim 94\%$ collection. There can be several such indicators that can help us determine the completeness check or in other words if our algorithm is capturing all the necessary information and not missing out on crucial datapoints.

This reinforces that the analysis and algorithm presented in this project is able to select nearly all the targeted quasars and immaculately groups them into clusters based on similar properties which can be easily identified by human help.

4.2.1 Future Scope

- We plan to use data from various surveys along with multi-epoch SDSS data to find and understand the exact physical process occurring in those distant quasars that made them to behave weirdly. In the case of unavailability of data, observation proposals would be placed at relevant advanced observatories such as HST and JWST.

- We plan to create a catalog of the unique objects found and present it for further investigation by the astrophysics community. A list of corrupted spectra would also be delivered to the SDSS team for redacting or removal.

- Finally, we would also wish to create semantically sound models based on the population of different quasars we find in the anomalies. This would involve our analysis expanding to most of the available redshift values in small window steps.

Bibliography

[1] Robert Antonucci. Unified models for active galactic nuclei and quasars. *Annual Review of Astronomy and Astrophysics*, 31:473–521, 1993.

[2] Roger D. Blandford and Roman L. Znajek. Electromagnetic extraction of energy from kerr black holes. *Monthly Notices of the Royal Astronomical Society*, 179:433–456, 1977.

[3] A. C. Carnall. SpectRes: A Fast Spectral Resampling Tool in Python. *arXiv e-prints*, page arXiv:1705.05165, May 2017.

[4] C. D. Dermer. High-energy gamma-ray emission from active galactic nuclei: Production and absorption of gamma rays in massive jets. , 400:L57–L60, December 1992.

[5] Moshe Elitzur. The physics and diagnostics of active galactic nuclei. *Monthly Notices of the Royal Astronomical Society*, 421:L112–L116, 2012.

[6] Bernard L Fanaroff and Julia M Riley. The morphology of extragalactic radio sources of high and low luminosity. *Monthly Notices of the Royal Astronomical Society*, 167:31P–36P, 1974.

[7] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21(3):768–769, 1965.

[8] Paul J. Francis, Paul C. Hewett, Craig B. Foltz, Frederic H. Chaffee, Ray J. Weymann, and Simon L. Morris. A High Signal-to-Noise Ratio Composite Quasar Spectrum. , 373:465, June 1991.

[9] Andrew D Goulding and David M Alexander. Are there any starbursts in nearby seyfert 2 galaxies? *Monthly Notices of the Royal Astronomical Society*, 398:1165–1173, 2009.

[10] Zhiyuan Guo and Paul Martini. Classification of broad absorption line quasars with a convolutional neural network. *The Astrophysical Journal*, 879(2):72, jul 2019.

[11] Fred Hamann and Gary Ferland. The Age and Chemical Evolution of High-Redshift QSOs. , 391:L53, June 1992.

[12] Timothy M Heckman. The nature of liners. *Astronomy and Astrophysics*, 87:152–162, 1980.

[13] Walter Heitler. The quantum theory of radiation. *Clarendon Press*, 1954.

[14] Brad W. Lyke, Alexandra N. Higley, J. N. McLane, Danielle P. Schurhammer, Adam D. Myers, Ashley J. Ross, Kyle Dawson, Solène Chabanier, Paul Martini, Nicolás G. Busca, Hélion du Mas des Bourboux, Mara Salvato, Alina Streblyanska, Pauline Zarrouk, Etienne Burtin, Scott F. Anderson, Julian Bautista, Dmitry Bizyaev, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Johan Comparat, Paul Green, Axel de la Macorra, Andrea Muñoz Gutiérrez, Jiamin Hou, Jeffrey A. Newman, Nathalie Palanque-Delabrouille, Isabelle Pâris, Will J. Percival, Patrick Petitjean, James Rich, Graziano Rossi, Donald P. Schneider, Alexander Smith, M. Vivek, and Benjamin Alan Weaver. The sloan digital sky survey quasar catalog: Sixteenth data release. *The Astrophysical Journal Supplement Series*, 250(1):8, August 2020.

[15] David L. Meier. Relativistic jets from active galactic nuclei. *International Journal of Modern Physics D*, 20:2201–2264, 2012.

[16] Dimitri Mihalas and Barbara Weibel Mihalas. *Foundations of Radiation Hydrodynamics*. Oxford University Press, 1978.

[17] Donald E Osterbrock and Gary J Ferland. Astrophysics of gaseous nebulae and active galactic nuclei. *University Science Books*, 2006.

[18] Bradley M. Peterson. *Taxonomy of Active Galactic Nuclei*, page 21–31. Cambridge University Press, 1997.

[19] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[20] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, January 1964.

[21] Maarten Schmidt. 3c 273: A star-like object with large red-shift. *Nature*, 197:1040–1040, 1963.

[22] M Scourfield, A Saintonge, D de Mijolla, and S Viti. De-noising of galaxy optical spectra with autoencoders. *Monthly Notices of the Royal Astronomical Society*, 526(2):3037–3050, September 2023.

[23] Carl K Seyfert. Nuclear emission in spiral nebulae. *Astrophysical Journal*, 97:28, 1943.

[24] Pulkit Sharma. The ultimate guide to k-means clustering: Definition, methods and applications, Feb 2024.

[25] G. A. Shields. Broad emission lines in seyfert galaxies and active galactic nuclei. *Nature*, 272:706–707, 1978.

[26] Matthew J Temple, Gary J Ferland, Amy L Rankine, Marios Chatzikos, and Paul C Hewett. High-ionization emission-line ratios from quasar broad-line regions: metallicity or density? *Monthly Notices of the Royal Astronomical Society*, 505(3):3247–3259, June 2021.

[27] Robert L Thorndike. Who belongs in the family? In *Psychometrica*, volume 18, pages 267–276. Springer, 1953.

[28] C Megan Urry and Paolo Padovani. Blazars in synchrotron inverse compton models. *Publications of the Astronomical Society of the Pacific*, 107:803–845, 1995.

[29] Y. Yousef and S. Davis. Testing the Standard Model of AGN Accretion Disks. In *American Astronomical Society Meeting Abstracts #235*, volume 235 of *American Astronomical Society Meeting Abstracts*, page 369.09, January 2020.

[30] Andrzej A Zdziarski. Pair production and annihilation in black hole accretion flows. *Monthly Notices of the Royal Astronomical Society*, 423:663–675, 2012.

[31] W. Zheng, G. A. Kriss, R. C. Telfer, and J. P. Grimes. The ultraviolet continuum emission in active galactic nuclei: The role of accretion disks. *The Astrophysical Journal*, 475:469–475, 1997.