Training Report

A new selection method of Double Peak emission galaxies

Prepared by:

Youssef Temmam

2nd year Master : Theoretical Physics
Aix-Marseille Université

Supervised by:

Pr. **Amram Philippe**

Laboratoire d'Astrophysique de Marseille (LAM)

Academic Year:
2022/2023

Abstract

Double Peak galaxies were studied for the first time among the population of Active Galactic Nucleus galaxies. This study was later expanded to include all types of galaxies, leading to a new perspective on galaxy evolution. These galaxies have been associated with galaxy mergers, which is one of the most extensively studied topics in astrophysics.

Traditionally, the identification of Double Peak galaxies involves fitting double Gaussian profiles. However, in this research project, we aim to explore a promising method that relies on measuring statistical quantities (Skewness and Kurtosis) to describe the shape of emission lines, eliminating the need for additional Gaussian fits.

To begin, we will create a list of Double Peak galaxy candidates based on the **SDSS** and **RCSED** catalogs. We will then apply **AMAZED**, a redshift calculating software currently being developed by the Laboratoire d'Astrophysique de Marseille. This will also serve as an opportunity to test and document the software.

Additionally, we will explore an alternative approach to fitting emission lines by utilizing an AI-based library called GaussPy. We will develop a program capable of accurately fitting emission lines. Finally, we will analyze the Double Peak galaxy candidates using Kurtosis and Skewness measurements in different emission lines.

Contents

Chapter 1

Introduction

1.1 Scientific context

On a moonless, clear night in the heart of summer, in a remote region far from cities in the northern hemisphere, you gaze up at the sky and see a hazy band of light stretching vertically. This is the *Milky Way*, our home galaxy. As old as the universe itself, the Milky Way is not alone. There are more than 100 billion galaxies like ours in the universe. The closest neighbor to the Milky Way is *Andromeda*, a spiral galaxy that is nearly twice as large. Andromeda and the Milky Way are on a collision course and are expected to merge in the distant future. But this won't be the first merger for the Milky Way. Our galaxy is part of the *Local Group*, a collection of more than 80 galaxies that interact gravitationally with each other, leading to future galaxy mergers.

The Local Group is also part of the *Virgo Super-cluster* which is itself a component of the *Laniakea Super-cluster*. The Laniakea Supercluster contains over 100000 galaxies [1] at the size of $150\,Mpc$. Unlike its subgroups, The Laniakea Super-cluster is not bound by gravity and it is expected to disperse [2]. However, at smaller scales (around $\sim 3\,Mpc$ which represents the diameter of the Local Group), galaxies are bound by gravitation and they are expected to merge into one big elliptical in about a dozen Gyrs [3]. Yet, this does not mean that there haven't been any mergers in the past. In fact, considering only our galaxy alone, it has merged many times in the past including with *Gaia Sausage* [4] around $10\,Gyr$ ago and is merging currently with some dwarf galaxies (*Sagittarius*, *Canis Major*, *Large* and *Small Magellanic Clouds*).

Galaxy mergers are the most violent and common type of galaxy interactions and they have many consequences for galaxies. The gravitational interactions and the friction between matter affect galaxies on both macroscopic and mesoscopic scales. On the macroscopic scale, galaxy mergers have a significant impact on the morphology of galaxies and their type too [5]. Major mergers, depending on the mass ratio of the merged galaxies, can alter the global shape, galactic halos, and lead to the appearance of Active Galactic Nuclei. For instance, they can transform a spiral galaxy into an elliptical galaxy. Whereas minor mergers often change the structure of the galaxy

such as galactic arms, disks thickening and heating, bulge and internal bars, dark matter distribution...

On the mesoscopic level, galaxy mergers introduce more matter to the galaxy providing them with more fuel to form stars. In minor mergers, the Star Formation Rate (SFR) increases and can even change the galaxy sequence and make it more blue [6] thanks to gas accretion. Galaxy mergers can also change the dynamics of matter in galaxies, increasing its velocity and radiation, and therefore altering galaxies spectra. In some particular cases, galaxies have a particular spectrum with double emission lines, which can be a sign of galaxy mergers.

Early, galaxy mergers have been identified through observations of galaxy mythologies or dynamics. In a study by [7], 1716 dynamically close merging galaxies have been counted in the **SDSS** (Sloan Digital Sky Survey). This number was further extended by including Active Galactic Nucleus (AGN) where astronomers associated galaxy mergers with Double Peaked (DP) galaxies [8].

1.2 Project goal

Until today, many questions remain regarding DP emissions. Answering these questions could provide valuable insights into galaxy evolution and mergers, which is one of the most researched topics in extragalactic astrophysics. In the literature, only a few methods have been explored to detect DP emissions. Maschmann et al [9] developed a program based on double Gaussian fits for emission lines to distinguish DP galaxies. In this research project, we will investigate a different promising method to identify these galaxies.

Our approach focuses on a straightforward method to detect DP emissions. Specifically, we measure statistical quantities (Skewness and Kurtosis) that characterize the shape of emission lines, without the need for additional Gaussian fits. Firstly, we measure these quantities using the results of the **AMAZED** emission line fits. **AMAZED** is a software currently being developed by the Laboratoire d'Astrophysique de Marseille for use with Euclid mission and also the Prime Focus Spectrograph of the Subaru telescope. Secondly, we will explore an alternative to **AMAZED** for emission line fits. We will utilize an AI-based library called **GaussPy**[10] to create a program capable of fitting emission lines, from which we will compute Skewness and Kurtosis.

In the second chapter, we will introduce the SDSS and RCSED galaxy catalogues that we will be using throughout this project. The third chapter will provide a brief review of DP galaxies and discuss previous work done in the field. In the fourth chapter, we will introduce both **AMAZED** and **GaussPy**, the tools we used to select DP galaxies, and discuss their advantages and limitations. The fifth chapter will explain the DP galaxy selection method and the underlying concepts. Finally, in the sixth chapter, we will discuss the results and draw conclusions.

Additionally, we have taken this opportunity to create complementary documentation for the laboratory cluster, a comprehensive documentation for **AMAZED**, and a comment on the **GaussPy** library, all of which have been sent to the developers.

1.3 Hosting laboratory

Laboratoire d'Astrophysique de Marseille (LAM) is a research laboratory specializing in astrophysics and related instruments. It is affiliated to CNRS, CNES, and Aix-Marseille University. The laboratory has established partnerships with organisations as ESA, NASA, ESO... The LAM has also contributed to many space experiments like Euclid, Herschel, Rosetta... and developed some instruments for those satellite telescopes, as well as ground telescopes like VLT.

With its members over 200, LAM is operating in diverse range fields of astrophysics:

Cosmology, galaxies formation and evolution, stars and interstellar medium,

Planetary systems formation and the solar system,

Optical instruments,

which are operated respectively the three research groups: GECO (Galaxies, Étoiles et COsmologie), GRD (Groupe Recherche et Développement) and GSP (Groupe systèmes planétaires) within the LAM facility.

Chapter 2

SDSS and RCSED

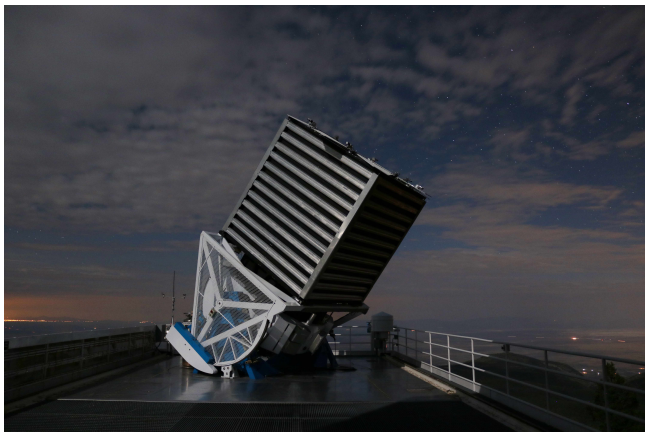
2.1 Presentation

In this project, I will use the Reference Catalog of galaxy Spectral Energy Distribution (RCSED) [11]. This catalog counts 800299 galaxies gathered by 3 telescopes (cf figure 2.1): the Sloan Digital Sky Survey (SDSS) data release 7, which is multi-spectral imaging and spectroscopic redshift survey, UK Infrared Telescope Deep Sky Survey (UKIDSS) and the Galaxy Evolution Explorer (GALEX). All the spectra of this catalog had been analysed, fitted and k-corrected in the ultraviolet, optical and infrared. Moreover, it provides 2 fits and their corresponding χ^2 for all emission lines: Gaussian fit and non-parametric fit. This latter is a fit with an arbitrary function that is more suited for DP emission lines since it is the combination of 2 Gaussians.

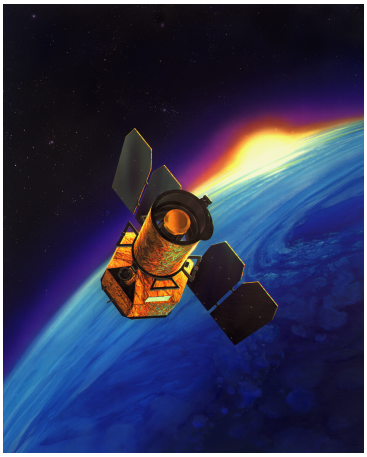
2.2 Spectra identification

Each spectrum is identified by a unique ID in the catalog. This ID contains information about the observation plate, the MJD (Modified Julian Date), and the observation fiber on the telescope. For instance, the spectrum "$spec-1311-52765-0254.fits$" is observed through the plate number 1311 on the day 52765 by the fiber 0254.

On the cluster, spectra are grouped in folders by their plates number.



(a) The SDSS telescope at night, Image Credit: Patrick Gaulme.



(b) Artist's impression of GALEX spacecraft, Credits: NASA/JPL-Caltech.

(c) The United Kingdom Infrared Telescope, Credit: Paul Hirst.

Figure 2.1: RCSED data sources telescopes.

Chapter 3

Double peak emission galaxies

3.1 Context of Double peak emission galaxies

In spectroscopy, we can identify different types of lines:

- absorption lines are the present pockets in spectra. They originate from the transition of electrons between energy levels in atoms or ions and they represent wavelengths of light are absorbed by those electrons to transition from a lower energy level to a higher one.
- Whilst emission lines are release of energy in the form of photons of specific wavelengths. They are also a form of electrons transition but from a higher energy level to a lower one.

Both of those types of lines provide valuable information about galaxies: like mass (Tully–Fisher Relation that we will see later), metallicity and age, interstellar medium and more generally about the evolution of galaxies [12] . In our project, I will focus only on the second type of lines.

At first order, those emission lines can be modelled by Gaussian functions. Yet some galaxies show some peculiar emission lines that cannot be precisely modelled by a Gaussian, but by 2. Those galaxies were called Double Peak galaxies. Figure 3.1, displays an example of the DP emission line fitted by a single and the composition of 2 Gaussians.

In literature, DP galaxies were first studied in the context of AGNs and specifically quasars. It was until 2008 when it was linked to galaxy mergers [13] namely for dual AGN galaxies. Maschmann et al [9] extended this link to all type DP galaxies at redshifts $z < 0.34$. They have shown that 7% of them have dual AGNs or more, 10% morphologically confirmed merging galaxies, where the rest includes galaxies with significant velocity dispersion galaxies with very large bulge, and S0 galaxies that are formerly merged galaxies.

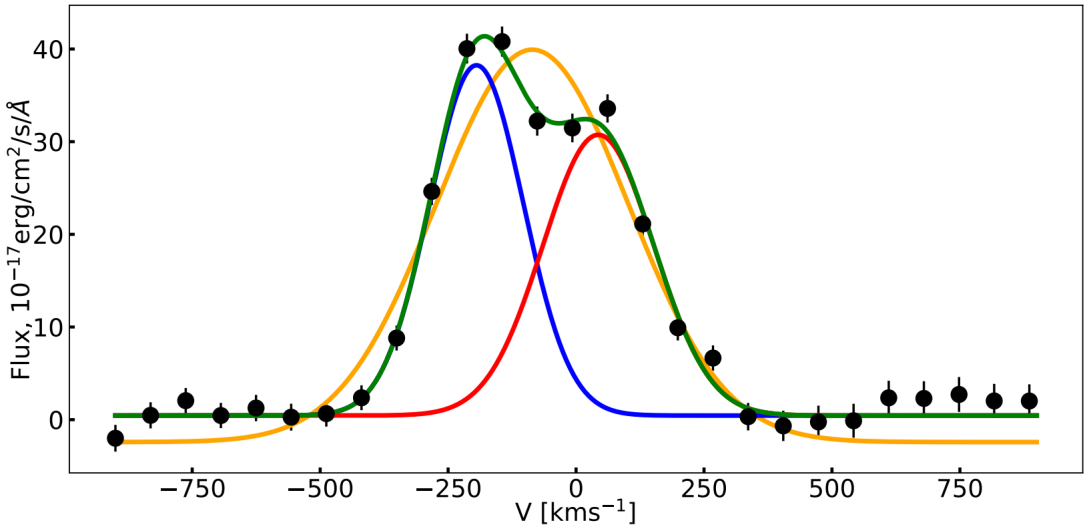


Figure 3.1: An example of DP emission line. Black dots represent the spectrum, the yellow curve is the one Gaussian fit, and the green curve is the composed fit of the 2 Gaussians red and blue. Image Credit: Maschmann et al [9].

3.2 Characteristics of Double Peak emission galaxies

The authors of [9] made an outstanding study of DP galaxies. They have studied all their aspects from morphology, emissions, galaxy classification, stellar velocity, environmental characteristics... in order to better understand the origin of DP galaxies and possibly uncover some correlations between those galaxies and the aforementioned aspects. We shall see each of those aspects in following. In their study [9], they created in the beginning 2 pools of galaxies (using the algorithm described in figure 2 of [9], we shall see it thoroughly in section 5.1) : Double Peak galaxies Sample (DPS) as their name suggests, and Control Sample (CS) consisting of the single peak galaxies.

