Unsupervised classification of large samples of galaxy spectra

Julien Dubois

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : PHYS - Physique
Spécialité : Astrophysique et Milieux Dilués
Unité de recherche : Institut de Planetologie et d'Astrophysique de Grenoble

Classification non supervisée de gros échantillons de spectres de galaxies

Unsupervised classification of large samples of galaxy spectra

Présentée par :

Julien DUBOIS

Direction de thèse :

Didier FRAIX-BURNET CHARGE DE RECHERCHE HDR, CNRS DELEGATION ALPES	Directeur de thèse
Jihane MOULTAKA Toulouse III, Université Toulouse III - Paul Sabatier	Co-encadrante de thèse

Rapporteurs :

STEPHANE ARNOUTS
DIRECTEUR DE RECHERCHE, CNRS DELEGATION PROVENCE ET CORSE
IRYNA VAVILOVA
PROFESSEUR, Main Astronomical Observ NAS of Ukraine

Thèse soutenue publiquement le **5 septembre 2023**, devant le jury composé de :

STEPHANE ARNOUTS DIRECTEUR DE RECHERCHE, CNRS DELEGATION PROVENCE ET CORSE	Président
CHARLES BOUVEYRON PROFESSEUR DES UNIVERSITES, UNIVERSITE COTE D'AZUR	Examinateur
MARC HUERTAS-COMPANY MAITRE DE CONFERENCES HDR, UNIVERSITE PARIS 7 - DENIS DIDEROT	Examinateur
ESTELLE MORAUX MAITRE DE CONFERENCES HDR, UNIVERSITE GRENOBLE ALPES	Examinatrice
IRYNA VAVILOVA PROFESSEUR, Main Astronomical Observ NAS of Ukraine	Rapporteure

Invités :

JIHANE MOULTAKA
ASTRONOME ADJOINT, UNIVERSITE TOULOUSE 3 - PAUL SABATIER

Université Grenoble Alpes

Institut de Planétologie et d'Astrophysique de Grenoble
Equipe SHERPAS

Astrophysique et milieux dilués

Unsupervised classification of large samples of galaxy spectra

Julien Dubois

1. Reviewer Irina Vavilova
Department of Astronomy and Space Physics
Taras Shevchenko National University of Kyiv

2. Reviewer Stéphane Arnouts
LAM
CNRS, Aix Marseille Université

Supervisors Didier Fraix-Burnet and Jihane Moultaka

September 05, 2023

Julien Dubois

Unsupervised classification of large samples of galaxy spectra

Astrophysique et milieux dilués, September 05, 2023
Reviewers: Irina Vavilova and Stéphane Arnouts
Supervisors: Didier Fraix-Burnet and Jihane Moultaka

Université Grenoble-Alpes
Institut de Planétologie et d'Astrophysique de Grenoble
Equipe SHERPAS
414, Rue de la Piscine
38400 St-Martin d'Hères, France

Abstract (English)

In astrophysics, classifications prove to be particularly useful to retrieve the typical archetypes of one's objects of interest and facilitate population analysis. Among others, this applies to galaxies. Soon after their discovery, they were classified in morphological schemes that are still prevalent today due to how they correlate with certain physical properties. However, thanks to numerous technological breakthroughs, it is possible nowadays to tackle galaxy classification through the prism of spectroscopy. Spectra encompass a lot more information and features than morphology and could hence be used to reach a finer and more meaningful classification. This next-to-unexplored yet promising prospect is precisely what my thesis was anchored around.

Contrary to images of galaxies, spectra are not prone to visual classification due to their complexity. Not only that, but the quantity of observations is such that many processes have to be automated; this is where machine learning and statistical methods come into play. During my thesis, I focused on one particular classification algorithm called Fisher-EM and explored its application to this astrophysical problem. This algorithm, which uses a Gaussian mixture model in a discriminant latent subspace, was chosen for its unsupervised nature, and its ability to deal with high-dimensional data, hence making it a perfect fit for the matter.

The first part of my thesis consisted in investigating the power and limitations of Fisher-EM on simulated data. I used a sample of 12 000 optical galaxy spectra that was previously generated with the code CIGALE, and I adapted it for this study. The simulated nature of the data allowed me to draw direct conclusions on the physical relevance of the classifications produced by studying the distribution of the simulation parameters within the classes. I initially focused on noiseless data and concluded that Fisher-EM was able to successfully extract useful physical information within the spectra to build a meaningful classification. Then, I investigated the effect of noise on the classification process by adding artificial noise to the simulated data. I showed that the results remained very robust down to a signal-to-noise ratio (S/N) of 3 and that meaningful discrimination was still retrieved down to an S/N of 1.

Next, I went on to work on observed spectra and focused on 80 000 galaxies of redshift $0.4 < z < 1.2$ from the VIMOS Public Extragalactic Redshift Survey. Observed data is intertwined

with instrumental effects, noise, and contamination by the atmosphere, which all had to be addressed carefully for the best results. The sample was split into 26 subsamples by bins of redshift, and each of these subsamples was classified independently of one another. Links between classes were then retrieved using the k-Nearest Neighbours algorithm, hence creating a tree-like structure revealing the evolutionary pathways of the classes. Three sub-trees emerged, separating quiescent galaxies, from ones with moderate star formation, and others with ongoing intense star-formation events. Finer discrimination is made down the branches, separating, among others, different levels of star-formation rate and stellar masses.

This thesis work highlights a fully automated and data-driven approach capable of providing physically meaningful spectroscopic classifications of galaxies. For the first time, the diversity of optical spectra of galaxies within redshifts $0.4 < z < 1.2$ was mapped in an unsupervised and automated manner, and their spectral evolution was investigated. Together with new spectroscopic observations of very high redshift galaxies that the JWST will provide, this opens novel and promising perspectives to unravel the unanswered questions surrounding the formation and evolution of galaxies.

Abstract (French)

Peu après leur découverte, les galaxies ont été classées selon des schémas morphologiques qui sont toujours d'actualité en raison de leur corrélation avec certaines propriétés physiques. Cependant, grâce à de nombreuses avancées technologiques, il est aujourd'hui possible d'aborder la classification des galaxies à travers le prisme de la spectroscopie. Les spectres contiennent beaucoup plus d'informations que la morphologie et pourraient donc être utilisés pour obtenir une classification plus fine et pertinente. Cette perspective encore inexplorée, mais pourtant prometteuse, est précisément l'objet de ma thèse.

Contrairement aux images de galaxies, les spectres ne se prêtent pas à une classification visuelle en raison de leur complexité. De plus, la quantité de données est telle que de nombreux processus doivent être automatisés ; c'est là que l'apprentissage automatique et les méthodes statistiques entrent en jeu. Au cours de ma thèse, je me suis concentré sur un algorithme de classification appelé Fisher-EM et j'ai exploré son application à ce problème astrophysique. Cet algorithme, qui utilise un modèle de mélange gaussien dans un sous-espace latent discriminant, a été choisi pour sa nature non supervisée et sa capacité à traiter des données de haute dimension, ce qui le rend parfaitement adapté au problème.

La première partie de ma thèse a consisté à étudier la puissance et les limites de Fisher-EM sur des données simulées. J'ai utilisé un échantillon de 12 000 spectres optiques de galaxies, généré précédemment avec le code CIGALE, et je l'ai adapté pour cette étude. La nature simulée des données m'a permis de tirer des conclusions directes sur la pertinence physique des classifications produites en étudiant la distribution des paramètres de simulation au sein des classes. Je me suis d'abord concentré sur les données sans bruit et j'ai conclu que Fisher-EM était capable d'extraire avec succès des informations physiques utiles dans les spectres pour construire une classification pertinente. J'ai ensuite étudié l'effet du bruit sur le processus de classification en ajoutant du bruit artificiel aux données simulées. J'ai montré que les résultats restaient très robustes jusqu'à un rapport signal-sur-bruit (S/N) de 3 et qu'une discrimination significative était toujours obtenue jusqu'à un S/N de 1.

Par la suite, j'ai étudié un échantillon de 80 000 spectres optiques de galaxies de redshift 0,4 $< z <$ 1,2 provenant du VIMOS Public Extragalactic Redshift Survey. Aux données observées se mélangent des effets instrumentaux, du bruit et la contamination par l'atmosphère, qui ont dû être traités avec soin pour optimiser les résultats. L'échantillon a été divisé en 26

sous-échantillons afin de limiter la perte d'information due au redshift, et chacun de ces sous-échantillons a été classé indépendamment des autres. Les liens entre les classes ont ensuite été déterminés à l'aide de l'algorithme k-Nearest Neighbours, créant ainsi une structure arborescente révélant les chemins evolutifs des classes. Trois sous-arbres ont émergé, séparant les galaxies passives, de celles avec une formation stellaire modérée, et d'autres traversant des événements de formation stellaire intense. Une discrimination plus fine est faite le long des branches, séparant notamment différents niveaux de formation d'étoiles et de masses stellaires.

Ce travail de thèse met en évidence une approche entièrement automatisée, capable de produire des classifications spectroscopiques de galaxies physiquement pertinentes. Pour la première fois, la diversité des spectres optiques des galaxies à des redshifts $0.4 < z < 1.2$ a été cartographiée de manière non supervisée et automatisée, et leur évolution au cours de l'histoire de l'Univers a été étudiée. Avec les nouvelles observations à très haut redshift que le JWST fournira, ce travail ouvre des perspectives nouvelles et prometteuses pour élucider les questions restantes atour de la formation et l'évolution des galaxies.

Remerciements

Il me semble important de commencer mes remerciements en notant l'encadrement idéal que m'ont fourni Didier et Jihane, sans quoi je ne serais très certainement pas arrivé à bout de ma thèse. Au-delà d'un encadrement scientifique, vous avez su faire preuve de beaucoup de patience, de gentillesse et de bienveillance envers moi ; je vous en suis très reconnaissant.

Par ailleurs, si j'ai pu mener à bien cette thèse, c'est aussi grâce à tous les membres de l'IPAG qui m'ont accompagné d'une manière ou d'une autre dans cette aventure. Merci à Jean-Philippe, que j'ai eu la chance d'avoir en tant qu'enseignant lors de ma formation à Phelma, sans qui je n'aurais très probablement jamais mis les pieds à l'IPAG ; merci aux membres de l'équipe SHERPAS de m'avoir si gentiment accueilli malgré mon sujet de recherche atypique ; merci à Bruno, qui mène encore aujourd'hui un combat effréné contre Notilus, pour son efficacité et son aide précieuse pour mes missions ; merci à David et Ghislain pour leur assistance en cas de pépin (ou baignade improvisée) informatique. Merci également à ma collaboratrice Gosia ...