By comparing stellar mass and redshifts of those pools, it turned out that the stellar population of CS is younger with less stellar mass ratio due to selection criteria. Therefore, this CS cannot serve its purpose to compare between those two populations and reveal DP characteristics. Consequently, the authors derived from the CS a pool having the same age and stellar to mass ration as DPS and they called it Non Biased Control Sample (NBCS).

3.2.1 Galaxy population classification

The NBCS contains 55% of Star Forming galaxies (SF), 24% of COMP galaxies (Composite galaxies exhibiting the same properties as AGN and SFs) with 13% classified as AGN or LINER (Low-Ionization Nuclear Emission-line Region) galaxies. Even though DPS contains fewer SF galaxies (45%) with a similar AGN fraction (11%)

and a lower fraction of LINER galaxies in the DPS (3% vs. 6% in the NBCS), radio observations suggest a significant excess of radio sources in the DPS compared to the NBCS, indicating increased AGN excitation and SF. This is due to COMP galaxies that represents 45% of DPS, nearly twice the fraction observed in the NBCS (24%). Also, the DPS does not appear to be dominated by double-peak emission-line AGNs, as only about 7% are classified as double AGN or LINER.

3.2.2 Morphological Classification

The authors use a machine-learning-based algorithm to classify galaxies into different morphological types, including disc features, S0 galaxies, edge-on orientation, visual mergers, and bar structures. In the DPS and the NBCS, a similar percentage of galaxies remains unclassified (35% and 38%, respectively). The DPS contains a lower fraction of LTG (Late-Type Galaxy) galaxies (16%) compared to the NBCS (30%), while more S0 galaxies are present in the DPS (36%) than in the NBCS (20%). Both samples have a similar merger rate of approximately 10%, suggesting that DP galaxies might be minor mergers, post-mergers, or hidden mergers. Visual inspection confirms the merger rate and consistency with the classification results. There is no significant excess of edge-on galaxies in the DPS compared to the NBCS.

3.2.3 Emission characteristics

Regarding emission lines, Hα luminosity differences between DPS and NBCS are small but statistically significant, with DPS galaxies having slightly higher luminosities. Balmer decrement-based extinction is slightly higher in DPS galaxies than in NBCS galaxies, particularly for LTG and S0 galaxies. Both NBCS and DPS galaxies have, on average, older populations, with a majority being younger than 5 Gyr and exhibiting significant SF activity. This latter is prevalent in all three samples (DPS, CS, NBCS), but AGN activities (or galaxies classified as COMP) are also present in the NBCS and DPS.

3.2.4 Stellar velocity dispersion and environmental characteristics

The DPS exhibits a higher stellar velocity dispersion σ compared to the NBCS. This could indicate that the ionized gas in DP galaxies is either strongly perturbed or has an external origin, providing insights into the mechanisms responsible for double-peaked emission lines. The gas velocity dispersion (σ_{gas} is higher for both NBCS and DPS aligning with the Tully-Fisher relation and indicating more massive galaxies. In the NBCS, Gas velocity dispersion (σgas) and galaxy inclination angle correlate, supporting the rotating disc scenario. This correlation is absent in DPS galaxies, suggesting other mechanisms for DP emission line shapes. Regarding galaxies population, the TFR studies of DPS and NBCS shows good agreement for elliptical galaxies in all pools. However, S0 galaxies in the NBCS exhibit systematically lower σ values, while

the DPS shows velocities consistent with those of elliptical galaxies with the same stellar mass. This supports the notion that the DPS galaxies may represent a larger system composed of the superposition of two systems.

Environmental analysis shows no significant connection between galaxy environment and morphological type. S0 galaxies in both DPS and NBCS are mainly located in less dense environments, contradicting the idea that spirals evolve into S0s due to their environment. Sérsic index suggests larger bulges in LTGs of the DPS compared to NBCS, indicating potential differences in merger rates, especially indicating a merging scenario for DP galaxies. However, it is also noted that DP galaxies might be dominated by post-coalescence mergers, which can escape detection in analyses targeting ongoing mergers. Also, minor merger rates are slightly higher for DPS galaxies, especially LTGs and S0 galaxies.

3.3 Summary

To sum up, no significant differences are found between the DPS and NBCS in terms of merger-related characteristics, suggesting major mergers may not be the dominant mechanism for DPS galaxies. They may result from sequential gas-rich minor mergers or gas accretion, explaining observed characteristics, including the presence of S0 galaxies. Other observations support that S0 galaxies can have complex kinematics and are not strictly in the red sequence, consistent with ongoing or recent interactions and star formation. Alternative explanations, such as AGN-driven outflows, apply to a small fraction of DPS galaxies.

Chapter 4

Presenting Softwares

4.1 AMAZED

4.1.1 Introduction

The LAM contributed to the development of both hardware and software of the Prime Focus Spectrograph (PFS) module that would be attached to Subaru telescope 4.1. This module comes with a fiber positioner that point each fiber of 1 arcsec diameter on a single galaxy or star. The fiber positioner enables to take exposures of 2,400 astronomical objects simultaneously. The light from stars and galaxies are dispersed and recorded as spectra simultaneously covering a wide range of wavelengths ranging from the near-ultraviolet, through the visible, and up to the near-infrared regime, with a spectral power ranging from 2500 to 4500 from 0.38 to 1.30 μm.

Figure 4.1: The Subaru telescope, Image Credit: Patrick Gaulme.

On the software aspect, the CESAM team is developing **AMAZED** program and I mainly worked with them in this part of the internship because **AMAZED** is still

under development and I worked with them as a software tester where I contributed to the identification of many issues. I also made a documentation for this program (section A.2), and a complementary documentation for the LAM cluster.

The main goal of **AMAZED** is to determine redshift of galaxies, quasars... from their spectra. Nevertheless, it can have other usages and applications (like studying line emissions of galaxies and their correlation with galactic morphology, history...) since redshifts are calculated using emission lines fits. In a few words, the program run through 2 steps:

- in the first step, a global fit is applied to all galaxy spectrum with least square fitting and calculates roughly the redshift.
- In the second step, knowing approximately the wavelength localisation of each line, it fits its continuum with second degree polynomial, then applies a Gaussian fit to the line. Subsequently, it determines the redshift from all lines more precisely.

4.1.2 Tests of AMAZED with a SDSS sub-samples of galaxies

4.1.2.1 Test of AMAZED on the 48 brightest galaxies in Halpha from SDSS

I selected the 48 brightest galaxies in the Hα line in **SDSS**. I run **AMAZED** on them with the aim of comparing the results (namely the parameters of this emission line) with those provided by the **SDSS**. Since **AMAZED** can run up to $z = 10$ and I work in our case only galaxies with redshifts lower than $z = 1$, this comparison would enable us to test **AMAZED** for lower redshifts, and adjust fitting parameters of the file "parameters.json" aforementioned in order to obtain the optimum results namely for line fits.

Before running **AMAZED** and exploiting its results, I have to test and calibrate it for our usage. By setting the redshift range upper limit at 1, and the redshift step for the first phase and second phase at 0.001. I set the resolution at 2500 and I run the program on those galaxies. We notice that **AMAZED** has worked on only 39 galaxies out of 48 and fitted Hα line. Also, only 26 of 39 that have direct integrated flux and fitted flux difference under 10%. To have a clearer idea about the exactitude of Hα fluxes, I plot the figure 4.2. We see that most of the 39 galaxies are under the red line, meaning that the fitted flux is underestimated and major corrections need to be done to the program.

After discussing with the builder team of **AMAZED**, they released a newer version "amazed_0.44-RC3". With this version, the program still worked only on 39 out 48. The 9 galaxies that the software did not compute have null pixels. Also, I plotted the flux relative difference in figure 4.3 and we see more galaxies closer to the red line. Precisely, I obtained 36 out of the 39 galaxies with a flux relative difference for Hα line under 10%. This is clearly a significant improvement, but there still more to be done. When we look at the 3 galaxies, whose flux difference is larger than 10%, they

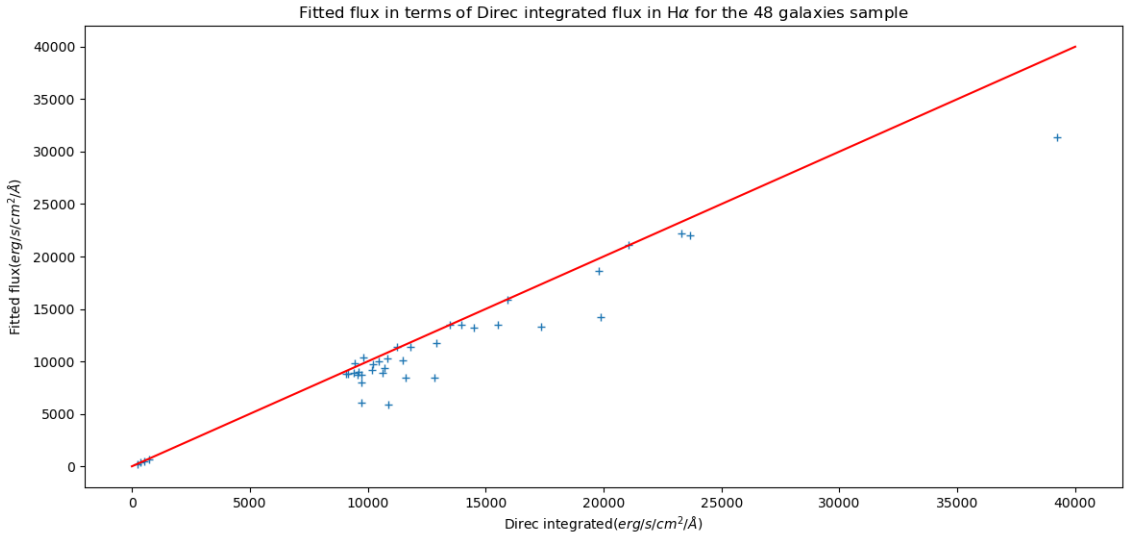


Figure 4.2: Distribution of the 48 brightest galaxies with the version "amazed_0.44-RC1". The red line represents the Fitted Flux = Direct Integrated Flux.

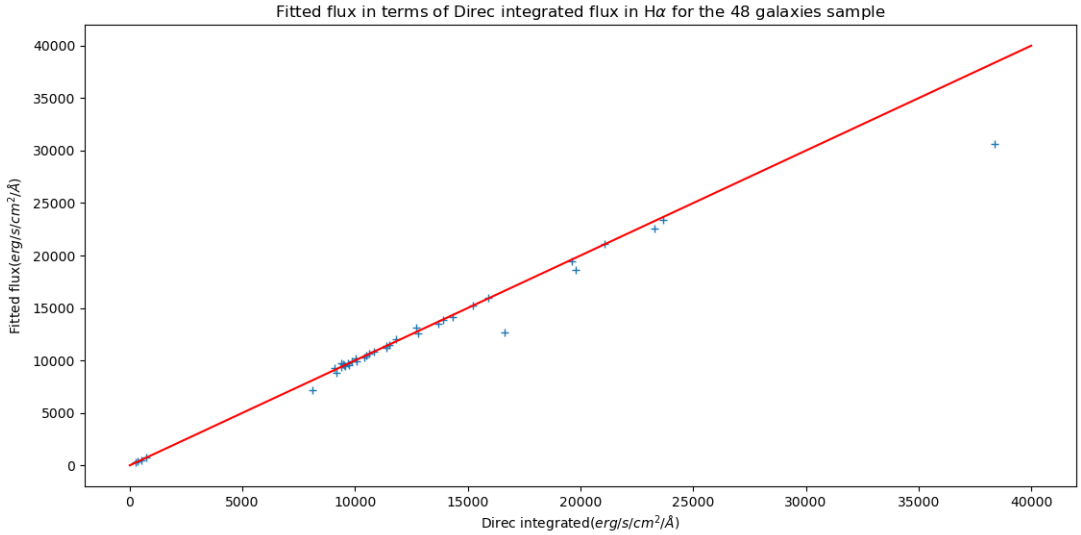


Figure 4.3: Distribution of the 48 brightest galaxies with the version "amazed_0.44-RC3". The red line represents the Fitted Flux = Direct Integrated Flux.

all have NII lines close to Hα line. Thus, when the program fits all the lines with the same velocity dispersion and offset for Hα and NII and not individually. So, the Hα fit is a little bit off-centered and can lead to a bigger difference.

Moreover, if those emission lines NII and Hα are close, the direct integrated flux would include all of them when calculating flux for Hα using direct integration method. As a result, flux relative difference would grow. Especially, this occurs when the velocity dispersion increases. To better visualize this, I plot the module

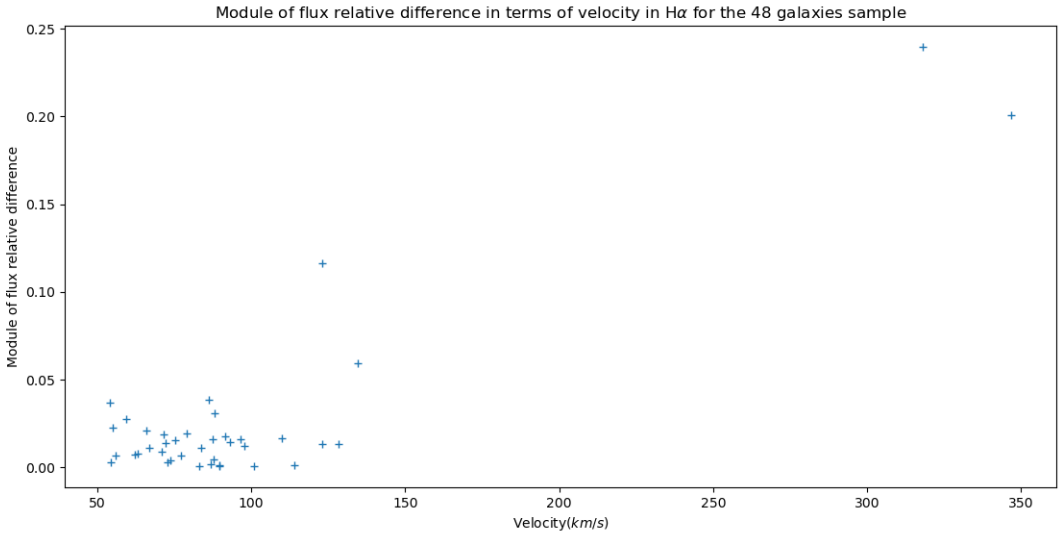


Figure 4.4: Distribution of the 48 brightest galaxies with the version "amazed_0.44-RC3" data plot.

of flux relative difference in terms of velocity in figure 4.4. We see clearly all the 3 galaxies whose flux relative difference beyond 10% have larger velocities. However, the first galaxy with a flux difference at 12% has a velocity of 136 $km.s^{-1}$, less then other galaxies with higher velocity and less flux difference. This proves the software works for galaxies with even big velocities up to 150 $km.s^{-1}$ and further updates for **AMAZED** should be done. I shall also check this bias with bigger samples of galaxies.