Merci à Stéphane Arnoult et Iryna Vavilova d'avoir gentiment accepté le rôle de rapporteur.e et d'avoir fait partie de mon jury de thèse, par ailleurs aussi constitué d'Estelle Moraux, Marc Huertas-Company et Charles Bouveyron, que je remercie également pour leur investissement.

J'ai eu la chance d'effectuer ma thèse dans un environnement idéal, entouré, certes, de montagnes, mais aussi de nombreux doctorants et post-doctorants géniaux. Cela a été extrêmement épanouissant pour moi d'évoluer avec vous. J'ai sincèrement apprécié votre bienveillance, votre empathie, votre richesse d'esprit, votre sens de l'humour et votre diversité. Étant trop nombreux pour tous vous citer ici, je me contenterai de remercier l'équipe du RU et du bassin d'avoir partagé avec moi tous ces somptueux repas proposés par le CROUS ; les valeureux membres de la communauté du "Bonk" pour ces moments solennels ; mes collègues de bureau pour m'avoir supporté, moi et mon café, pendant aussi longtemps ; tous les grimpeurs, occasionnels et réguliers, qui ont partagé des séances d'escalade avec moi ; les botanistes en herbe qui m'ont généreusement fourni boutures et conseils ; les membres réguliers des pauses thé de fin d'après-midi ; les rôlistes, confirmés et néophytes, pour avoir partagé avec moi la découverte de ce monde. Enfin, il me tient tout

particulièrement à cœur de remercier ma compagne d'infortune pour ses mots justes et son soutien précieux durant ces trois années.

Dans cette épopée, j'ai également eu la chance d'être entouré et soutenu par les membres de ma famille, mes proches et mes amis. Ma réussite découle en grande partie de l'environnement dans lequel j'ai grandi et évolué pour en arriver là, ainsi que des personnes qui m'ont accompagné jusqu'ici. Je remercie en particulier mes parents, ma sœur et Coralie pour avoir été tant là pour moi ; mes amis de longue date, Adrien, Nicolas, Kévin, Johan et Raphaël, qui restent à mes côtés malgré les années ; mon partenaire de Pomodoro et d'échecs Alexandre ; Servan, mon modèle pianistique et cycliste ; mon ami et grimpeur intrépide Mathieu ; mes amis d'école Robin, Sébastien, et Thomas ; mon amie et ... Océane ; mes amis et compagnons de grimpe Alice et Thomas. ...

Contents

Préface – Résumé français

Introduction

Bien qu'il semble aujourd'hui impossible d'envisager l'Univers sans les galaxies, leur découverte ne remonte qu'à un siècle. Auparavant, la distinction entre les nébuleuses galactiques et les "nébuleuses extragalactiques" n'était pas encore établie, et l'on pensait que l'Univers se limitait essentiellement à notre propre galaxie, la *Voie lactée*. Ainsi, la prise de conscience que certaines des nébuleuses observées étaient en fait situées en dehors de notre galaxie a complètement bouleversé notre conception de l'Univers et de sa taille.

Depuis, nous savons l'Univers peuplé de milliards de galaxies, ces larges structures liant gravitationnellement étoiles, gaz interstellaire, poussière et matière noire. Ces structures, dont le diamètre est typiquement de quelques dizaines de milliers de parsecs, contiennent des milliards d'étoiles, et sont aussi nombreuses dans l'Univers qu'elles sont diverses, présentant des formes et des structures variées ainsi qu'une large panoplie de caractéristiques spectrales.

En 1936, Hubble publia "The Realm of Nebulae" (Hubble, 1936), dans lequel il exposa la nécessité de classer les galaxies, et proposa alors la célèbre *séquence de Hubble*. Dans cet œuvre, il affirme : "La première étape consiste évidemment à étudier les caractéristiques apparentes des systèmes étudiés. Les nébuleuses peuvent être membres d'une même famille ou représenter un mélange d'objets de nature totalement différente. [...] Les nébuleuses sont si nombreuses qu'elles ne peuvent pas toutes être étudiées individuellement. Il est donc nécessaire de savoir s'il est possible de constituer un bon échantillon à partir des objets les plus visibles et, dans l'affirmative, quelle est la taille de l'échantillon nécessaire. La réponse à cette question, et à beaucoup d'autres, se trouve dans la classification des nébuleuses". En effet, en astrophysique comme dans de nombreux autres domaines, la classification est essentielle. Les exoplanètes, par exemple, sont classées en différents types en fonction de leur masse, de leurs paramètres orbitaux et de leur composition ; les étoiles, elles, sont généralement classées en fonction de leur type spectral, et des sous-populations sont distinguées sur le diagramme de Hertzsprung-Russell, mettant en évidence les différents

stades de l'évolution stellaire. En sommes, l'immensité de l'univers et l'innombrabilité de son contenu font qu'il est essentiel de rassembler les objets astrophysiques (et notamment des galaxies !) en groupes et d'étudier les propriétés inhérentes à ces groupes.

Cette thèse étant rédigée en anglais, cette préface propose un résumé concis et en français du travail réalisé, et présenté plus en détail dans le reste du manuscrit. Dans un premier temps, dans la Sect. 1 est présenté un court résumé des travaux déjà existants de classifications des galaxies. Cette section met notamment en évidence le point de départ de cette thèse, à savoir la nécessité de développer une nouvelle classification sur la base des spectres. Après cela, la Sect. 2 aborde le sujet des méthodes automatiques de classification —point central de ce travail de thèse—, et présente notamment l'algorithme utilisé pour ce projet. Par la suite, les sections Sect. 3 et Sect. 4 présentent les résultats principaux obtenus, respectivement sur des données simulées, puis des données réelles. Enfin, une brève conclusion du travail réalisé est présentée en Sect. 5.

1. Classification des galaxies

Historiquement, les galaxies ont d'abord été classées en catégories morphologiques. Ces classifications ont été élaborées à une époque où les astronomes n'avaient essentiellement accès qu'à des clichés des galaxies ; les premières classifications ont donc été élaborées à partir d'une inspection visuelle des caractéristiques morphologiques apparentes des galaxies. En outre, ces classifications ont été établies à l'origine sur quelques centaines d'images de galaxies de l'Univers local, et ont été améliorées par la suite.

La première —et probablement la plus connue— classification morphologique de galaxies est la séquence de Hubble (Hubble, 1926, Hubble, 1936 ; Fig. 0.1). Cette classification se compose de quatre types principaux :

- **Elliptiques** (E0 à E7) – Ces galaxies ont, comme leur nom l'indique, une forme elliptique. Elles sont caractérisées par leur ellipticité $e = (a-b)/a$, avec a et b les axes de l'ellipse. Les ellipticités observées sont comprises entre 0.0 et 0.7, correspondant à la nomenclature allant de E0 à E7.
- **Spirales** (Sa à Sc, et SBa à SBc) – Ces galaxies ont des caractéristiques morphologiques plus complexes que les elliptiques, et sont caractérisées par un bulbe central, un disque, et des bras spiraux. On distingue les spirales (S) et les spirales barrées (SB). Une séparation supplémentaire est faite sur la base de la taille du bulbe et des bras. Les galaxies des classes Sa et SBa ont généralement un bulbe important et des bras compacts, tandis que les classes Sc et SBc se caractérisent par un bulbe plus petit, ainsi que par des bras ouverts et davantage parsemés.

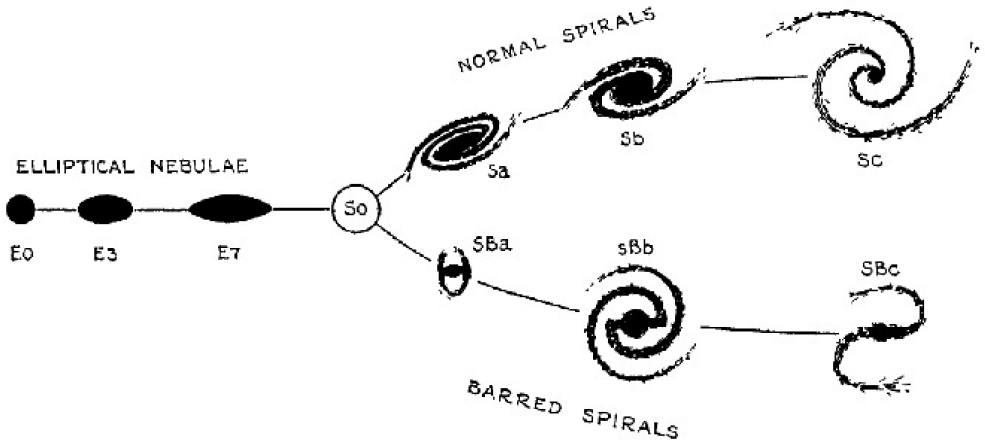


Fig. 0.1.: La séquence de Hubble, telle qu'initialement publiée dans Hubble (1936)

- **Lenticulaires** (S0) – Lorsqu'elles ont été introduites par Hubble, les galaxies lenticulaires étaient considérées comme un état de transition nécessaire entre les galaxies elliptiques et les galaxies spirales. Elles sont ainsi caractérisées par un bulbe et un disque, mais pas de bras spiraux.
- **Irrégulières** (Irr) – Elles sont caractérisées par une structure irrégulière et quelque peu chaotique, qui ne correspond à aucun des types précédents.

Au fil des années, la classification de Hubble fut améliorée, et de nouvelles classifications morphologiques virent le jour (e.g.). À l'heure actuelle, ces classifications sont encore communément utilisées, notamment, car elles s'avèrent être plus ou moins bien corrélées avec certaines caractéristiques physiques telles que la luminosité, la couleur, la masse et le taux de formation d'étoiles (e.g.). Cependant, les classifications morphologiques sont, par nature, limitées : puisqu'elles ne sont construites que sur la base de quelques caractéristiques géométriques et structurelles, elles ne peuvent naturellement pas refléter toute la diversité et complexité physique qui caractérisent une galaxie. Le caractère informatif d'une classification découle directement des critères utilisés pour la construire, et si l'on souhaite obtenir une classification physique des galaxies, leur morphologie n'est peut-être pas le meilleur critère à utiliser. Une alternative évidente —mais difficile à mettre en place— est de baser leur classification sur leurs spectres. Ces derniers sont directement définis par la physique des galaxies : ils sont fonction de l'histoire de formation stellaire, de l'age, de la métallicité, de l'activité du noyau galactique, de l'atténuation interstellaire, etc. Les spectres sont indéniablement une source extrêmement riche d'information, et, exploités correctement, ils peuvent donner lieu à une classification très informative d'un point de vue physique.

2. Une approche automatisée et non supervisée

1 Galaxy classification

1.1 Introduction

Galaxies are complex gravitationally bound structures made of stars, interstellar gas, dust, and dark matter. They typically are tens of thousands of parsecs in diameter, and contain tens of billions of stars. They are as numerous in the Universe as they are diverse, displaying various shapes and structures as well as spectral features.

Although it seems impossible to envision the field of astrophysics without galaxies nowadays, their discovery only dates back to a century. Before that, the distinction between galactic nebulae and 'extragalactic nebulae' was not yet made, and the Universe was thought to be essentially limited to the Milky-Way. Thus, the realisation that some of the observed nebulae were in fact located outside our Galaxy completely reshaped our conception of the Universe and its size.

Soon after the settlement of the *Great Debate*, Hubble published *The Realm of Nebulae* (Hubble, 1936), in which he exposes the need to classify galaxies and proposes the well known Hubble sequence, he states: "*The first step is obviously a study of the apparent features of the systems under investigations. The nebulae might be members of a single family or they might represent a mixture of utterly different kinds of objects. [...] The nebulae are so common that cannot all be studied individually. Therefore, it is necessary to know whether a fair sample can be assembled from the more conspicuous objects and, if so, the size of the sample required. The answer to this question, and to many others, is*

sought in the classification of nebulae.". In fact, taxonomy is essential. Exoplanets are categorised in different types based on their mass, orbital parameters and composition; stars are commonly classified based on their spectral type, and subpopulations are distinguished on the Hertzsprung–Russell diagram, highlighting different stages of stellar evolution. The fact is, the immensity of the universe and the innumerability of its content makes it essential to gather astrophysical objects in groups and study the inherent properties of such groups.

In this first chapter, I provide a concise overview of the history of galaxy classification, starting in Sect. 1.2 with morphological classifications, their use and limitations. Then, in Sect. 1.3, I discuss alternatives to a morphological approach, and conclude on the need for spectroscopic classifications of galaxies.

1.2 Morphological classification of galaxies

Historically, galaxies were first and foremost classified into morphological categories. These classifications were developed at a time when astronomers virtually only had access to photometric plates, and the first classification schemes were thus built from visual inspection of the apparent morphological characteristics of galaxies. Additionally, these classifications were originally built on a few hundreds images of galaxies within the Local Universe, and were later improved.

In this section, I present a description of the original "Hubble sequence", and the subsequent works of classification that it led to. This is not an exhaustive review by any means, but rather an overview of the genesis of galaxy classification. While morphology is not involved per se in my thesis work, which focuses on spectroscopy instead, it is obviously a staple aspect of galaxy classification which has to be addressed.

1.2.1 The Hubble sequence

Hubble originally introduced his morphological classification in Hubble (1926), which he then extended and organised in a sequence (Hubble, 1936; Fig. 1.1). The scheme uses various morphological characteristics (bulge, disk, central bar, spiral arms), and is composed of 4 following main types:

- **Elliptical** (E0 to E7) – These galaxies have, as their name suggests, an elliptical shape. They are characterised by their ellipticity $e = (a - b)/a$, with a and b the major and minor axis of the ellipse. Ellipticities are observed to range from 0.0 to 0.7, hence leading to the nomenclature E0 to E7.
- **Spiral** (Sa to Sc, and SBa to SBc) – These galaxies have more complex morphological features than ellipticals, and are characterised by a central bulge, a disk, and spiral

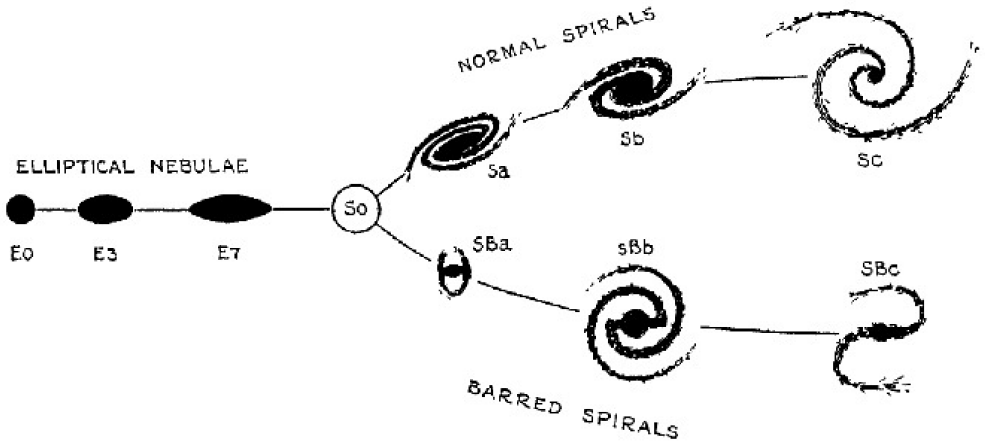


Fig. 1.1.: The "Hubble sequence", also called "Hubble's tuning-fork diagram", published in Hubble (1936)

arms. A distinction is made between normal spirals (S) and barred spirals (SB). Further separation is made on the basis of the size of the bulge and arms. Galaxies in classes Sa and SBa typically have a large bulge and compact arms, while Sc and SBc are characterised by a smaller bulge, as well as open and patchy arms.

- **Lenticular** (S0) – When introduced by Hubble, lenticular galaxies were thought to be a necessary transition state between elliptical and spiral galaxies. They are characterised by a bulge and a disk, but no spiral arms.
- **Irregular** (Irr) – They are characterised by their patchy and somewhat chaotic structure, which do not fit in any of the previous types.

Originally, the "Hubble sequence" was thought to describe the evolution path of galaxies, which supposedly formed as elliptical, and evolved into more complex structure along the sequence. Even though this has been disproved, galaxies are still commonly called "early-type" (elliptical and lenticular) or "late-type" (spiral) for this reason.

1.2.2 Other morphological classification schemes

Over the years, the Hubble classification system was improved, most notably by Vaucouleurs (1959), Sandage (1961), Sandage and Tammann (1981), and Sandage and Bedke (1994). The Sc galaxies were sub-divided into the Sc and Sd types; the existence of the S0 galaxies was confirmed by new observations; and new morphological classifications schemes were also devised.

The de Vaucouleurs' system (Vaucouleurs, 1959) provides an extended version of the Hubble sequence, which is often represented as a 3-dimensional classification scheme. The first

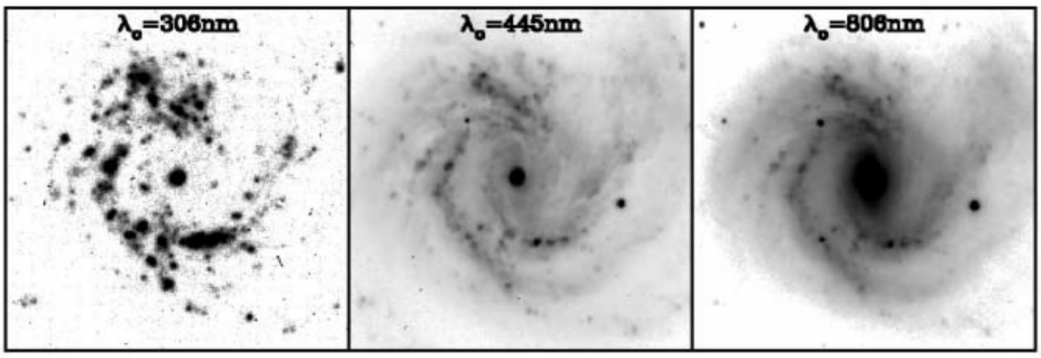


Fig. 1.2.: Illustration of how the appearance of a galaxy changes with wavelength. The images show the galaxy NGC 4303 observed in UV (left), B-band (middle), and I-band (right). Figure taken from Sheth et al. (2003).

dimension is the Hubble type (from early to late type), the second describes the presence (B) or absence (A) of bar, and the third, the presence of spiral arms (s) or rings (r). Compared to the Hubble scheme, this classification an additional layer of division to better reflect the diversity of features seen in spiral galaxies.

Some also noted that the characteristics of the spiral arms should be taken into consideration (e.g. Reynolds, 1927), which led to the development of a refined classification of spiral arms by van den Bergh (Bergh, 1960c; Bergh, 1960a; Bergh, 1960b) and Elmegreen & Elmegreen (Elmegreen and Elmegreen, 1982; Elmegreen and Elmegreen, 1987).

Lastly, another classification that can be mentioned is Morgan's classification scheme (Morgan, 1958; Morgan, 1959), or also called Yerke's classification scheme, which uses the central light concentration, the morphological type, and the inclination angle of the galaxy.

1.2.3 Usefulness and limitations

Nowadays, The Hubble sequence remains widely used in the field of extragalactic astronomy, mainly due to the fact that morphological types correlate well with various physical properties of galaxies, such as luminosity, colour, mass, as well as star formation rate (e.g. Kennicutt, 1992; Roberts and Haynes, 1994; Buta et al., 1994; Strateva et al., 2001).