Now considering the Hβ emission line to see if those flux bias remains even for a separate emission without any close one. We see in figure 4.5 that the direct integrated flux is close to the fitted one. Also, we see in figure 4.6 that the relative difference between fitted and direct integrated fluxes is under 6%. This means that one of the main reasons why the relative difference was bigger for Hα is due to the presence of NII lines that are included in the direct integrated flux. We cannot either confirm or deny the impact of velocity on the relative difference, because we have in both high and low velocities some galaxies with relative difference around 5%. One path to follow is to reproduce the same figure, but for larger number of galaxies.

4.1.2.2 Reconstructing an emission line and its fit with AMAZED results

I aim to verify the fit visually, by plotting a portion of an arbitrary line emission and an arbitrary galaxy spectrum alongside its fit from **AMAZED**. To this end, I make a program that chooses randomly a galaxy out of the 48, reads the output file of **AMAZED** containing fits and recovers the fit data of the chosen galaxy. Then it reads spectrum file corresponding to that galaxy, cut the interval around the emission line at the $[\lambda - 10\sigma, \lambda + 10\sigma]$ where λ is the fitted emission line wavelength and σ its standard deviation. Then I plot the spectrum portion and its fit in figure 4.7.

Figure 4.5: Distribution of the 48 brightest galaxies with the version "amazed_0.44-RC3". The red line represents the Fitted Flux = Direct Integrated Flux.

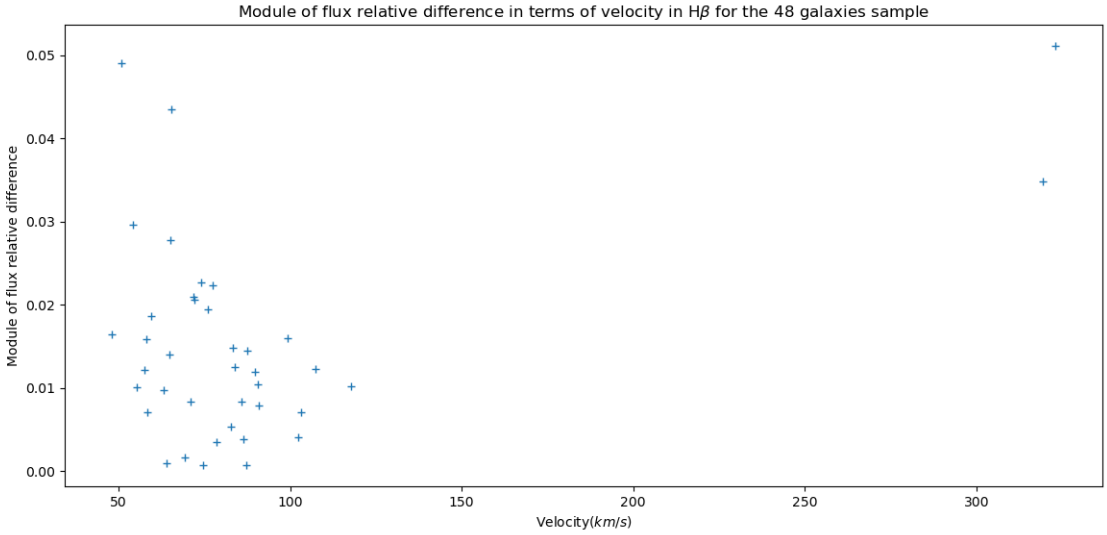


Figure 4.6: Distribution of the 48 brightest galaxies with the version "amazed_0.44-RC3".

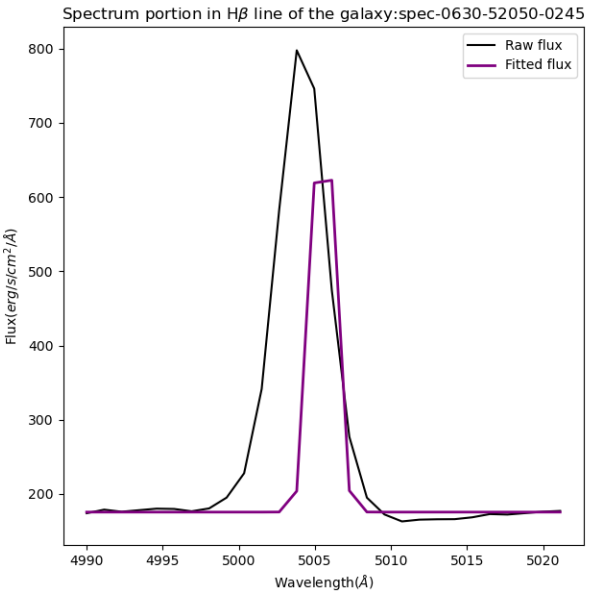


Figure 4.7: Hβ line fit of the galaxy:"spec-0630-52050-0245". Black (respectively purple) represents raw (respectively fitted) flux.

We see in this figure that there are two issues: the width and the deviation of the fit that are not compatible with the line. I recreated this plot other galaxies and all of have the same issue. So I decided to plot the fitted flux in terms of the one computed from the amplitude C.2 in the figure 4.8.

We observe that there is a factor 2 approximately between the fitted flux and the computed one. After investigation, it turned out that "LinemeasLineWidth" is indeed the standard deviation and its name is incorrect. Consequently, we do not need to "LinemeasLineWidth" by $2\sqrt{2\ln(2)}$ and figure 4.9 confirms it. I re-plot again the fit and the spectrum around Hβ line for the same galaxy "spec-0630-52050-0245" and I obtain figure 4.10.

Nevertheless, I have to solve also the issue of the fit deviation. By examining carefully the deviation for all emission lines, we see that the fit is deviated by around 1 Å. One of the possible reasons can be air to vacuum conversion. **AMAZED** program works only on vacuum spectra. The 48 galaxies spectra come originally from **SDSS**, they were converted in order to be read by **AMAZED**. When reading headers of those spectra, we remark that they are compute in air. Hence the fit deviation we observe in the previous figures.

In order to overcome this issue, I had two options: either reconvert all the spectra and those I would be working on, or building a version of **AMAZED** that can work directly on **SDSS** spectra. After discussing with developing team of **AMAZED**, we decided to go with the second option which is more interesting since it can work on **SDSS** spectra too. From that moment, I have been using a non-official release of

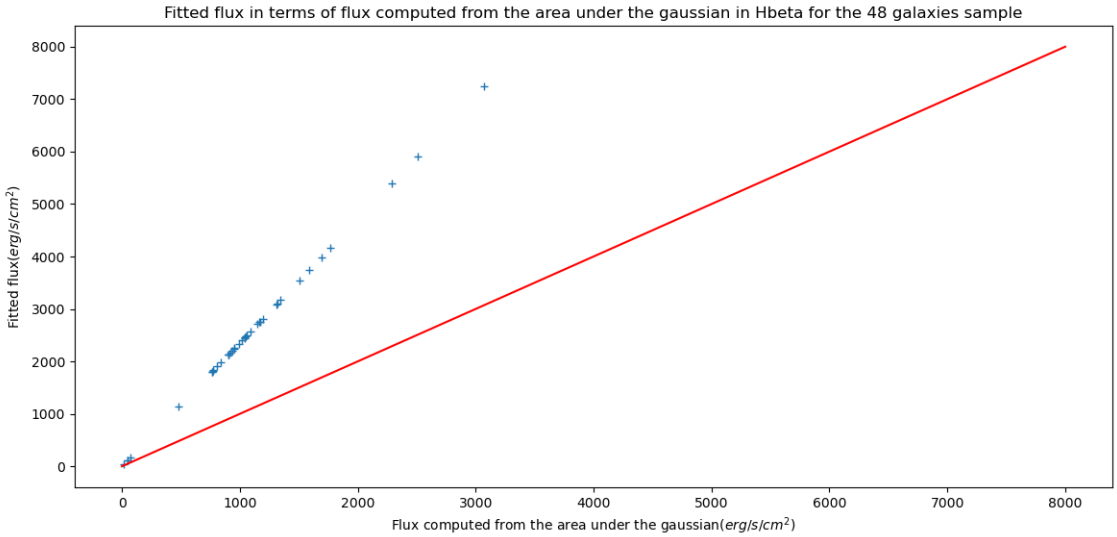


Figure 4.8: Distribution of the 48 brightest galaxies. The red line represents Fitted Flux = Computed Flux from C.2.

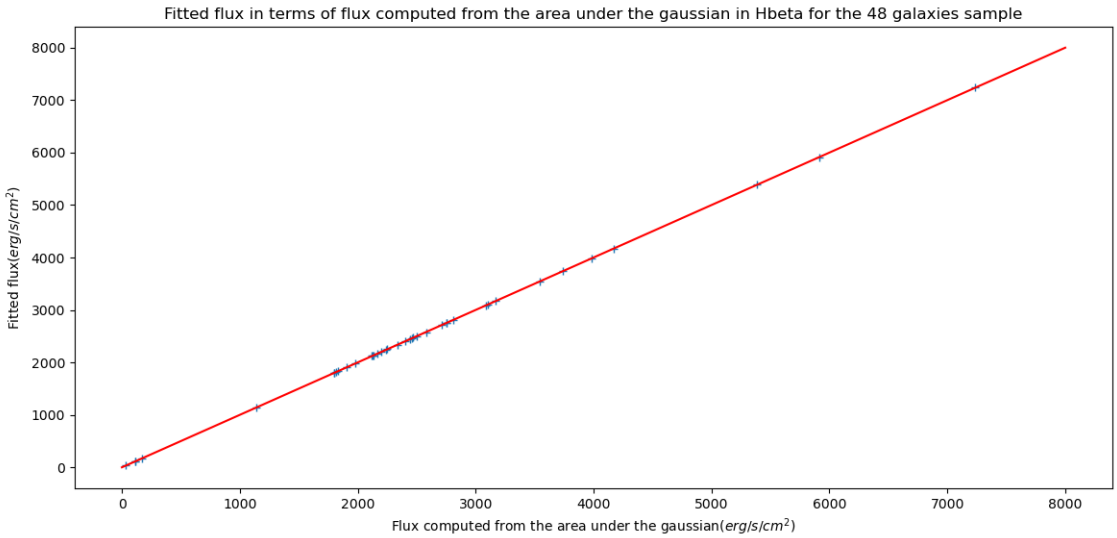


Figure 4.9: Distribution of the 48 brightest galaxies after the correction of the standard deviation. The red line represents Fitted Flux = Computed Flux from C.2.

AMAZED in the directory on the cluster:
"/net/CESAM/amazed/aallaoui/venvs/test/fix_issue_8015/40949022/" (cf A.2.2).

I re-run **AMAZED** again, on the same 48 galaxies, but on **SDSS** spectra. I check all fits for all galaxies and I obtain figure 4.11 for Hβ line of "spec-0630-52050-0245".

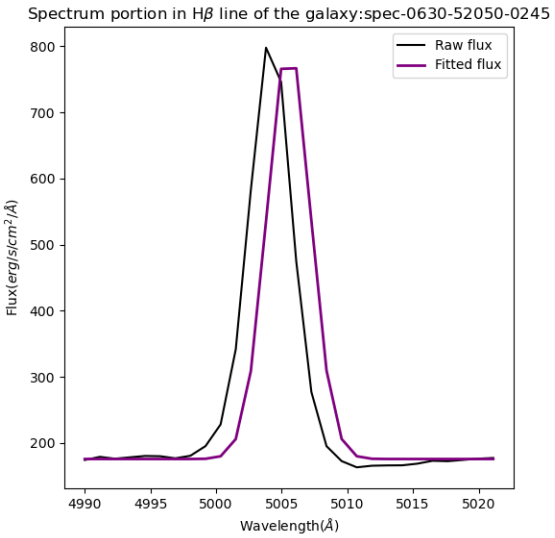


Figure 4.10: Hβ line fit of the galaxy:"spec-0630-52050-0245" after the correction of the standard deviation. Black (respectively purple) represents raw (respectively fitted) flux.

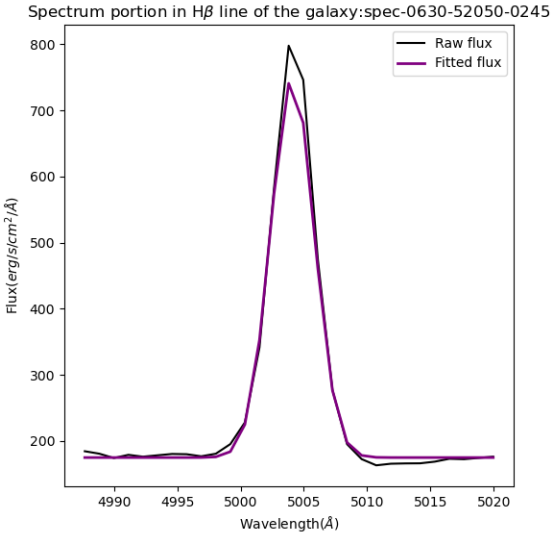


Figure 4.11: Hβ line fit of the galaxy:"spec-0630-52050-0245" using the non converted **SDSS** spectrum. Black (respectively purple) represents raw (respectively fitted) flux.