Nevertheless, morphological classifications face significant limitations. Such classifications remain somewhat subjective; one galaxy may be attributed different types by two different observers. Additionally, the appearance of a galaxy significantly depends on the wavelength at which it is observed, and the morphological features may differ at one wavelength and another. For instance, Fig. 1.2 illustrates how the bar of the NGC 4303 galaxy is essentially invisible in the UV, faint in B-band, and clearly visible in I-band. This approach also

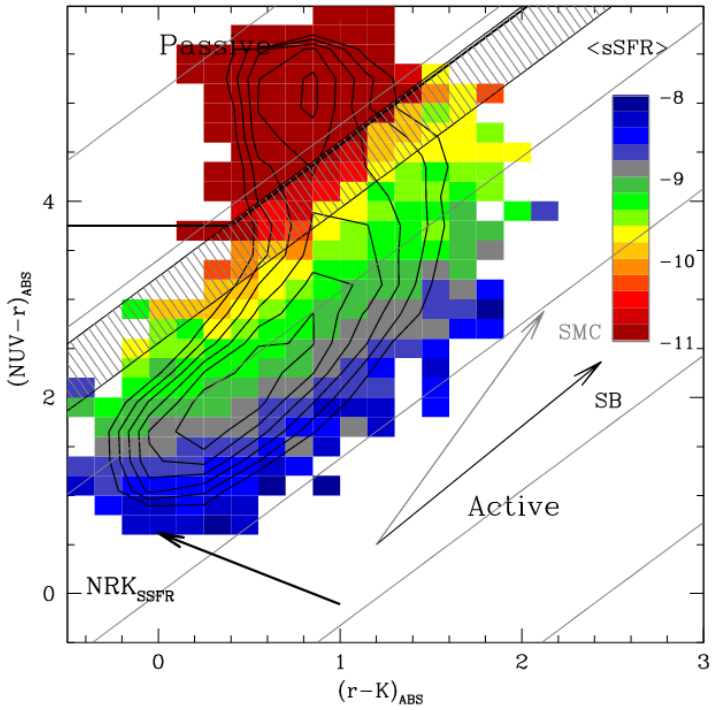


Fig. 1.3.: Distribution of galaxies from the COSMOS sample in the NUVrK diagram (black contour lines), with the corresponding mean value of specific star formation rate, as derived from spectral energy distribution fitting techniques. Figure taken from Arnouts et al. (2013).

encounters challenges due to orientation effects, where a galaxy's appearance can significantly differ based on whether it is viewed edge-on, face-on, or in between. Furthermore, visual classifications tend to be less reliable for faint or distant galaxies, where shapes and structures are not so well resolved and defined.

In the end, the main issue is that these classifications are based on subjective views of two-dimensional images, which only indirectly reflect their true physical properties. Instead, we would ideally want a classification that is directly built on of these intrinsic physical characteristics, should this be feasible.

1.3 Beyond morphology

Morphological classification of galaxies have clearly been extremely impactful on extragalactic astrophysics. It is still an active field of research, especially with the advent of machine-learning, which can be used to automatically classify large samples of galaxies (e.g. Shamir, 2009; Huertas-Company et al., 2015; Schutter and Shamir, 2015; Cheng et al., 2021; Bom et al., 2021; Walmsley et al., 2022; Vavilova et al., 2021; Khramtsov et al., 2022; Vavilova et al., 2022; Fraix-Burnet, 2023). Nevertheless, morphology cannot solely depict all the diversity in the galaxies' physical characteristics, and there is, in that sense, a need to go beyond morphological classifications.

1.3.1 Parametric classification

Various properties such as photometric parameters, spectral features or stellar mass have proven to be useful attributes to separate different populations of galaxies (e.g. Kauffmann et al., 2003; Bell et al., 2004). Rest-frame colour magnitudes are, for instance, used to separate red passive, "green valley" galaxies and blue star-forming galaxies in colour-colour diagrams such as the NUVrK diagram introduced in Arnouts et al. (2013) (Fig. 1.3).

The Baldwin-Phillips-Terlevich (BPT) diagram (Baldwin et al., 1981) is widely used to distinguish star-forming galaxies, Seyfert galaxies, and low-ionisation nuclear emission-line regions (LINERS) using OIII, NII, Hα and Hβ line ratios. However, the application of such traditional diagnostics is limited at redshifts $z > 0.4$ since optical emission lines such as Hα and NII are shifted out in the near-infrared. Other diagnostic diagrams such as the Mass-Excitation diagram (Juneau et al., 2011) have been introduced as an alternative for higher redshifts (Fig. 1.4).

In fact, galaxies display bimodal distributions in a number of parameters, including colours, morphological criterion and spectral features. This is what makes parametric spaces useful tools for classification, especially at higher redshifts, where classical morphological classification typically struggle. Nonetheless, these methods can hardly distinguish more than a few classes, and it goes to say that this cannot realistically encompass all the diversity of galaxies in the Universe. Arguably, the main criticism that can be made against these 2D parameter-space methods is that they only make use of a few numbers of parameters, which cannot be enough to reflect all the physics of the galaxies. But, just like morphological classifications are benefitting from the development of automated classification methods, the latter has opened up the possibility of building classifications on multi-parameter spaces (e.g. Siudek et al., 2018a; Siudek et al., 2018b; Chattopadhyay et al., 2019). This amounts to bringing more discriminant information to the table, hence leading to a richer, more meaningful, resulting classification.

1.3.2 Spectroscopic classification

Multi-parameter classification allow the combination of multiple pieces of information to build the classes. However, it begs the question of the parameter choice; if bringing more and more parameters on the table equates to bringing more and more relevant information, one should aim to combine as many parameters as possible. Arguably, there may be some redundancies between some parameters, but nothing that modern statistical tools cannot tackle. Alternatively, the following point can be made: if adding more and more information increases the richness of the classification, should we have appropriate methods to deal with it, then we might as well base the classification on the galaxies' spectra. Most certainly, a

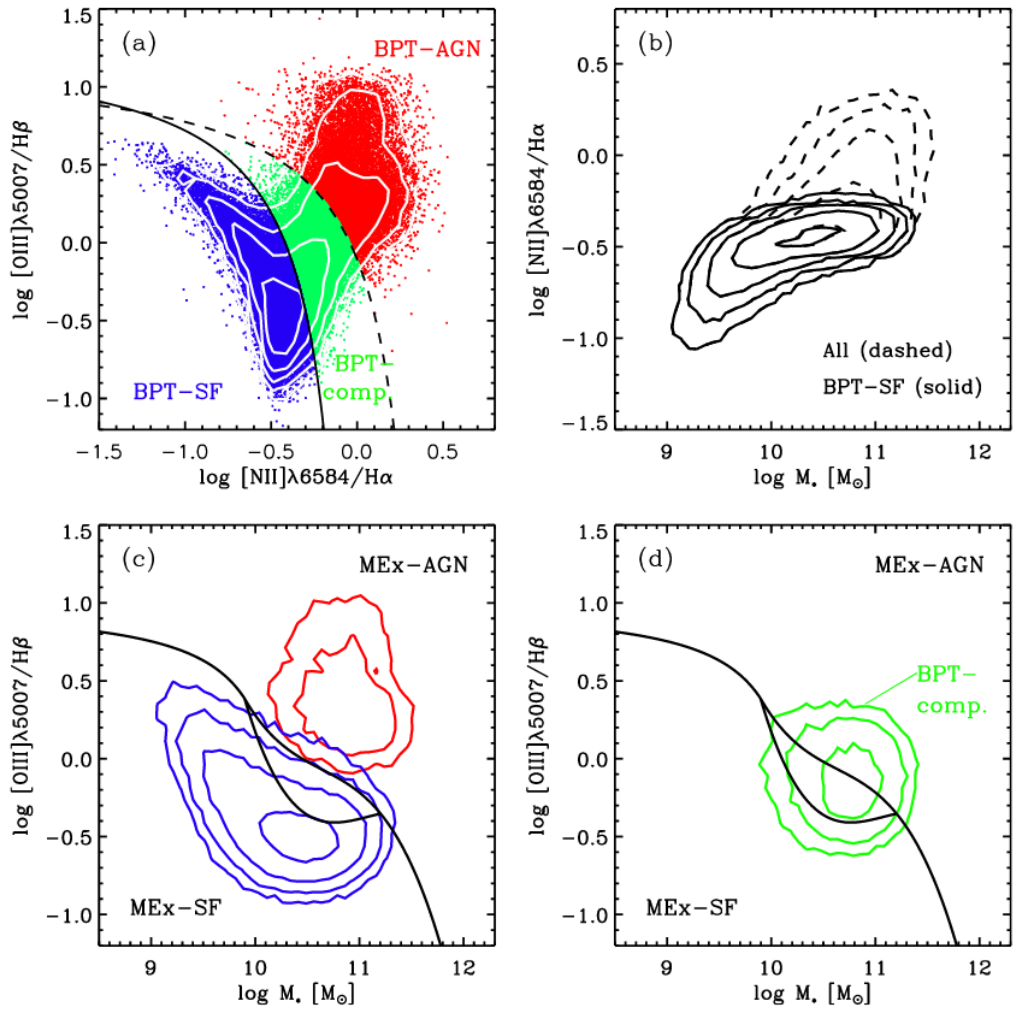


Fig. 1.4.: (a): Illustration of the BPT diagram. (b): Stellar mass versus log([NII]/Hα), which illustrates how stellar mass can be used as a substitute of the NII and Hα line ratio at higher redshifts. (c) and (d): Resulting MEx diagram with the distribution of the AGN, SF and composite classes. Figure taken from Juneau et al. (2011).

tremendous amount of information about the physics of the galaxies is encompassed within the spectra, more than any multi-parameter space can ever achieve.

This idea is obviously not new, but up until recently, it was mostly limited by the difficulty of the task, which cannot realistically be accomplished manually through visual inspection due to the vast diversity and complexity of a galaxy's spectrum. Perhaps the first building block towards a spectroscopic classification of galaxies is the work of Kennicutt (1992), in which the author provides an atlas of galaxy spectra, categorised by morphological type. This work does not introduce a new classification scheme per se, since it uses morphological types instead, but it highlights the richness of a galaxy's spectrum and the contribution it can have to galaxy classification.

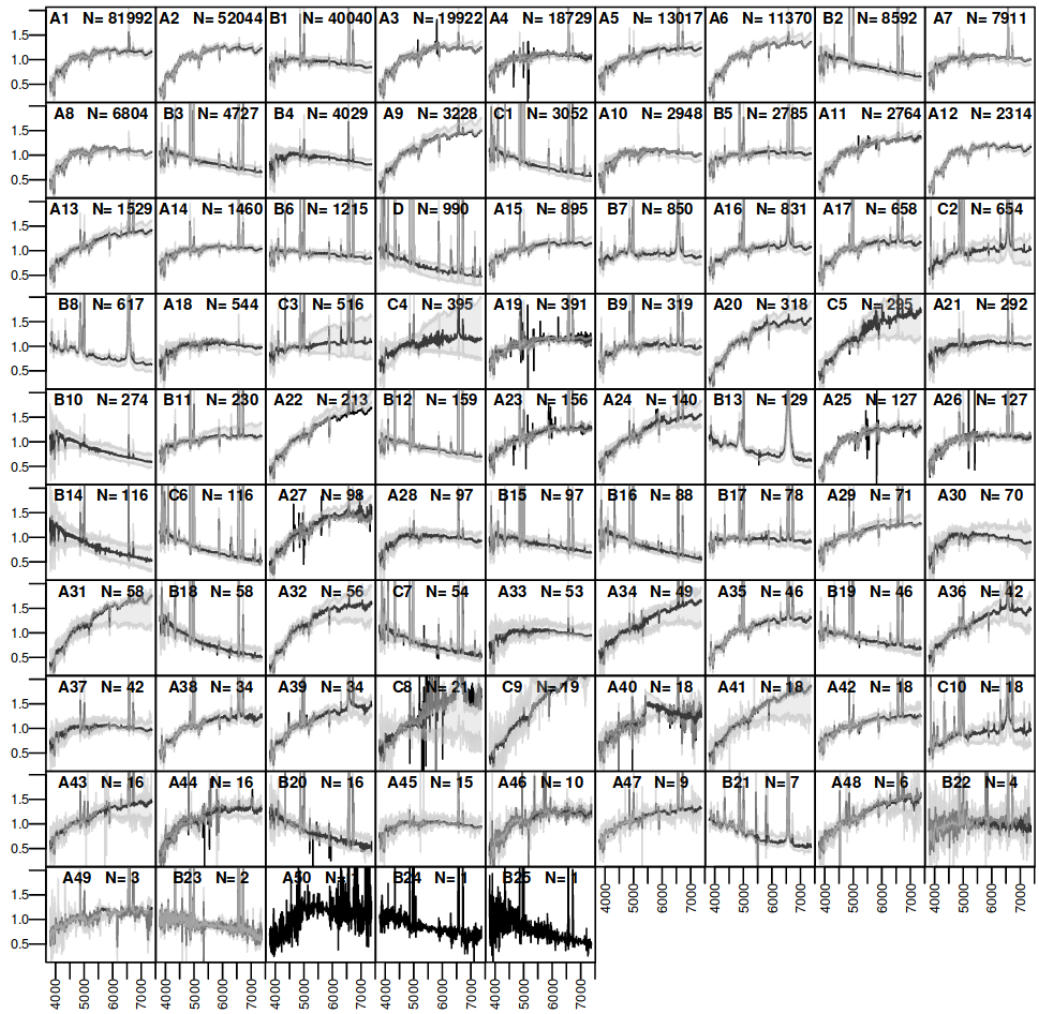


Fig. 1.5.: Spectroscopic classification of $z < 0.25$ galaxies from the SDSS (Fraix-Burnet et al., 2021). Each panel show a class, whose name is displayed in the upper-left corner and size in the upper-right corner. The black lines show the mean spectrum of the class, and the grey area the 10% and 90% quantiles.

It was not until the advent of modern statistical methods that significant progress was made in the quest towards a spectroscopy-based classification of galaxies. In Dobos et al. (2012), the authors build a template of stacked galaxy spectra using the Sloan Digital Sky Survey (SDSS). However, once again, they do not build their classification on the spectra directly, but instead use a combination of parameters (colour, morphology, nuclear activity and spectral lines strengths) inspired from the work of Lee et al. (2008). Similarly, Wang et al. (2018) propose an atlas of composite spectra of galaxies, but base their classification on spectral features and diagnostic diagrams.

In fact, basing a classification on "raw" spectra is a challenging task due to the large amount of parameters (i.e. the monochromatic fluxes), which requires the use of advanced machine-learning techniques and statistical methods specialised in performing such tasks (Fraix-Burnet et al., 2015). However, the choice of the classification algorithm is absolutely

crucial since its role is fundamental in the process. One of the main risks lies in the fact that most classification algorithms will be capable of providing a classification no matter what; however, only a few of them are appropriately designed to tackle high-dimensional data and produce relevant results with physical meaning. For instance, in Sánchez Almeida et al. (2010) and Sánchez Almeida and Allende Prieto (2013), the authors propose a spectroscopy-based classification of galaxies, but they use the k-means algorithm, which is unfortunately not well suited to tackle high-dimensional data and large sample sizes, as argued by De et al. (2016). More recently, the same task was tackled in Fraix-Burnet et al. (2021), but with an algorithm that is specifically designed for high-dimensional clustering. Their work led to an 86-class spectroscopic classification obtained on a sample of close to one million spectra (Fig. 1.5).

1.4 Aim of this work

I have shown in this chapter that the field of spectroscopic classification of galaxies is in its infancy, but it has the potential of leading to a much more physically meaningful end result than morphology. It is precisely in this context that my work was anchored. However, the complexity of the task requires the use of an automated classification method, which, as we have seen, has to be chosen carefully; this will be the subject of Chapter 2.

The contribution of this work is two-fold. First, thanks to a sample of simulated spectra, I investigated the capacity of the chosen classification algorithm to automatically produce, on the sole basis of their spectra, classes of galaxies sharing common physical characteristics. I simultaneously investigated the robustness of the method, and its behaviour when different levels of noise are added to the data. This piece of work constitutes the Chapter 3 of this thesis and has been published (Dubois et al., 2022).

Lastly, I studied and classified a sample of 80 000 galaxies of intermediate redshifts (0.4–1.2). This is essentially the extension of the work of Fraix-Burnet et al. (2021), which focuses on nearby galaxies ($z < 0.25$). This part of my work is discussed in Chapter 4, and is the subject of a forthcoming paper that is in preparation.

2 Automated classification techniques

2.1 Introduction

With the advent of technology and telecommunications, science now rhymes with data, and so does astrophysics. The James Webb Space Telescope (JWST), launched in late 2021 and operational since July 2022, was designed to gather and transmit back to Earth up to 34 GB of data every day (Johns et al., 2008). To put things into perspective, my personal computer can store up to 1 TB of data, and would thus be full within one month of receiving raw data from the JWST. With an estimated operational duration of 10 years, it will have produced enough data to fill up my computer more than a hundred times; and this is just one telescope, sending in data from outer space, that is. The Vera C. Rubin Observatory will conduct the Legacy Survey of Space and Time (LSST), expected to start in 2024, and

will produce an overwhelming 20 TB of raw data per night. Within 10 years of operation, it will have produced up to 70 PB of data: 70 000 times the current storage capacity of my personal computer. Suffice it to say, astrophysicists nowadays have no choice but to learn to efficiently store, manipulate, process and analyse data to make the most out of all the information we are gathering about the universe.

This thesis, as it tackles the complex problem of galaxy classification, is fully anchored in this context. Automated methods are absolutely crucial for modern classifications of galaxies. This chapter is here to provide an overview of the statistical tools commonly used for classification purposes. We will start by having a look at some elements of formalism in Sect. 2.2 before discussing some clustering methods. We will distinguish two types of algorithms: supervised (Sect. 2.3), and unsupervised (Sect. 2.4). Some of the methods presented are linked to my work and will thus be described more thoroughly. Some others were deemed important enough to the field not to omit them (e.g. k-means, neural networks). Finally, the last section will be specifically dedicated to Fisher-EM, the method we chose to use for this project (Sect. 2.5).

Two books helped me immensely write this chapter: Hastie et al. (2009) and Feigelson and Babu (2012). The former provides a quite rigorous and mathematical description of state of the art statistical learning techniques, while the latter focuses on applications to astrophysical problems. Of course, they both cover a lot more than what is discussed in this chapter, and I strongly encourage the reader to leaf through them, should they be interested in statistics and possible applications to their field of research.

2.2 The fundamentals

2.2.1 Formalism

Let us consider n observations $\{x_1, ..., x_n\}$, whatever they are, with p variables each. In the particular context of this thesis, the observations consist of galaxy spectra, so the p variables are simply the p monochromatic fluxes that make up the complete spectrum, and thus, $x_i \in \mathbf{R}^p$. But they can be anything, ranging from images, to parameter vectors, or sets of photometric measurements. The only requirement is for them to be written as p-dimensional vectors. Generally, a 2D image is written as a matrix, but it can easily be flattened into a single vector to meet this requirement, and in this case the p variables would be all the individual pixels.

Within the total population $\{x_1, ..., x_n\}$ very often lie several subpopulations, distinct from each other, with characteristic properties. We call these subpopulations, classes or clusters. The concept of clustering or classification simply consists of recognizing or finding the

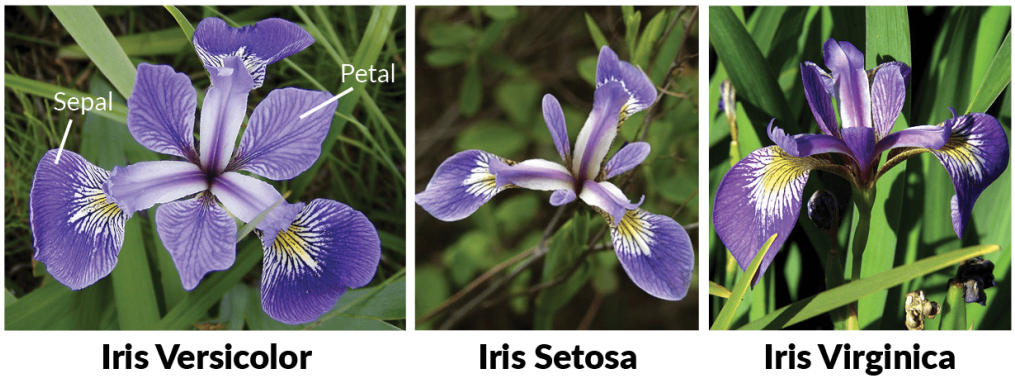


Fig. 2.1.: Pictures of *iris setosa* (first panel), *iris versicolor* (second panel), and *iris virginica* (third panel). This figure is taken from https://www.datacamp.com/tutorial/machine-learning-in-r

.

clusters within a dataset, and attributing each observation to the correct cluster. To put it another way, clustering algorithms are designed to produce K partitions $\{C_1, ..., C_K\}$ of $\{x_1, ..., x_n\}$.

However, not all classification problems are the same, and there are in fact mostly two paradigms. Sometimes, we have a-priori knowledge of the clusters. You may, for example, be working with a dataset containing images of either cats or dogs. You know for a fact there are no horses, dolphins, or unknown mystical creatures in $\{x_1, ..., x_n\}$, and each x_i belongs either to C_1 (dogs) or C_2 (cats). The problem can therefore be simplified to (i) figuring out the typical properties of C_1 and C_2 and (ii) finding out whether each x_i fits the archetypes of C_1 or C_2. Generally, (i) is achieved with a sample of labelled data which we call the training sample, from which we train the algorithm to recognize the features of C_1 and C_2. Based on the information it learned from the training sample, the algorithm is then capable of achieving (ii) and can thus be used to classify the unlabelled dataset. These types of algorithm are called **supervised**, in the sense that we supervise them during the learning process, until they are capable of working on their own and do what we want them to do (Sect. 2.3).