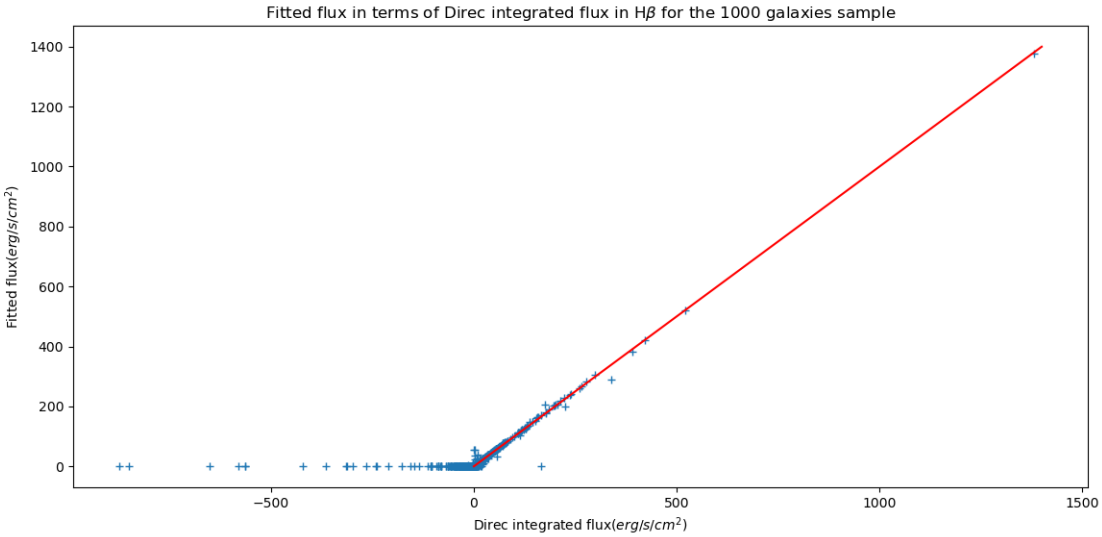


Figure 4.12: Distribution of the 1000 galaxies sample in Hβ line.

4.1.2.3 Test of AMAZED on 1000 galaxies from SDSS

I aim to test **AMAZED** on 1000 random spectra of **SDSS**. First I make a program that selects randomly 1000 of the existing 1843200 spectra of **SDSS**. In this program I create an array containing all the spectra of all the plates since **SDSS** spectra are nested under the plate that they have been observed with. Then I choose randomly 1000 spectra and I feed it to the program creating "input.spectrumlist" file for **AMAZED**. After running **AMAZED**, I recover the output file "redshift.csv" containing all data about emission lines and their fits. I made a program that reads (with **Pandas** library) and cleans data (removes Nan values...) using the library Pandas and returns some information about the run. On the 1000 galaxies, the software was able to work on 977, where it fitted Hα line of 920 galaxies where 752 have errors under 10%. I plot again the fitted flux in terms of the integrated in figure 4.12.

In this figure, we see that there are negative direct integrated fluxes and the software could not fit the lines, which explain the null values of the corresponding fitted fluxes. So I modified our program to count the number of those galaxies with negative fluxes and I found 464 galaxies. For instance the galaxy "spec-1876-54464-0193" Hα line displayed in figure 4.13, **AMAZED** could not fit it because of the low spectrum quality.

Thus, I clean again data by removing all null and negative values in fitted flux and direct integrated flux. After this manipulation, I end up with 392 galaxies with fitted Hα line where 247 have a relative fit error under 10%. I re-plot again the fitted flux in terms of the direct integrated flux in figure 4.14 and most of galaxies are near the red line meaning that the fit is more or less accurate.

The figure 4.15 represents the module of relative difference is higher and around the velocity 10 km/s . I inspected the galaxy "spec-1983-53442-0525" with the highest rel-

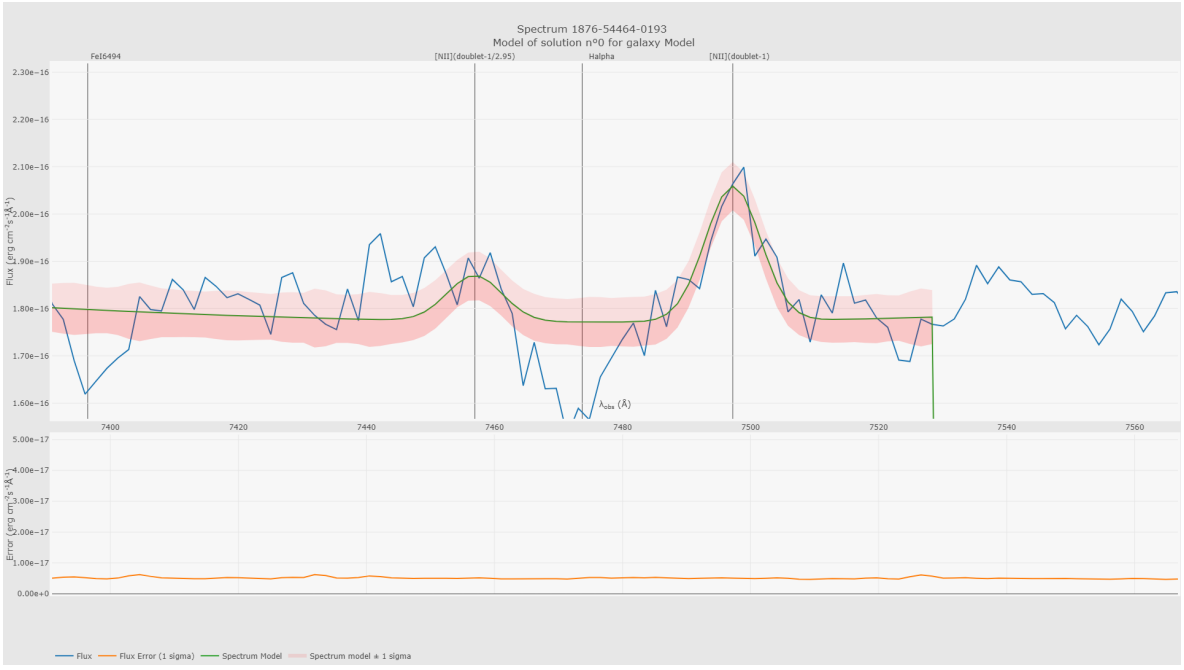


Figure 4.13: Visualisation with **VIZU** the spectrum "spec-1876-54464-0193" and its fit with **AMAZED**.

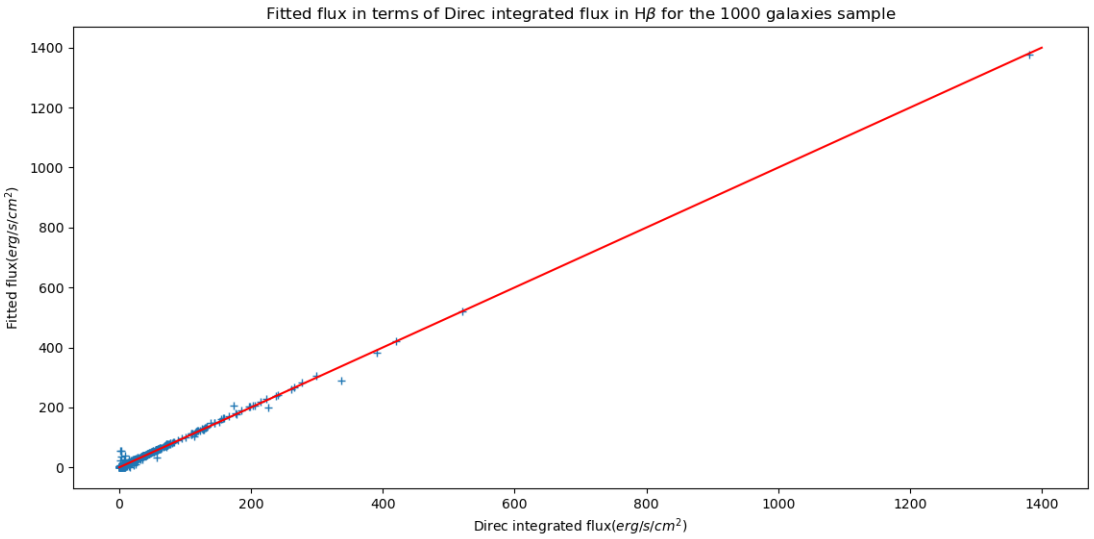


Figure 4.14: Distribution of the 1000 galaxies sample in Hβ line after removing negative and null fluxes.

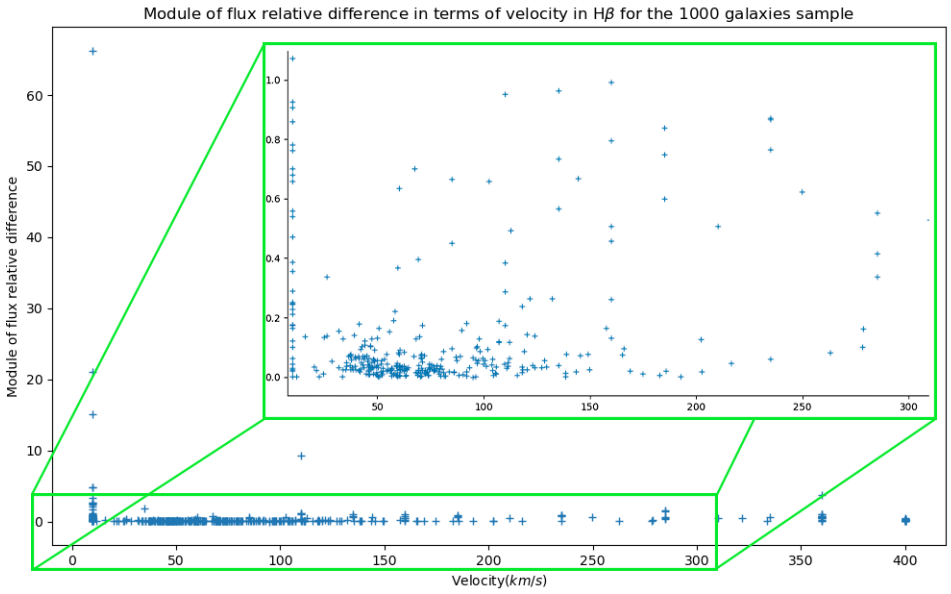


Figure 4.15: Distribution of the 1000 galaxies sample in Hβ line after cleaning data.

ative flux difference = 66.305 and I found the direct integrated flux is 0.813 $erg/s/cm^2$ whereas the fitted flux is 54.76690 $erg/s/cm^2$. Since I computed the relative difference as:

$$\text{Module of relative flux difference} = \frac{\text{abs(Fitted flux} - \text{Direct Integrated flux)}}{\text{Direct integrated flux}}, \tag{4.1}$$

we observe a relative flux difference greater than 1.

By looking at the spectrum and fits of this galaxy with **VIZU**, we found that this galaxy spectrum as the previous one has low quality.

Returning to figure 4.15, we see that there is a concentration of galaxies of higher accuracy at velocities under 150 km/s. This can be explained by the closeness of those galaxies with lower redshift since it is proportional to velocity.

I plot in figure 4.16 module of relative flux difference in terms of direct integrated flux. We observe that the galaxies with higher flux have more accuracy. Indeed, higher fluxes implies a better SNR, eventually a better fit.

4.1.3 AMAZED attributes

For each galaxy, I determine the redshifts and the main lines parameters (flux, velocity, velocity dispersion, line-width...) using the **amazed** software. In the output file, each of them has its own attribute that is affected to every emission line. I represented the most relatives ones in the table 4.1.

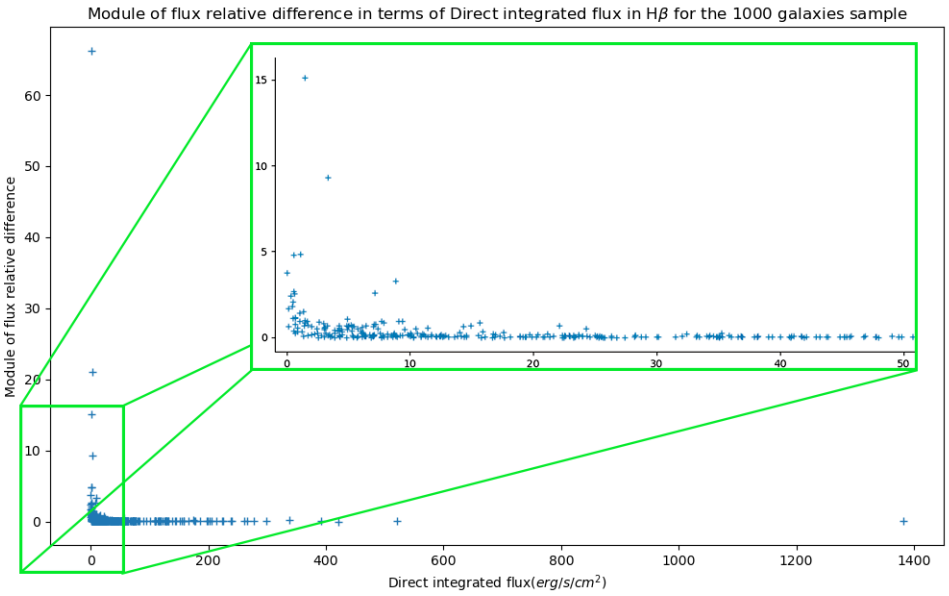


Figure 4.16: Distribution of the 1000 galaxies sample in Hβ of the 1000 galaxies sample after cleaning data.

Name	**SDSS parameters**	**AMAZED parameters**
Line wave length	SDSS_Line_LINEWAVE	LinemeasLineLambda
Line redshift	SDSS_Line_LINEZ	LinemeasRedshift + (1 + LinemeasRedshift)*LinemeasOffset/c
Line disperssion	SDSS_Line_LINESIGMA	LinemeasLineWidth
Flux	SDSS_Line_LINEAREA	LinemeasLineFluxDirectIntegration
Flux computation error	SDSS_Line_LINEAREA_ERR	LinemeasLineFluxDirectIntegrationError

Table 4.1: Table of parameters in **SDSS** and their corresponding in **AMAZED**. "LinemeasOffset" is the velocity offset, and "c" the speed of light.

As I mentioned earlier, the line dispersion attribute is mistakenly named "LinemeasLineWidth".

4.2 GaussPy

4.2.1 Introduction

GaussPy is a Python library designed to implement an algorithm known as Autonomous Gaussian Decomposition (AGD) [10]. AGD employs computer vision and machine learning techniques to automatically and efficiently provide optimized initial estimations for the parameters of a multi-component Gaussian model. Even though it was intended for radio astronomy, AGD can be used on spectra thanks to its speed

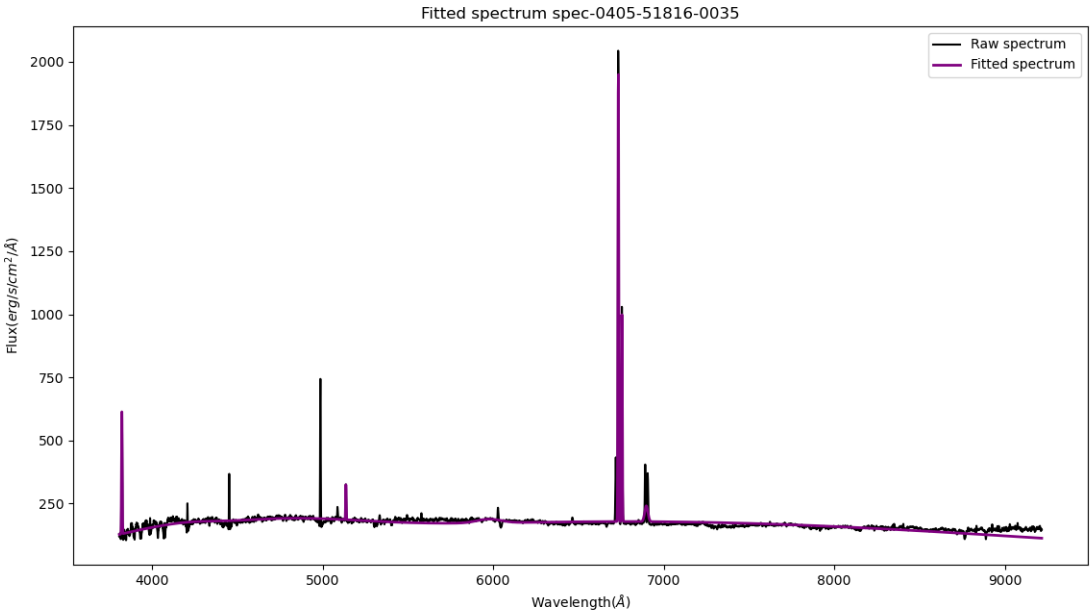


Figure 4.17: Fitted spectrum by **GaussPy**. Raw spectrum is represented in black, where the fitted spectrum is in purple.

and adaptability that makes it well-suited for interpreting substantial volumes of spectral data. I made a comment on this library in section B to understand how the fitting process works. I shall use it to fit emission lines of galaxies and compare it to **AMAZED** fits.

4.2.2 Running GaussPy

In the first test, I try **GaussPy** on the spectrum "spec-0405-51816-0035" and I obtain figure 4.17. We see in this figure that the program does not fit all lines even though I changed the fitting parameters in the AGD program. This problem persisted even on a portion of the spectrum with one emission line. After a discussion with Claire E. Murray, contributor to [10], I understood that **GaussPy** is continuum sensitive. To remove it, I use another library **Specutils** [14] to fit the continuum of the whole spectrum. The use of this library is fairly simple. I build from the wavelength and flux array the object "Spectrum1D". I pass it to the method "fit_generic_continuum()" whose output is an object that can be converted to an array. I subtract it from the spectrum, store the spectrum in a binary file ".pickle", run **GaussPy** again and I obtain the plot in figure 4.18. To check closely the fit quality, I zoom in on the emission lines individually and we observe that the fit quality is not consistent enough for emission lines that coincide with absorption lines. For instance, I zoom in on the Hβ emission line in figure 4.18 and the fit is not accurate despite removing the continuum.

Now I explore the library **Pyplatefit** [15] to better fit the continuum with its absorption lines. This library is promising since I was able to produce the continuum

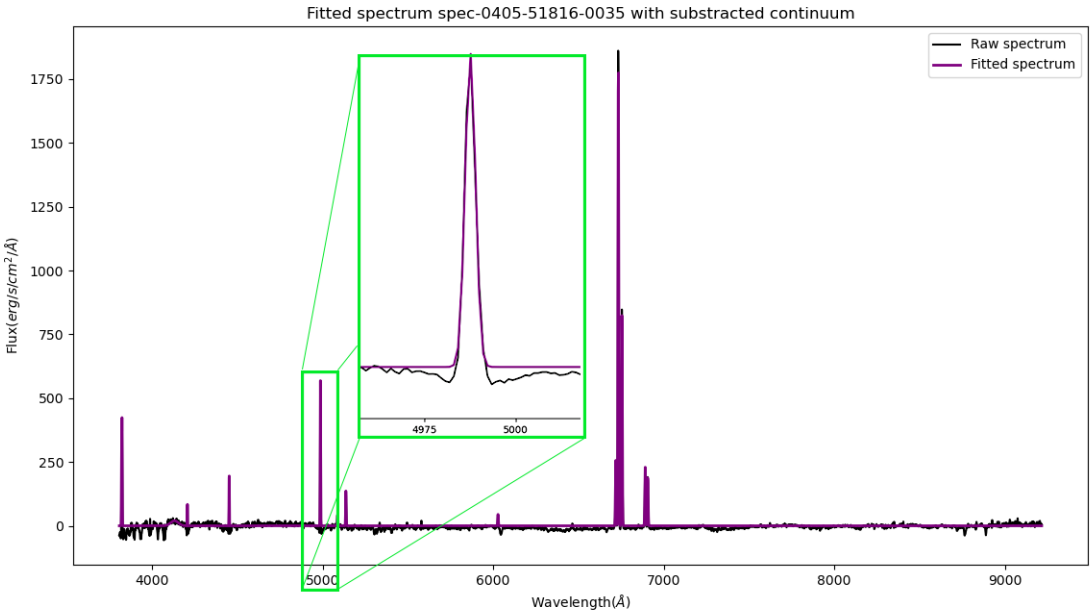


Figure 4.18: Fitted spectrum without continuum by **GaussPy**. Raw spectrum is represented in black, where the fitted spectrum is in purple. The zoom in concerns the Hβ emission line.

fit and the absorption line feature in figure 4.19 from the given spectrum in the test data folder of the library GitHub repository.

To use this library, the spectrum must be stored under the object "Spectrum" from **MPDAF** [16] library. There are 2 ways to build such an object: either from the spectrum file ".fits" (but it must have the same structure of the given test spectrum), or manually with the spectrum flux and the wavelength as "WaveCoord" object. In our case, the only possible option is the latest since **SDSS** ".fits" files have a different structure and **MPDAF** cannot recognize it properly. As soon as I tried to build "Spectrum" object, I was confronted to a new difficulty. In particular, the "WaveCoord" object produces a wavelength series evenly spaced. But in **SDSS**, it is not the case since wavelength is not linear (quadratic) in terms of pixels number in figure 4.20.

In order to overcome this issue, we thought about re-sampling **SDSS** spectra with a wavelength step smaller twice than the minimum step between 2 pixels in **SDSS** to build a proper "Spectrum" object and respect Nyquist–Shannon sampling criterion.

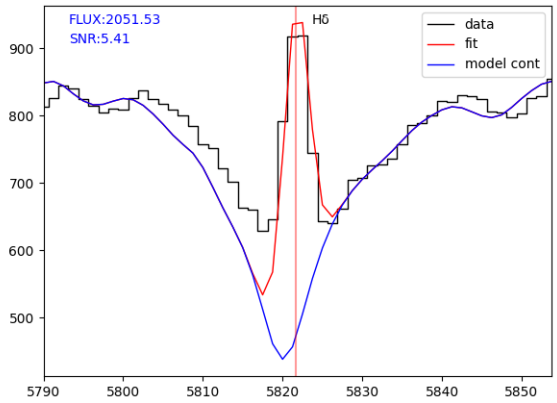


Figure 4.19: Different fits of a spectrum portion from the test example in **Pyplatefit**. Raw spectrum portion is represented in black, where the fitted continuum is in purple and the fitted line in red.

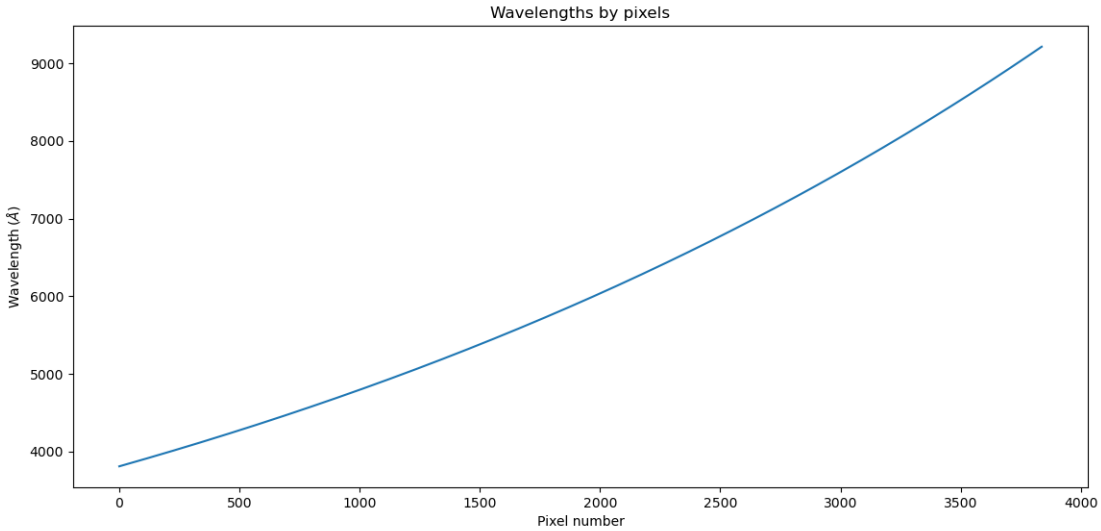


Figure 4.20: Wavelength in terms of the pixel number in **SDSS** specta.

Chapter 5

Selection of the DP galaxies

5.1 First 99740 DP candidates

Géogal Guichard (former intern of Pr. Amram) made a program generating a list of 99740 DP galaxies and I am using this list in this project. He selected them from the 800299 RCSED galaxies using the same method described in stage 1 of figure 2 [9]. Indeed, he conducted a first selection of galaxies at $z < 0.34$ with signal to noise ratio (SNR) superior to 10 either for Hα or O$III\lambda$5008. He computed those SNR and other ones from the Gaussian and non-parametric fits of emission lines in the RCSED given in the files "rcsed_lines_gauss.fits" and "rcsed_lines_nonpar.fits". After, he carried out a second selection by requiring at least 3 SNRs of the lines Hα, Hβ, Hγ, O$III\lambda$5008, O$I\lambda$6302, N$II\lambda$6550, N$II\lambda$6585 to be above 3. Then, he did a third selection under the condition $\chi^2_{non\_parametric} < \chi^2_{Gaussian}$ where he ended up with a total of 99740 DP galaxies.

5.2 Running AMAZED on the 99740 DP candidates

Running **AMAZED** on a big number of galaxies is a delicate task even in high performance computing cluster like the one of LAM. In fact, the wrong settings to run the software can extend multiply running time by 5.

Before running it on the 99740 DP candidates, I made a run test with 1000 galaxies and I measured the needed time to operate on one galaxy is 2 minutes with overestimation (since the cluster processors are not the same). Therefore, in a time limit of 2 hours, a processor would operate on 60 galaxies at least. In this regard, I choose to run **AMAZED** on the queue "short" of the cluster. This queue has fast processors, but it can be available only for short periods of time.

In the settings file of the software "config.json" (A.2.1), I limit the number of processors at 120 because a user is limited to 128, and I kept 8 processors to run **AMAZED** and for other tasks. I set the walltime (time limit for the cluster by the

user) at 2 hours to allow the availability of the cluster for other users. Also, the cluster prioritize shorter tasks over longer ones and walltime limit is 6 hours on the queue "short". I fix the bunch size (A.2.1) at 60, meaning that each processor job would deliver 60 galaxies to work on for 2 hours. Afterwards the cluster works on total of 7200 galaxies per task. A rough estimation of running time on the DP candidates is around 27 hours. I extend this time by 30% to take account of delays that might occur due to the merging of results, and I end up with running time of 35 hours.

To run such a task, I connect to the cluster via terminal. I open **tmux**, and lunch an interactive session for 96 hours, even the estimated running time is 35 hours, since there might be concurrently other users, which would delay execution of our jobs. Then I lunch **AMAZED** on the 99740 galaxies. To continue using freely the cluster, I detach **tmux** window with "CTRL"+"b"+"d".

From here on, after running both **AMAZED** programs and I obtained all fits of all the emission lines of our interest for the 99740 galaxies, we move to the next step to select the DP galaxies based on our own method with Kurtosis and Skewness which represents the core approach of this work. But first let us review those statistical quantities in the following section.

5.3 Skewness

Skewness and Kurtosis are two statistical measures that provide information about the shape and characteristics of a probability distribution or dataset [17], which would be very useful for DP emission lines detection. I compute them using a python program that reads the output fit files of **AMAZED** and **GaussPy** programs and complete Skewness and Kurtosis for each emission line of both output fit files.

Skewness measures the asymmetry of the probability distribution or dataset compared to the normal distribution. It is also known as the third statistical moment and it indicates whether the data is skewed to the left or right of the mean. A positive Skewness value indicates that the distribution is right-skewed, meaning that the tail on the right side of the distribution is longer or fatter than the left tail. In a right-skewed distribution, the mean is typically greater than the median.