Sometimes, instead, you may find yourself wandering into uncharted territory without much reliable and unbiased a-priori knowledge to train your classification algorithm onto. You are not sure whether to expect cats, dogs, or bears in your dataset. It is unclear whether you are working with images of resolved stellar surfaces or chorizo slices[1]. But you know for a fact you do not want to inject wrong information in your algorithm, in which case the safest bet is not to assume anything, and let the algorithm do the investigation on its own. This is what

[1] French physicist and philosopher E. Klein posted a picture of a chorizo slice on Twitter pretending it was a resolved image of the surface of Proxima Centauri taken by the JWST, which surprisingly deceived many people. Source: https://twitter.com/EtienneKlein/status/1553765864553472003

we call **unsupervised** classification, and some algorithms are specifically designed for such a task (Sect. 2.4).

To illustrate these concepts, let us consider a simple case of application with the "Iris flower dataset", a well known multivariate set of data first used in Fisher (1936). It is a rather small sample, with low-dimensional data, hence convenient for visualisation and pedagogical purposes. It consists of measurements of 4 morphological properties of flower (sepal length, sepal width, petal length and petal width) for a sample size of 150. There are 3 species of iris flowers in this dataset: *setosa*, *versicolor*, and *virginica* (illustrated in Fig. 2.1), and each observation is labelled with its known corresponding species. In other words, the sample has already been classified into 3 clusters. Because the number of feature is small, we are able to easily visualize the data in the parameter space and figure out the clusters properties with simple visual inspection. I provide in Fig. 2.2 a visualisation done with R[2], the software for statistical computing and data visualisation that I have used throughout my thesis. We can see that the clusters occupy different regions of this parameter space. This sample could be used to train a supervised algorithm to recognise the three species. It would probably pick up that *setosa* tend to have very narrow petals and wide sepals, while *versicolor* typically have petal widths ranging between 1.0 and 1.5 cm, and anything larger than that is very likely to be a *virginica*. Instead, if the dataset was not labelled in the first place, we could use an unsupervised algorithm to analyse the sample. Ideally, the algorithm should figure out on its own that the distribution in the parameter space is trimodal, and retrieve three clusters with good accuracy.

With this example in mind, it may instinctively appear clear that classification algorithms very often rely on a metric defining the distance between objects to distinguish the clusters, which leads us to the next subsection.

2.2.2 Multivariate distances

Looking for clusters in the parameter space amounts to finding groups of objects that are close to each other and far away from the rest of the overall population. But being *far* or *close* only makes sense through the prism of a metric that characterises the distance between objects in the p-dimensional space, i.e. a multivariate distance. Formally speaking, a distance on the set $\mathbf{E}$ is a function d characterised by the following properties:

- $d : \mathbf{E} \times \mathbf{E} \longrightarrow \mathbf{R}$
- $\forall x_i \in \mathbf{E},\ d(x_i, x_i) = 0$
- $\forall (x_i, x_j) \in \mathbf{E}^2 \mid x_i \neq x_j,\ d(x_i, x_j) > 0$
- $\forall (x_i, x_j) \in \mathbf{E}^2,\ d(x_i, x_j) = d(x_j, x_i)$

[2] https://www.r-project.org/

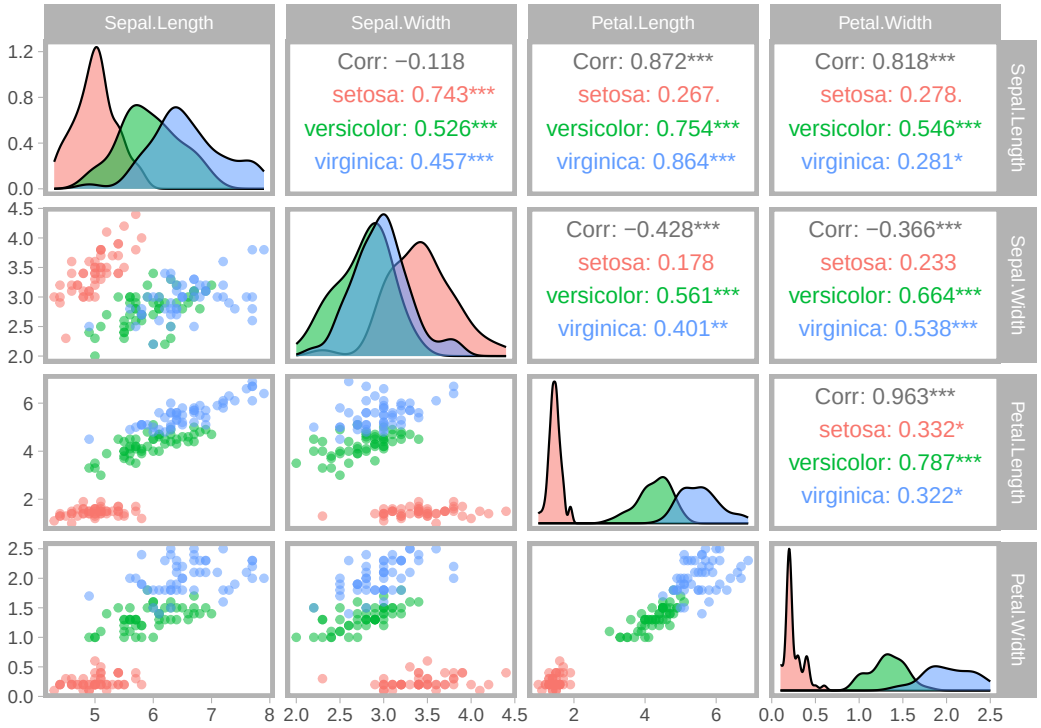


Fig. 2.2.: Visualisation of the Iris flower dataset. The six scatter plots show the distribution of the data in the 2D parameter spaces corresponding to pairs of `{sepal length, sepal width, petal length, petal width}`. The diagonal plots show the distribution of the variables within the species, and the upper panels indicate the general and species-specific correlations. In all the plots, *iris setosa* are displayed in red, *iris versicolor* in green, and *iris virginica* in blue

- $\forall (x_i, x_j, x_l) \in \mathbf{E}^3,\ d(x_i, x_j) \leq d(x_i, x_l) + d(x_l, x_j)$

The most basic and frequently used metric is the Euclidean distance (eq. 2.1):

$$d(x_i, x_j) = \sqrt{\left(\Sigma_{k=1}^{p}(x_{i,k} - x_{j,k})^2\right)} \tag{2.1}$$

Although there is a plethora of metrics available, it would be irrelevant to provide an exhaustive list of distance functions here; the important aspect to grasp, however, is that the distance used should be thoughtfully chosen because the classification results and the conclusions that can be drawn from them may strongly differ from one metric to another (Abu Alfeilat et al., 2019).

The Euclidean distance, like many other metrics, only applies to data where the p features can be compared with one another, but it is not always the case. Say, for instance, the first feature is an apparent magnitude and the second is a radius in parsec. Obviously, using the Euclidean distance on the data-set as is would not make sense, first, from a dimensional standpoint (adding parsec^2 to a unit-less quantity), but also and mostly because both features could have totally different orders of magnitudes. This issue is generally overcome by standardizing the p variables to centre them around zero and to re-scale their distribution to

unit variance, hence making them comparable for distance computing. Distances are used, for instance, in the k-Nearest-Neighbours or the k-means algorithms, presented respectively in Sect 2.3.2 and Sect 2.4.1.

2.2.3 High-dimensional data

As the number of variables p increases, the p-space grows more and more empty, and it becomes increasingly harder to find groupings within it. This, along with several other issues that high-dimensional data suffer from, is called the "curse of dimensionality".

In Beyer et al. (1999), they show that as the dimension increases, the distance of the nearest data point and the distance of the furthest observation from a given point in the p-space become more and more similar. In other words, as counter-intuitive as it may be, in higher dimensions, the p-dimensional space is mostly empty, and there is not much difference in terms of distance from the closest data point and the furthest data: everything is spread apart. They argue that this phenomenon car occur for as little as 10–15 dimensions.

This can be visualised with a simple toy model that I made with R (Fig. 2.3). In this model, for dimensions ranging from 1 to 1000, I simulated 100 data points from a multivariate Gaussian distribution centred around $\mathbf{0} = (0, ..., 0)$ with the identity matrix as the covariance matrix. Then, I computed the evolution of the ratio between the furthest and the closest point from $\mathbf{0}$, and we see that it appears to converge towards 1. In other words, as the dimension increases, the difference between "close" and "far" becomes less and less well-defined.

As a result, when dealing with high-dimensional data, it is often necessary to reduce the dimension before proceeding to the classification. This is called "dimensionality reduction". Obviously, we want to keep most of the discriminative information in the process, or else it would be pointless. The process amounts to finding so-called latent variables, inferred from the data, where the majority of the information lies. They can be simple linear combinations of the p observed variables, or originate from more complex and non-linear models, but they often are non-physical—simply meaningful from a statistical standpoint. Principal component analysis (PCA; Pearson, 1901, Hotelling, 1933), linear discriminant analysis (LDA; Sect. 2.3.1), or autoencoders (Kramer, 1991, Kramer, 1992), for instance, are common techniques used for dimensionality reduction.

The spectra of galaxies involved in my thesis are made of 10^2 to 10^3 monochromatic fluxes, and as a result, definitely belong to the realm of high-dimensional data. It was thus necessary to choose an approach that includes a dimensionality reduction aspect. More details are discussed in Sect. 2.5.