A null Skewness value indicates that the distribution is approximately symmetric. Whereas a negative Skewness value indicates that the distribution is left-skewed, meaning that the tail on the left side of the distribution is longer or fatter than the right tail. In a left-skewed distribution, the mean is typically less than the median. Those 3 cases are represented in figure 5.1

To compute the Skewness for an emission line I have to select first an interval of data where I will compute it. For isolated emission lines like Hβ, Hγ and [OIII] I consider the interval $[\mu - 4\sigma, \mu + 4\sigma]$ where μ is the mean of the Gaussian fit and σ its standard deviation. For close doublet emission lines like [NII] or Hα, I choose $[\mu-2\sigma, \mu+2\sigma]$. Regarding the [S$II$] emission line, I take the interval $[\mu-2.5\sigma, \mu+2.5\sigma]$ since the doublet lines are relatively separated.

After choosing the interval, I construct the continuum under the emission line

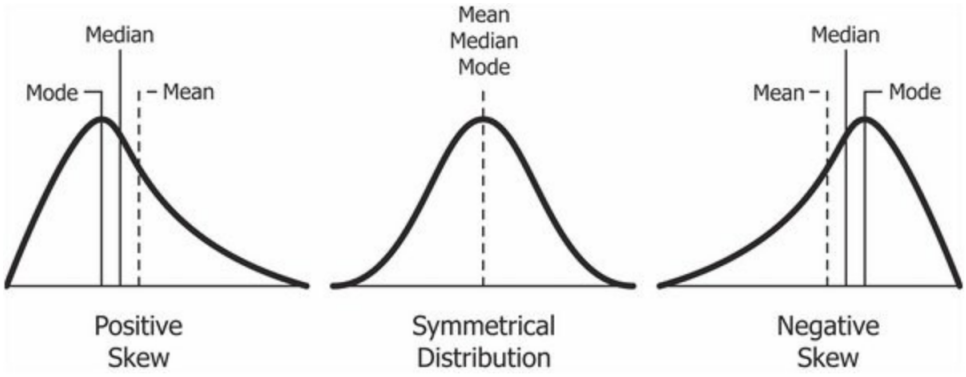


Figure 5.1: A general relationship of mean and median under differently skewed unimodal distribution, Image Credit: Diva Jain.

thanks to the continuum fit as a second-degree polynomial. Then I recover the corresponding the raw unfitted flux values and I subtract the continuum in order to have exact results. Then I compute first the first statistical moment (statistical mean) as:

$$\text{Mean} = \frac{\sum_{i=1}^{N} f_i}{N}, \tag{5.1}$$

where N is the number of pixels inside the selected interval and f_i is the flux subtracted continuum at the "i"th pixel. I calculate the second statistical moment (the variance):

$$\text{Variance} = \frac{\sum_{i=1}^{N} (f_i - \text{Mean})^2}{N}. \tag{5.2}$$

After, I compute the Skewness as:

$$\text{Skewness} = \frac{\sum_{i=1}^{N} (f_i - \text{Mean})^3}{N \times \text{Variance}^{3/2}}. \tag{5.3}$$

5.4 Kurtosis

Kurtosis measures the "tailedness" or peakedness of the probability distribution or data-set related to the normal distribution. It is also known as the fourth statistical moment and it provides information about the presence of outliers and the relative weight of tails compared to the rest of the distribution.

Positive kurtosis (excess kurtosis greater than 0) indicates that the distribution has heavier tails and a more peaked central region compared to a normal distribution. It implies a greater likelihood of extreme values or outliers.

A kurtosis value of 0 (excess kurtosis equal to 0) indicates that the distribution has the same tail behavior as a normal distribution (i.e., it's mesokurtic).

Negative kurtosis (excess kurtosis less than 0) indicates that the distribution has lighter tails and a flatter central region compared to a normal distribution. It suggests fewer extreme values than a normal distribution. Those 3 cases are depicted in

figure 5.2. One must note that Kurtosis is independent of distribution amplitude and characterizes only its peakedness and the heaviness of its tails.

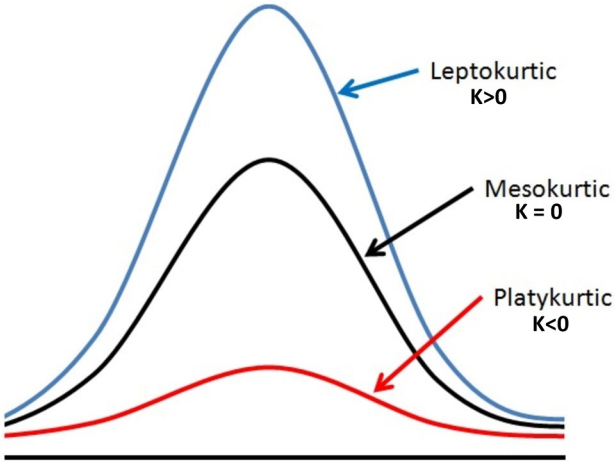


Figure 5.2: Distributions with different values of Excess Kurtosis and different amplitudes, Image Credit: Ajay Mehta.

To calculate it, I follow the same steps as for the Skewness, by selecting the intervals for all statistical moments, and I use the computed mean and variance to work out the Kurtosis as:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N} (f_i - \text{Mean})^4}{N \times \text{Variance}^2}, \tag{5.4}$$

with Excess Kurtosis = K = Kurtosis − 3

5.5 Signal to Noise Ratio

Note that also I use the same program to add the Signal to Noise Ratio (SNR) for all emission lines in both output files. It represents the strength of the signal relative to the level of noise or background in our data. A higher SNR indicates a stronger and more significant signal with less interference from noise. There are many ways to define the SNR, and each of them has its own pros and cons. I calculate it as:

$$\text{SNR} = \frac{\text{Fitted integrated flux}(ergs/s/cm^2)}{\text{Fitted continuum at the Gaussian mean}}. \tag{5.5}$$

This method ensures taking account of the both standard deviation σ and amplitude A of the fitted emission line because the Fitted integrated flux = $A\sigma\sqrt{2\pi}$ (sectionC.1). If I considered the emission line peak instead of the integrated flux, a broad emission line with lower amplitude would have a smaller SNR compared to a narrow emission line with higher amplitude.

5.6 Computing SNR, Skewness and Kurtosis

I made a Python program that reads **AMAZED** output file "redshift.csv" with **Pandas** and removes galaxies upon which the software did not run. Then it computes for each emission line respectively SNR, Skewness and Kurtosis according to the methods described previously, and store them in the data frame. I tested this program on the 1000 galaxies sample and I was confronted to the issue of some incorrect fits or completely unfitted lines for certain spectra. To resolve this issue, I put 2 conditions on the fit of each line before computing the SNR, Skewness and Kurtosis: the measured emission line wave different from Nan, and under it has to be under 9000 because the spectra wavelength range is $[3800, 9222]$Å.

Then I save this new data frame in a new updated "redshift_updated.csv" file. Since the output file of the 100000 candidates is large, I run the program on the cluster. To do so, I made a "slurm" script shell to run this program on cluster session of 120 cores and 32 GB or ram using the queue "short".

Chapter 6

Results and discussion

In this chapter, I shall discuss the results I obtained with our selection method. In order to better interpret DP galaxies results, I created control sample (CS) of 100000 galaxies I selected randomly from **SDSS**. This CS would serve us to compare Double Peaked Sample (DPS) to Non Double Peaked galaxies in different ways so that I can assess the effectiveness of our method.

I started by plotting a frequency distribution in Skewness and Kurtosis for both DP and CS. I obtained figure 6.2 for the Hα line.

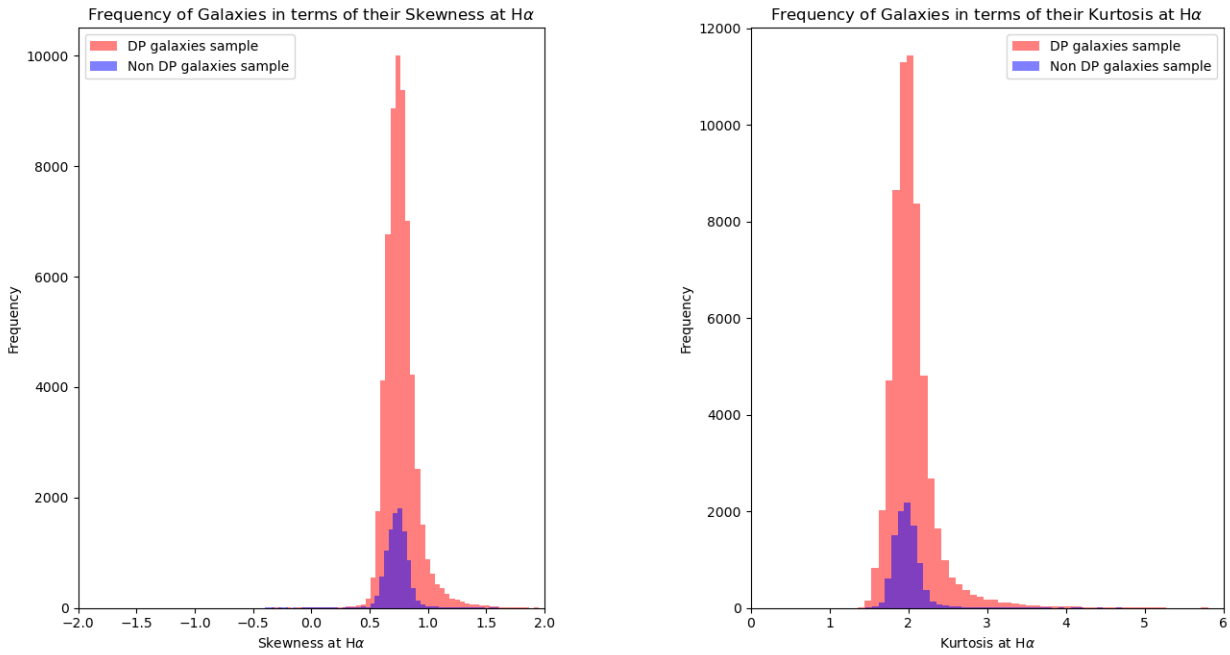


Figure 6.1: Frequency distributions of Skewness and Kurtosis for DPS sample and CS at Hα. In red (respectively blue), it is the frequency distribution of DPS (respectively CS).

By observing this figure, the first remarkable point is the amplitude difference of histograms. Visually, the total number of galaxies in DPS is much larger than CS one in both Skewness and Kurtosis figures. To verify that, I calculate the total

number of galaxies that have been computed in the DPS and CS and I found 60829 in contrast to 10069 for CS. In fact, before making the histograms, I selected only galaxies with positive SNR (that I personnaly computed) in all the lines [1] Hα, Hβ, Hγ, [OIII]λ5008, [OI]λ6302, [NII]λ6550, and [NII]λ6585 (1) in both DPS and CS.

Also, I considered only galaxies with positive fitted and direct integrated fluxes (2). Indeed I made those criteria (1) and (2) even after selecting galaxies DPS when I made it previously in section 5.1. On one hand, first selection was performed on **RCSED** catalog data, namely on the SNRs computed by the algorithms of catalog. On the other hand, the SNRs I computed were done based solely on **AMAZED** data. Since it is a different algorithm with different fits methods, SNR computational method (cf section 5.5)... I cannot expect to have all the 99740 galaxies in DPS and 100000 in the CS. Moreover, (1) and (2) criteria were made only to ensure obtaining consistent results.

In order to exploit those histograms despite DPS and CS have different galaxy numbers, I normalized the distributions of Skewness and Kurtosis such that the area under the histogram is equal to 1 and it yeilds to figure 6.2.

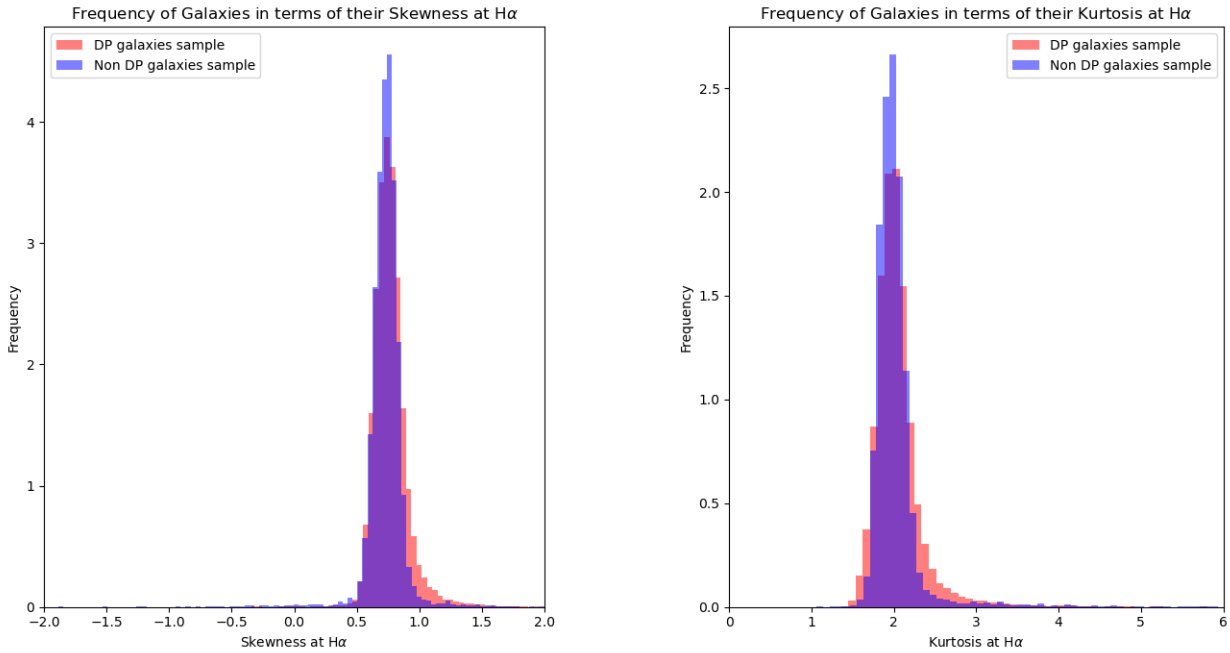


Figure 6.2: Frequency distributions of Skewness and Kurtosis for DPS sample and CS at Hα after normalization. In red (respectively blue), it is the frequency distribution of DPS (respectively CS).

By examining the Skewness histograms in this figure, Both CS and DP are not

[1] Actually, for [NII]λ6550, I could not retrieve it because of the line nomination problem in **AMAZED** that has not been solved until this moment of writing the report. The issue lies within the point in the name of this line "[NII](doublet-1/2.95)" which **AMAZED** interprets as the attribute "95" of the attribute "galaxy.linemeas.[NII](doublet-1/2" since it uses object oriented programming. Therefore I limited our study to those 6 lines out of 7 that remains also relevant thanks to the high SNR of the 6 lines.

centered at 0. For CS, the mean Skewness is at 0.731 and it means that most of CS galaxies have positive skewed Hα line even they were supposed to have single emission lines, and their emission line shape is not symmetric and further different from the Gaussian profile. There can be many reasons to interpret this results: first the closeness of Hα line to NII lines, second the spectrum resolution... I shall verify the first hypothesis the same plot for Hβ since this latter is not surrounded by another emission line. Furthermore, the CS Skewness distribution is not symmetric and its Skewness (of the frequencies) is -2.389, which implies that there are more galaxies with Kurtosis lower than 0.731 with more symmetric Hα line.

Alternatively, the DPS presents a mean at 0.768, not very far from that of CS, but we can see clearly that the distribution is not symmetric and its Skewness (of the frequencies) reads 1.376. Indeed, it means that there are more galaxies with Skewness higher than 0.768 which is consistent with DP galaxies.

On the right side of figure 6.2, the CS Kurtosis mean is around 2.02, which represents a less peaked emission line with thin tails. For DPS, the distribution presents a mean at 2.04, slightly larger than CS. This is the opposite of what I expected since DPS galaxies have broader emission line with fatter tail. It can also be due to the closeness of [NII] lines to Hα. Moving the frequency distribution symmetries, CS has a Skewness (Skewness of the frequenecies distribution) of 6.41, where DPS has 3.97. The bigger Skewness signifies that there are more galaxies with stronger peaks and closer to a single peak emission line Gaussian profile. For DPS, the lower value of Skewness transcribes the presence of less galaxies with higher peaks, which is in congruence with the selection pool since the emission lines of its galaxies are larger and less peaked.

To verify the discrepancies we observed in the previous figure for the values of Skewness and Kurtosis mean values for DPS and CS, I generate the same histograms for Hβ line in figure 6.3. This emission line would make a compelling evidence since it is the second brightest line emission after Hα and not surrounded by any emission line.

On the left side of this figure, both CS and DPS are closer to 0 compared to those at Hα, with a mean value respectively 0.58 and 0.63. The CS presents a lower mean than Hα CS. Therefore the closeness from NII lines contributes in making the asymmetry of the spectrum. Nevertheless, the mean CS frequencies distribution is not null, which means that the spectrum quality also contributes to the asymmetry. Measuring the Skewness of those frequency distributions reads -2.50 meaning that there are more galaxies on the left, closer to 0, whose emission lines are closer to the Gaussian profile. For DPS, the mean is higher than CS, but remains lower than Hα case. The Skewness of those frequencies is equal to -2.46. It is lightly higher than the CS, meaning that there are less galaxies than in the CS that are close to the Gaussian profile, but in general only few galaxies that presents a substantial asymmetry. Generally, DPS frequencies distribution confirms the deductions I made for CS.

On the right side, the mean of Kurtosis distribution for CS is at 1.94. This is

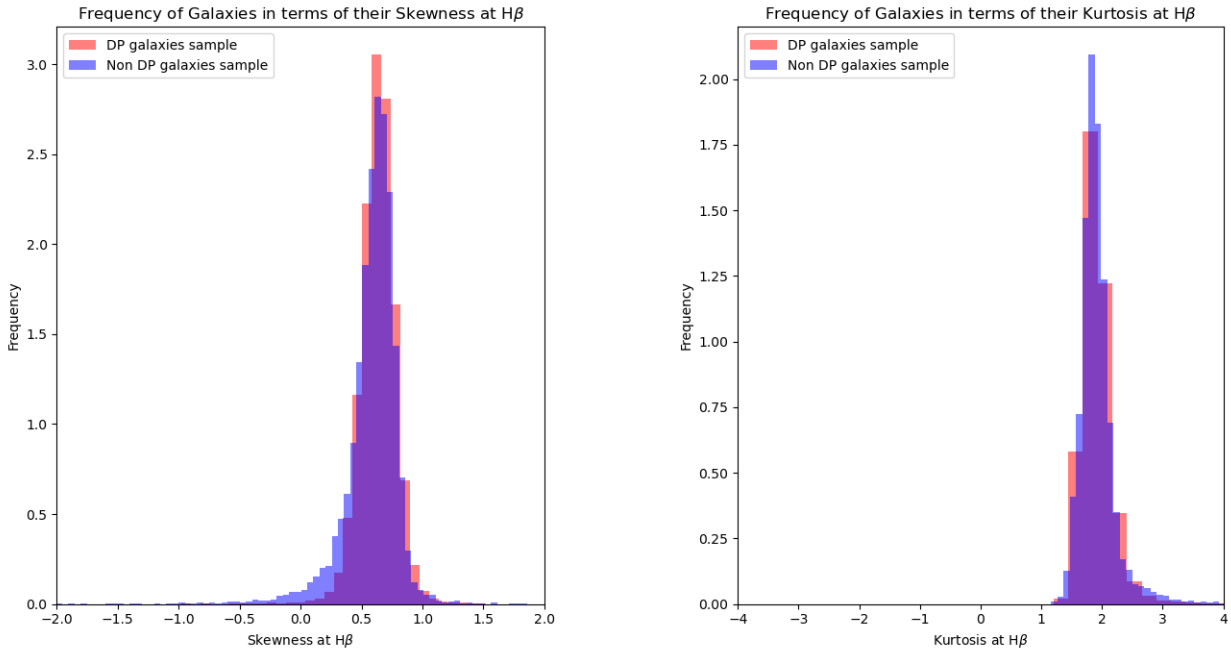


Figure 6.3: Frequency distributions of Skewness and Kurtosis for DPS sample and CS at Hβ. In red (respectively blue), it is the frequency distribution of DPS (respectively CS).

lower than value I obtained for Hα, and suggests that Hβ lines are flatter even for the single peaked galaxies. The Skewness of this distribution of frequencies is 6.67, slightly higher than for Hα meaning that more galaxies have emission lines with lower peakedness than higher. Regarding DPS, the mean value is 1.93 with a distribution Skewness at 13.64. The mean is also lower than Hα case, indicating that Hβ lines are considerably broader than Hα, but the Skewness is more than three times higher than for Hα, meaning most of DPS galaxies have lower peaked lines. The global conclusion we can make from this figure is that CS presents nearly the same distribution, but DPS presents larger and less peaked emission lines especially in Hβ than Hα, with more ratio of galaxies presenting hypothetically double emission lines.

A possible improvement to this study is performing the same preliminary selection on CS as I did for DPS, to create the CS in order to make a non biased CS. On the top of that, [9] mentioned that there is another bias of mass that I should account for in building the CS.

Chapter 7

Conclusion

To conclude, we introduced the context of DP galaxies and their significance in galaxy evolution and correlation with galaxy mergers. We discussed the **SDSS** and **RCSED** catalogs used in research on DP galaxies. We reviewed previous literature on DP galaxies, examining galaxy population, morphology, and emission characteristics. This allowed us to propose possible origins for these galaxies, including minor mergers, gas accretion, and S0 type galaxies.

To identify DP galaxies, we utilized the software **AMAZED** and the AI-based library **GaussPy** for fitting emission lines. However, before using these tools, we tested them on different galaxy samples and provided feedback to improve their performance. We found that **GaussPy** can fit spectra but requires the removal of the continuum. We used **Specutils** to remove the continuum, successfully fitting all emission lines. However, the fit quality was poor for emission lines close to absorption ones. We then tried **Pyplatefit**, which has excellent continuum fitting features. However, it only works with spectra in WCS coordinates, so we considered transforming **SDSS** spectra into WCS coordinates.

Next, we explained our approach to identifying DP galaxies, starting with 99740 candidates from the aforementioned catalogs. We described how we efficiently ran **AMAZED** on this large number of galaxies. Additionally, we briefly reviewed Skewness and Kurtosis, which are central to our approach for detecting DP galaxies. Finally, we discussed the results of Skewness and Kurtosis for two galaxy samples: DPS (containing DP candidates) and CS (containing single-peaked galaxies). Despite having different total counts, we made interesting deductions. Both DPS and CS exhibit asymmetric and relatively flat emission lines, potentially due to spectrum quality and the overlap of some emission lines as it was shown for Hα emission line. However, DPS has a higher ratio of galaxies with higher Skewness (higher asymmetry to the left) emission lines and a more significant ratio of lowly peaked galaxies especially in Hβ line, which confirms the presence of Double Peaked galaxies in DPS. We can improve this study by the same preliminary selection criteria for CS and DPS. We also need to consider galaxy masses as they can affect line widths and amplitudes.

To further advance this study, we can make the preliminary selection based on **AMAZED** emission line fits and computed SNRs without relying on **SDSS** or

RCSED data. We can then compare the results to those obtained in [9].

I regrouped some of the relevant programs I used in this project in a public GitHub repository. This repository is still under construction and I will update gradually its content.

A word about this internship

This internship provided me with the opportunity to deepen my knowledge about galaxy evolution. I also learned a lot about various aspects of galaxies, such as typology, morphology, and kinematics. On the technical side, I enjoyed working on programs, managing and analyzing data, testing software, libraries, and exploring their different aspects and capabilities. Attending conferences and seminars also broadened my insight into observational astrophysics.

Overall, the subject of this internship was very enriching and educational. I discovered many new things in terms of both knowledge and skills. In general, it went well, and I am grateful for the opportunity to have this experience at the LAM.

Appendix A

General documentation

We report here the documentation we personally created for **AMAZED**. While some readers already acquainted with the process might find this documentation too detailed, because it includes some information about the remote connection to a local computer and the cluster... but we could not find this piece of information on any third party documentation for new users like us. For convenience, we will use a colour code. We shall use green for file content.

A.1 Cluster and Linux complementary documentation

To use **AMAZED**, we need first to connect to the Cluster from a local computer. If it is not the case, we can connect a laboratory computer using a VPN, and from the following command on the shell with **LDAP** username:

```
ssh -Y _ldap_username_@_computer_name_.lam.fr
```

To forward the GUI of a remote computer, we use:

```
ssh -X _ldap_username_@_computer_name_.lam.fr
```

In the terminal of the remote machine, we can open the file manager as:

```
nautilus --new-window
```

To connect to the cluster, we use:

```
ssh _ldap_username_@cc.lam.fr
```

and we enter the password. To copy a file from the local computer to the user folder on the cluster, we use the following the command in the local computer terminal:

```
scp _file_directory_in_local_computer_
↪ _ldap_username_@cc.lam.fr:_directory_in_cluster_
```

and from the cluster to the LAM local computer:

```
scp _ldap_username_@cc.lam.fr:_file_directory_in_cluster_
↪ _directory_in_local_computer_
```

From an external computer, one must use "@cc.lam.fr:" instead of "@cc:". To copy a folder, the commands remain the same, and we add "-r" after "scp". It is useful to open many terminal tabs on the cluster. For this end, we use:

```
tmux
```

To open a new, we use "CTRL"+"b" then "c". To switch between tabs we use "CTRL"+"b" then "n". To scroll in up, we use "CTRL"+"b" then "[" and up button. For more information about **tmux**, visit this online documentation.

In the home folder, there is a folder under the user name where we can put all our files and run them on the cluster. In order to see the running tasks on the cluster we use:

```
usage.py
```

and displays the running tasks for each user, and the number of nodes is using in his tasks. Note that for running any task on the cluster (running a program, compiling...), we need to specify the node, otherwise it will be run on the cluster head node, which would create conflicts afterwards. Indeed, the cluster head node is dedicated only to OS tasks. To require a node:

```
srun -c1 -p mem -t 03:00:00 --pty bash
```

which will require a node under **mem** partition for 3 hours. This standard partition provides the user with a node having 24 cores and 512 GB or ram. More information about partitions is provided by the online documentation of the cluster. We can require a node at each **tmux** tab. To display all the running jobs on the cluster:

```
squeue
```

and to see all the running jobs in a specific queue, we can use pipe:

```
squeue |grep batch
```

To display number of used nodes used and the remaning time time the user, we use:

```
squeue -u _username_
```

To cancel a running job on the cluster:

```
scancel _job_id_
```

To cancel all user's jobs on the cluster:

```
scancel -u _username_
```

A.2 AMAZED documentation

A.2.1 Running AMAZED

In order to run **AMAZED**, one must goes to the folder "/net/CESAM/amazed" and create his own folder where all the results generated by **AMAZED** will be stored. "Venvs" is the folder containing all available **AMAZED** versions. There are many versions and the one we will use is "amazed_0.44-RC1". To set up the options, we create in our folder the file "config.json" having all options to run the program. Using Vim, we can access to this file and it contains:

```
"parameters_file":  "parameters.json",
"input_file":  "input.spectrumlist",
"spectrum_dir":  "spectra",
"calibration_dir":  "/net/CESAM/amazed/calibration",
"pre_command":  "source /net/CESAM/amazed/venvs/amazed_0.42.0/bin/activate"
```

This file involves many other files. "parameters_file:" receives the directory of parameters file "parameters.json" containing finer options concerning the spectra (fits parameters for the 1st and 2nd step of the program...), "input_file" receives the directory of the file "input.spectrumlist" with the spectra on which we run the program, "spectrum_dir": receives "spectra" the directory of the folder containing all the spectra. One should notice that we specified only the name of those 3 files because they are in the same folder where "config.json" is. Otherwise we have to write down all the whole directory of each file we use in "config.json". Last but not least, "calibration_dir": receives "/net/CESAM/amazed/calibration" folder directory of calibration. This file contains all calibration data used by **AMAZED**, such as absorption and emission lines, type of the spectra continuum of galaxies... in fact this file is used in both first and second steps of the program.