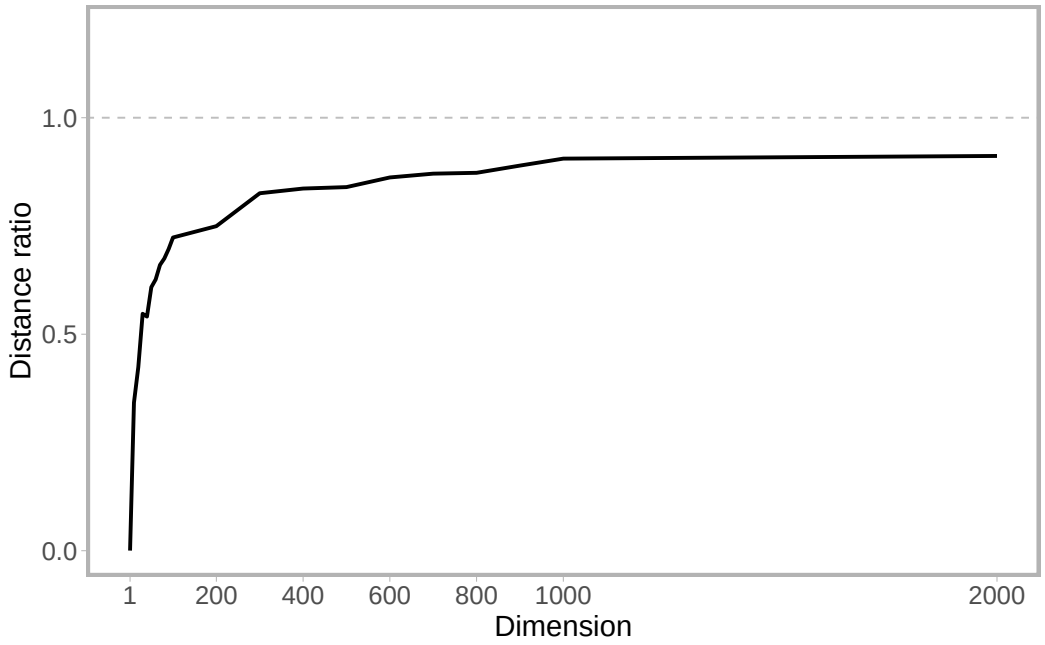


Fig. 2.3.: Ratio of the furthest and closest point to **0** among data generated with a multivariate Gaussian distribution centred around **0** and identity covariance. Datasets were generated at dimension ranging from 1 to 2000.

2.3 Supervised classification

Supervised methods were not suited for the main objective of my PhD project, as they require prior knowledge of what classes are there to expect. Nonetheless, some of these methods remained useful under certain circumstances, and are thus presented here. The linear discriminant analysis (Sect. 2.3.1) was used as a tool to better understand the unsupervised classification results in Chapt. 3. The k-nearest-neighbours method, presented in Sect. 2.3.2, was used in Chapt. 4 to find links between classes of galaxies at different redshifts using distance similarities. Lastly, neural networks are quickly discussed in Sect. 2.3.3. Despite not having used them for my work, I found it difficult to dedicate a section to supervised learning without at least mentioning neural networks, as they have become extremely popular in the last decade.

2.3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a supervised dimension reduction and classification method developed by Fisher (1936) that is still relevant and used nowadays, including in astrophysics (e.g. Leka and Barnes, 2007; Ferrari et al., 2015; Nevin et al., 2023). The key idea behind this is to find a subspace of the p-space that discriminates the clusters as much as possible. The simplest example consists of a two-cluster dataset. In this special case, the

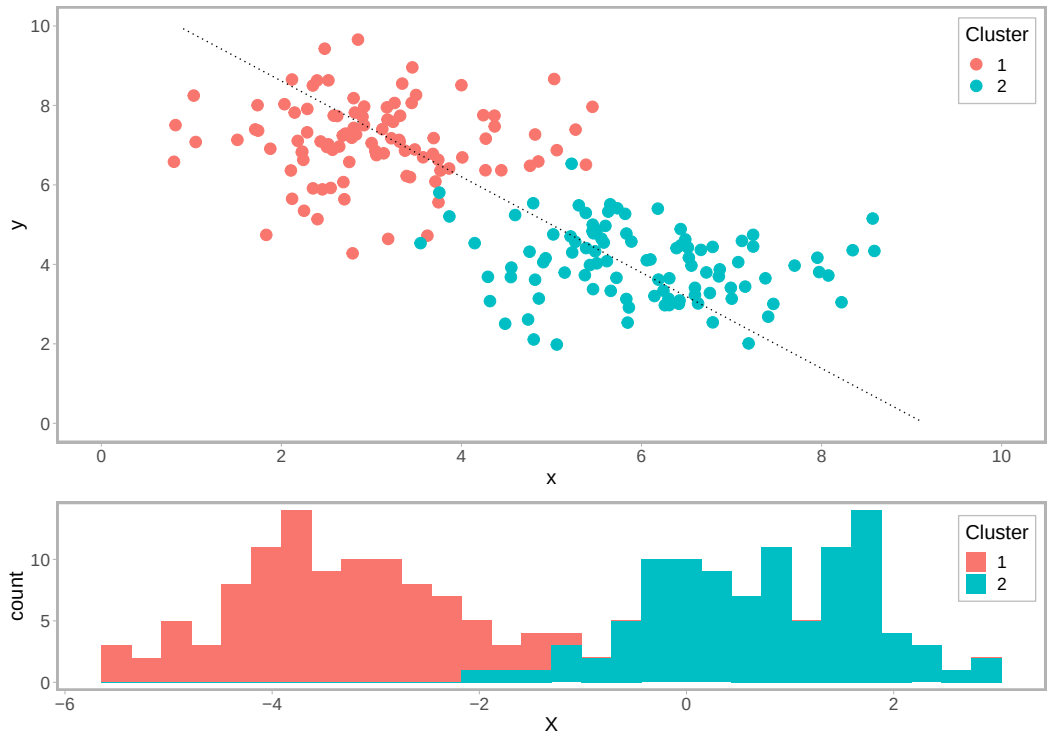


Fig. 2.4.: Illustration of an application of the LDA to a bimodal sample in two-dimensional space. The upper panel shows the scatter plot of the sample in the two-dimensional space, and in dotted line the projection axis resulting from the LDA analysis. The lower panel shows the histogram of the projection of the data on the LDA component.

discriminant subspace has only one dimension and is built from a linear combination of the p-dimensions. The observations x_i can therefore be projected on it as follows (eq. 2.2):

$$X_i = \Sigma_{k=1}^{p} \lambda_k x_{i,k} \tag{2.2}$$

where X_i is the projection (a scalar in this case), and λ_k the linear coefficients. This projection has to be optimized such that it maximizes the discrimination between the two clusters C_1 and C_2. In other words, we aim to find the λ_k-values that most separate the clusters from one another. Mathematically speaking, maximizing the separation amounts to maximizing the ratio of between-class variance to within-class variance. This problem can be solved analytically under the assumption that the populations are distributed normally, and that the covariance matrices of C_1 and C_2 are equal. For problems where the latter assumption cannot be made, an alternative is the Quadratic Discriminant Analysis (QDA).

Once the discriminative linear coefficients are known, a threshold c can be defined such that if $X_i > c$, then $x_i \in C_1$ and otherwise $x_i \in C_2$. This way, new unlabelled data can be classified based on the optimal λ_k that were processed using the labelled observations.

The concept behind the LDA can be easily visualised and understood in two-dimensional space with two clusters (Fig. 2.4), and essentially remains the same when generalised to higher dimensions and greater number of clusters. Higher dimensions will increase the number of λ_k to adjust, but will not affect the dimension of the projection subspace. Instead, the number of clusters K will be the defining factor in that regard; the dimension of the discriminative subspace is $K - 1$. So instead of having a single projection axis as in the $K = 2$ case, there will be an additional axis for every additional cluster.

2.3.2 k-Nearest Neighbours

Among supervised classification techniques, the k-nearest neighbours (k-NN) algorithm, developed by Cover and Hart (1967), is certainly one of the most well known. It is a discriminative technique that labels new data by looking at the k closest labelled observations in the p-dimensional space, and drawing its conclusion based on majority class among the k nearest neighbours (Fig. 2.5). It uses distance considerations in its process, and it is, therefore, necessary to choose an appropriate metric distance for the dataset (see Sect. 2.2.2).

Aside from the distance, k-NN has an obvious free parameter: the number of neighbours k. Choosing a high value for k will tend to smooth out the borders between classes, while a lower value will conserve smaller scale features but will be more sensitive to noise. A balance must be found empirically, and depends on the dataset.

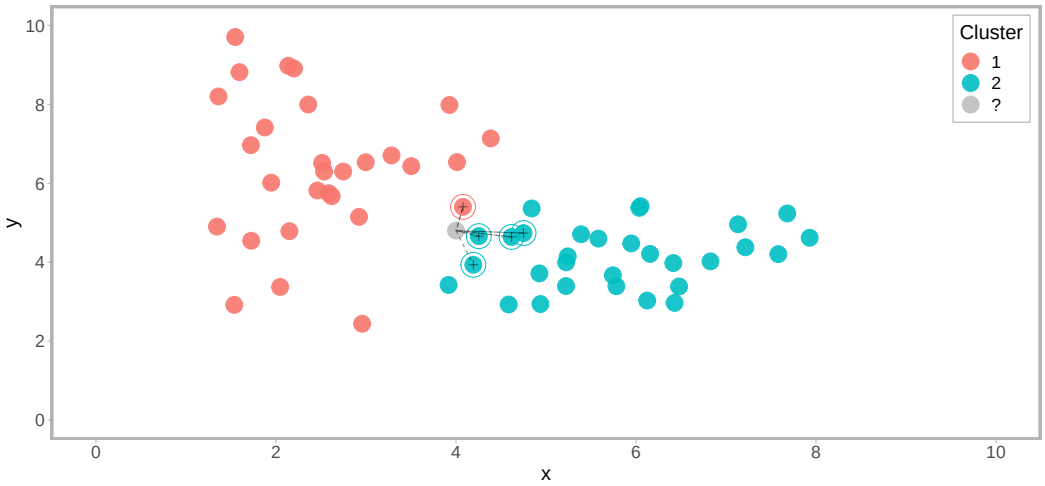


Fig. 2.5.: Illustration of the k-NN algorithm with a bimodal sample in a two-dimensional space. The scatter plot shows the two clusters as well as an unlabelled data point (in grey). The $k = 5$ closest neighbours of this data point are circled, and the grey line represents their Euclidean distance to the unlabelled data. The majority of the nearest neighbours belong to C_2, so the new observation would be labelled as belonging to C_2 too.

2.3.3 Neural networks

Neural networks have become extremely trendy lately, and are virtually used in every field possible, including astrophysics (see Huertas-Company and Lanusse, 2023 for a recent review of the impact of deep-learning on extragalactic astrophysics). They have proven to be powerful tools for certain tasks, but were not suited for this project. Neural networks are inspired from the human brain. They consist of a multi-layer structure made of nodes (neurons) and links (synapses), which are built and trained to perform a given task. The first layer of a neural network usually takes linear combinations of the input data, which are then used by the following layers to model the wanted output as a non-linear function of these inputs. The parameters of the resulting architecture are then adjusted using training data to match the expected behaviour.