To open the **AMAZED** program:

```
source /net/CESAM/amazed/venvs/amazed_0.44-RC1/bin/activate
```

and it has to be opened in the same tab of **tmux** where we selected the node. To shut down the program:

```
deactivate
```

To know the program version:

```
amazed --version
```

To run the program on spectra:

```
amazed -c ./config.json -o output
```

on our machine. We can also use:

```
amazed -c ./config.json -o output -w slurm
```

The additional command "-w slurm" is used to run the program with the integrated settings of **SLURM** task manager inside **AMAZED** and run it on the cluster, which would be faster. In fact running on the cluster cores would be faster, knowing that each user can have up to 128 cores using the partition **sbatch**. In this regard, we shall use some features in **AMAZED** to run it on the cluster many cores. For instance, if we wish to run **AMAZED** on 100000 spectra, we have 128 cores at our disposal on the cluster, and we execute the following command:

```
amazed -c ./config.json -o output -w slurm -s 1000 -j 80
```

it will run on each of the 80 processors 1000 spectra. In other words, it creates 100000/1000 = 100 bunches, each of them containing 1000 spectra with "-s 1000". Each of those 100 bunches will be run on 80 cores from 128 cores at our disposal because we limited cores number with "-j 80". The order of those additional features is irrelevant. We can also put them inside the "config.json" file by appending the following lines:

```
"output_dir":  "output",
"bunch_size":  "100",
"worker":  "slurm",
"concurrency":  "80"
```

To specify the running specifications of **AMAZED** on the cluster, respectively queue, wall-time, used memory, and bunch size for a run by **AMAZED**, we append respectively the following line in the file "config.json":

```
"queue":  "batch",
"walltime":  "02:00:00",
"mem":  "3G",
"bunch_size":  1
```

meaning that we use the queue "batch", for a maximum time of 2 hours (if the task takes more then 2 hours, it will stop at the first run, then continue later according to

the available resources), and memory size of 3 GB with a bunch size of 1 spectrum per processor, knowing that every user has 128 at his disposal.

To know the time and cluster resources that have been used to run **AMAZED** and make the output:

```
amzperfstats  _output_directory_
```

A.2.2 Running AMAZED on SDSS spectra

To run **AMAZED** on an **SDSS** spectra, we use a non-official release of the software as a virtual environment on the cluster path:
"/net/CESAM/amazed/aallaoui/venvs/test/fix_issue_8015/40949022/".

After lunching this version, we also update the **AMAZED** path in the configuration file "config.json" and we expend it by the following line:

```
"reader":  "sdss"
```

and we can run the program on **SDSS** spectra.

While working on **SDSS** spectra, one should always multiply all output results of **AMAZED** namely amplitudes and continuum by 10^17 because the software converts the spectra amplitude. On the top of that, we do not have to specify the path of galaxies folder, but we must define the plate number of the galaxy near to the galaxy ID in the "input.spectrumlist". To this end, we made a program that create such a file numbers from a list of spectra. For each spectrum, it looks for its corresponding plate in the **SDSS** path on the cluster
"/net/CESAM/amazed/dataset/SDSS/dr17/sdss/spectro/redux/26/spectra/lite" and write it in the "input.spectrumlist".

After running the program, it creates the the output folder "output" containing the results. It has the configuration file "config.json" with all its underlying files... "redshift.csv" is a compact table with red-shifts only with "errors.csv" the corresponding errors explaining the encountered problems by the program if it does not run for correctly, and the remaining files are log files. Also the folder "0000" holding the results on the spectra, in files under the name "resuts.hdf5". To read this file we use the commands:

```
h5dump resuts.hdf5
```

or:

```
h5ls resuts.hdf5
```

However, reading those files is very complicated and we will use a tool designed for this purpose that we shall describe in the next subsection.

A.2.3 Visualing results with VIZU

To better visualise the results, there is a built-in graphical user interface (GUI) in **AMAZED** called **VIZU**. It displays the spectra, with all fits for both line emissions and absorptions, in addition to the content of the file "redshift.csv". In a cluster terminal tab with **AMAZED** opened, we run **VIZU** on a port using:

```
vizu -p _port_number_ _output_directory_ -r _reference_file_directory
```

where "_output_directory_" is the output directory of **AMAZED** and "_reference_file_directory" is the directory of the file containing theoretical values.

Now in a terminal tab on our local computer we connect to the opened port on the cluster via a tunnel using

```
ssh -t -L _port_number_:localhost:_port_number_
↪ _user_name@node_node_number_.lam.fr
```

then we open in our web browser http://localhost:_port_number_ and we access to the **VIZU** GUI. For example, to visualise with **VIZU** a sample of 48 galaxies, we run it on the port 5055 using the command:

```
vizu -p 5055 output_48_sample -r
↪ /net/CESAM/amazed/aallaoui/pfs_weekly/science/catalog.csv
```

and we create the tunnel from our local computer using:

```
ssh -t -L 5055:localhost:5055 ytemmam@node20.lam.fr
```

From a Windows computer, we create a tunnel with:

```
ssh -t -J ytemmam@cc.lam.fr -L 5055:localhost:5055
↪ ytemmam@node20.lam.fr
```

A.2.4 Extracting data with AMZEXPORTATTRIBUTES

The ".hdf5" files are "containers" with classes including much information. "Galaxies" represent a class of objects, where each object is an individual galaxy. In order to extract the data we need, we use the built-in program **AMZEXPORTATTARIBUTES**. It creates from the output folder a file containing the data we extract following ".hdf5" files structure in the output model documentation, and the parameters in the parameters documentation of **AMAZED**. Before running this program, one need to create a temporary folder in the cluster at his session and add the following line in his "bashrc" file

```
export TEMP_DIR=_temporary_folder_directory_
```

To run the program:

```
amzexportattributes _output_directory_ -a
↪ _object_._attribute/method_... -b -1 -d _file_extension_ -o
↪ _file_name_
```

where the object class corresponds to the structure of ".hdf5" file. In our case, we use this program to extract data of galaxies. For instance, to extract Hα and Hβ line flux in the file "Halpha_beta_flux_48_sample.csv" file for our 48 galaxies sample, we use:

```
amzexportattributes output_48_sample -a
↪ galaxy.linemeas.Halpha.LinemeasLineFlux
↪ galaxy.linemeas.Hbeta.LinemeasLineFlux -b -1 -d csv -o
↪ Halpha_beta_flux_48_sample.csv
```

where "galaxies" is the class, "linemeas" is a method of "galaxies". Each absorption and emission line represents an attribute. The emission lines sub-classes name are found in the file "/net/CESAM/amazed/calibration/linecatalogs/linecatalogamazedvacuum_H1.tsv". "LinemeasLineFlux" is one of the possible parameters we can extract reported in the **AMAZED** parameters documentation.

A.2.5 Extracting Data inside of "redshift.csv"

Another way to extract the data we need for galaxies directly in the file "redshift.csv" wihtout relying on **AMZEXPORTATTRIBUTES**. The advantage of this method lies within a considerable gain of time especially for a large number of galaxies because it is done simultaneously in **AMAZED**. Firstly, we make a copy of the file "summary.conf" from "/net/CESAM/amazed/aallaoui/" in our directory. Then we add the data we desire to extract following the same syntax **__object__._attribute/method__...**:

```
galaxy.linemeas.Halpha.LinemeasLineFlux
galaxy.linemeas.Hbeta.LinemeasLineFlux
....
```

We add the following line in "config.json":

```
"summary_conf_path":  "summary.conf_directory"
```

and we run **AMAZED**.

Appendix B

Comment on GaussPy Library

This section is intended to shed more light about the GaussPy library. Indeed, the online documentation provides only a description of the Library through several examples without explaining the routines included, their arguments nor their outputs. For instance, the method "gausspy.gp.GaussianDecomposer()" takes the parameters:

- number of phases for the Gaussian fit [10],
- signal to noise ratio for fits,
- the value of the phases

To make the fits, there is a sub-method "batch_decomposition()" that take into argument only binary files under the extension ".pickle" and make Gaussians fits on its data. This latter must be stored as a dictionary, where spectrum is an array under the key 'data_list', the wavelength is an array under 'x_values' and spectrum errors is an array under 'errors'. Any additional data in ".pickle" files would create conflicts. The output of this routine is a dictionary containing the Gaussian fits parameters (amplitudes, mean, full width at half maximum...).

Appendix C

Mathematical concepts

C.1 Gaussian function

A Gaussian function is defined as follows:

$$f(x) = A \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \tag{C.1}$$

where A is the amplitude, μ is the mean and σ is the standard deviation. Some properties of Gaussian like the area under the Gaussian is:

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = A\sigma\sqrt{2\pi}, \tag{C.2}$$

and the Full Width at Half Maximum (FWHM):

$$\mathrm{FWHM} = 2\sqrt{2\ln(2)}\sigma. \tag{C.3}$$

Bibliography

[1] R. Brent Tully, Hélène Courtois, Yehuda Hoffman, and Daniel Pomarède. The laniakea supercluster of galaxies. *Nature*, 513(7516):71–73, sep 2014.

[2] Gayoung Chon, Hans Böhringer, and Saleem Zaroubi. On the definition of superclusters. *Astronomy & Astrophysics*, 575:L14, mar 2015.

[3] Fred C. Adams and Gregory Laughlin. A dying universe: the long-term fate and evolutionof astrophysical objects. *Reviews of Modern Physics*, 69(2):337–372, apr 1997.

[4] G. C. Myeong, N. W. Evans, V. Belokurov, J. L. Sanders, and S. E. Koposov. The sausage globular clusters. *The Astrophysical Journal*, 863(2):L28, aug 2018.

[5] Jennifer M. Lotz, Patrik Jonsson, T. J. Cox, and Joel R. Primack. Galaxy merger morphologies and time-scales from simulations of equal-mass gas-rich disc mergers. *Monthly Notices of the Royal Astronomical Society*, 391(3):1137–1162, dec 2008.

[6] Rahul Kannan, Andrea V. Macciò, Fabio Fontanot, Benjamin P. Moster, Wouter Karman, and Rachel S. Somerville. From discs to bulges: effect of mergers on the morphology of galaxies. *Monthly Notices of the Royal Astronomical Society*, 452(4):4347–4360, 08 2015.

[7] Sara L. Ellison, David R. Patton, Luc Simard, and Alan W. McConnachie. Galaxy pairs in the sloan digital sky survey. i. star formation, active galactic nucleus fraction, and the mass-metallicity relation. *aj*, 135(5):1877–1899, may 2008.

[8] Julia M. Comerford, Brian F. Gerke, Jeffrey A. Newman, Marc Davis, Renbin Yan, Michael C. Cooper, S.M̃. Faber, David C. Koo, Alison L. Coil, D.J̃. Rosario, and Aaron A. Dutton. Inspiralling supermassive black holes: A new signpost for galaxy mergers. *apj*, 698(1):956–965, jun 2009.

[9] Daniel Maschmann, Anne-Laure Melchior, Gary A. Mamon, Igor V. Chilingarian, and Ivan Yu. Katkov. Double-peak emission line galaxies in the SDSS catalogue. *Astronomy & Astrophysics*, 641:A171, sep 2020.

[10] Robert R. Lindner, Carlos Vera-Ciro, Claire E. Murray, Snežana Stanimirović, Brian Babler, Carl Heiles, Patrick Hennebelle, W. M. Goss, and John Dickey. Autonomous gaussian decomposition. *The Astronomical Journal*, 149(4):138, mar 2015.

[11] Igor V. Chilingarian, Ivan Yu. Zolotukhin, Ivan Yu. Katkov, Anne-Laure Melchior, Evgeniy V. Rubtsov, and Kirill A. Grishin. Rcsed—a value-added reference catalog of spectral energy distributions of 800,299 galaxies in 11 ultraviolet, optical, and near-infrared bands: Morphologies, colors, ionized gas, and stellar population properties. *The Astrophysical Journal Supplement Series*, 228(2):14, feb 2017.

[12] Lisa J. Kewley, David C. Nicholls, and Ralph S. Sutherland. Understanding galaxy evolution through emission lines. *Annual Review of Astronomy and Astrophysics*, 57(1):511–570, aug 2019.

[13] B. Epinat, T. Contini, O. Le Fèvre, D. Vergani, B. Garilli, P. Amram, J. Queyrel, L. Tasca, and L. Tresse. Integral field spectroscopy with SINFONI of VVDS galaxies. I. Galaxy dynamics and mass assembly at $1.2 < z < 1.6$. *Astronomy & Astrophysics*, 504(3):789–805, September 2009.

[14] Astropy-Specutils Development Team. Specutils: Spectroscopic analysis and reduction. Astrophysics Source Code Library, record ascl:1902.012, February 2019.

[15] Roland Bacon, Jarle Brinchmann, Simon Conseil, Michael Maseda, Themiya Nanayakkara, Martin Wendt, Raphael Bacher, David Mary, Peter M. Weilbacher, Davor Krajnović , Leindert Boogaard, Nicolas Bouché, Thierry Contini, Benoît Epinat, Anna Feltre, Yucheng Guo, Christian Herenz, Wolfram Kollatschny, Haruka Kusakabe, Floriane Leclercq, Léo Michel-Dansac, Roser Pello, Johan Richard, Martin Roth, Gregory Salvignol, Joop Schaye, Matthias Steinmetz, Laurence Tresse, Tanya Urrutia, Anne Verhamme, Eloise Vitte, Lutz Wisotzki, and Sebastiaan L. Zoutendijk. Tthe muse hubble ultra deep field surveys: Data irelease II. *Astronomy & Astrophysics*, 670:A4, jan 2023.

[16] Roland Bacon, Laure Piqueras, Simon Conseil, Johan Richard, and Martin Shepherd. MPDAF: MUSE Python Data Analysis Framework. Astrophysics Source Code Library, record ascl:1611.003, November 2016.

[17] R.M. Warner. *Applied Statistics II: Multivariable and Multivariate Techniques*. SAGE Publications, 2020.

[18] B. Epinat, L. Tasca, P. Amram, T. Contini, O. Le Fè vre, J. Queyrel, D. Vergani, B. Garilli, M. Kissler-Patig, J. Moultaka, L. Paioro, L. Tresse, F. Bournaud, C. López-Sanjuan, and V. Perret. MASSIV: Mass assembly survey with SINFONI in VVDS. *Astronomy & Astrophysics*, 539:A92, mar 2012.