They have a reputation of being mysterious or so-called "black-box" models because it can be difficult to really understand what a trained neural network is actually doing to perform the task it was trained to do. And as all supervised methods, they are sensitive to biases in the training data. In an effort to provide ways to better understand the predictions of such models, tools such as LIME (Ribeiro et al., 2016) were developed.

2.4 Unsupervised classification

Supervised methods, and neural networks in particular, are extremely popular due to their flexibility and how fast they are once the model is trained. However, as we showed in the previous section, they are limited to applications where the expected outcomes are known. When this condition is not met, one should consider unsupervised methods instead, and this is precisely why the algorithm used throughout my thesis (Fisher-EM) is unsupervised. In this section, I discuss three unsupervised methods that were used in my work. The first two are the k-means algorithm and hierarchical clustering, which are described respectively in Sect. 2.4.1 and Sect. 2.4.2. These two methods are relevant in my work as they are used as initialisation tools in Fisher-EM. The third method mentioned here is Gaussian Mixture Model (Sect. 2.4.3). This section is particularly important because Fisher-EM uses a similar model. Note that although Fisher-EM is indeed unsupervised, it is discussed in its own specific section (Sect. 2.5).

2.4.1 K-means

A simple approach to classifying a dataset in an unsupervised manner is the K-means method, introduced by MacQueen (1967). First, K points representing the cluster centroids are drawn in the p-space and each observation is attributed to the nearest centroid. Then,

the cluster centroids are updated by computing the mean position of the observations that populate the clusters. Finally, the latter step is repeated until the classification reaches its definitive state.

This algorithm does not demand a lot of resources and is a great tool to quickly explore datasets, but it suffers from significant drawbacks. It is not suited for high-dimensionality nor complex data distributions; as the dimension increases, distance based methods lose in efficiency as objects of the p-space become sparser. It is also sensitive to outliers, as they greatly affect the mean position of the centroids.

On a more pragmatic aspect, choosing the initial position of the centroid as well as the number of clusters is an issue on its own as it will greatly affect the resulting classification, and is not guided by a statistically objective criterion (Hastie et al., 2009).

2.4.2 Hierarchical clustering

Hierarchical clustering is another distance-based approach to unsupervised classification. This method is conceptually different from the others, as it does not partition a dataset into a given number of clusters, but rather builds a tree or dendrogram of similarity. Two paradigms are distinguished: agglomerative and divisive. In agglomerative hierarchical clustering methods, the process consists of starting with n clusters so that each of the n observations has its own unique class, and then iteratively merge together the pair of closest clusters until the whole dataset is contained in a single cluster. Divisive methods, on the other hand, start with a single cluster, which is divided iteratively into more and more clusters until there are as many classes as observations.

Each level of the dendrogram represents a particular classification, and the wanted number of classes can be met by going up or down the tree. It is, however, up to the user to decide which level is the best or "true" classification, assuming it so exists.

2.4.3 Mixture models

Mixture models (MM) are probabilistic models that map the data as a mixture of K distributions, with K being the number of clusters. Although any kind of distribution can be used for MM, Gaussian Mixture Models (GMM) are the most popular. In such models, each cluster is described within the p-space by a multivariate Gaussian probability density function Φ that is parameterised by a p-dimensional mean vector μ_k and a $p \times p$ covariance

matrix Σ_i. Thus, the whole sample can be modelled as the weighted sum of the K Gaussian distributions (eq. 2.3; illustrated in Fig. 2.6.).

$$f(K, \theta) = \sum_{k=1}^{K} \pi_k \Phi(\mu_k, \Sigma_k) \tag{2.3}$$

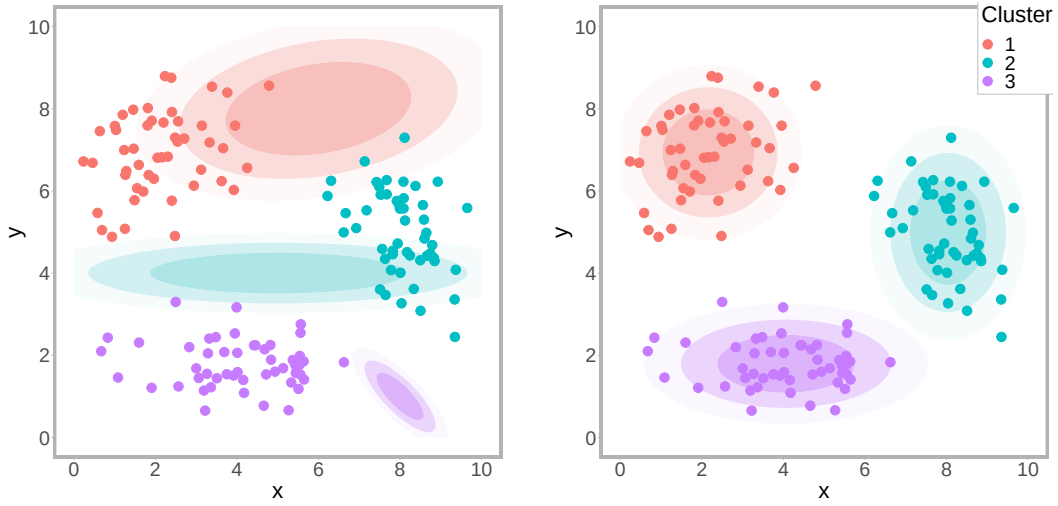
with K the number of cluster and $\theta = (\theta_1, ..., \theta_K)$ the normal distributions parameters, i.e. for each cluster k: its weight π_k, its mean μ_k and its covariance matrix Σ_k. The model has to be adjusted to fit the data, which includes finding the best number of clusters in addition to the best positions (μ_k) and shapes (Σ_k) for each normal distribution in the p-space. This process is achieved through maximum likelihood estimation, and for GMMs this is very often done with the expectation-maximisation (EM) algorithm. It works as follows:

1. **Initialisation:**
   - The user choses the number of clusters K. This parameter is fixed within the optimisation loop: the best model will be found for this given K value. The EM algorithm can be then run several additional times with different values of K to find the optimum number of clusters according to the likelihood.
   - The parameters θ are set to their initial values θ_0. This can be done by choosing θ_0 randomly, or by running a less complex classification algorithm first (e.g. k-means, see Sect. 2.4.1) to get a rough estimation of the clusters.
2. **E-step:**
   - The membership probabilities of all the observations $X = \{x_1, ..., x_n\}$ given the current estimation of θ are computed. At iteration j, the membership probability of x_i to a given cluster C_m is obtained by computing the following ratio (eq. 2.4):

$$P(x_i \in C_m \mid \theta^{(j)}) = \frac{\pi_m \, \Phi(x_i, \mu_m^{(j)}, \Sigma_m^{(j)})}{\sum_{k=1}^{K} \pi_k \, \Phi(x_i, \mu_k^{(j)}, \Sigma_k^{(j)})} \tag{2.4}$$

   - A label is attributed to each observation given the current estimation of their membership probabilities. This is typically done by selecting a probability threshold P_0 such that if $P(x_i \in C_m \mid \theta^{(j)}) > P_0$, then x_i is labelled as belonging in cluster C_m.
3. **M-step:**
   - The model parameters are optimized to maximize the likelihood. For GMMs, it can be solved analytically given the current membership probabilities.
   - The next estimates of the clusters' weights $\{\pi_1, ..., \pi_K\}$ are computed (eq. 2.5).

$$\pi_k^{(j+1)} = \frac{1}{n} \Sigma_{i=1}^{n} P(x_i \in C_m \mid \theta^{(j)}) \tag{2.5}$$



(a) Random initialisation of the GMM.

(b) Best-fit model of the GMM reached after a few iterations.

Fig. 2.6.: Illustration of the convergence of a GMM with the EM algorithm on a trimodal dataset in two-dimensional space that I simulated using R. The ellipses show the Gaussian probability density functions of the three clusters, and the different opacities highlight the 50%, 80% and 95% regions from inner to outer.

- The next estimates of the clusters' mean vectors $\{\mu_1, ..., \mu_K\}$ are computed (eq. 2.6).

$$\mu_k^{(j+1)} = \frac{\Sigma_{i=1}^{n} P(x_i \in C_k \mid \theta^{(j)}) x_i}{\Sigma_{i=1}^{n} P(x_i \in C_k \mid \theta^{(j)})} \tag{2.6}$$

- The next estimates of the clusters' covariance matrices $\{\Sigma_1, ..., \Sigma_K\}$ are computed (eq. 2.7).

$$\Sigma_k^{(j+1)} = \frac{\Sigma_{i=1}^{n} P(x_i \in C_k \mid \theta^{(j)})(x_i - \mu_k^{(j+1)})(x_i - \mu_k^{(j+1)})^\intercal}{\Sigma_{i=1}^{n} P(x_i \in C_k \mid \theta^{(j)})} \tag{2.7}$$

4. **Loop until convergence:**
   - The likelihood function $\mathcal{L}(\theta^{(j+1)} \mid X)$ is computed given the current estimate of the mixture parameters (eq. 2.8).

$$\mathcal{L}(\theta^{(j+1)} \mid X) = \Pi_{i=1}^{n} \Pi_{k=1}^{K} \left[P(x_i \in C_m \mid \theta^{(j+1)}) \right]^{\mathbb{1}(x_i \in C_k)}, \tag{2.8}$$

   where $\mathbb{1}(x_i \in C_k) = 1$ if $x_i \in C_k$ and $\mathbb{1}(x_i \in C_k) = 0$ otherwise.

   - Steps 2. and 3. are repeated until eq. 2.9 is satisfied or until the number of iterations reaches the chosen limit.

$$\mathcal{L}(\theta^{(j+1)} \mid X) < \mathcal{L}(\theta^{(j)} \mid X) + \epsilon, \tag{2.9}$$

   where ϵ is a chosen threshold.

2.5 Fisher-EM

2.5.1 Why Fisher-EM?

Given the scientific objective of the project, it was necessary to choose a classification method that met the following requirements:

(i) **Unsupervised:** supervised methods require a robust training sample, which we do not have for galaxy spectra. Of course, it would be possible to build a training sample based on morphological classification, but the point of this work is to go beyond that, and explore the possibility of creating a more complex classification from the spectra. Therefore, the chosen method must be unsupervised.

(ii) **Dimension reduction:** the spectra of galaxies involved have several hundreds to several thousands of monochromatic fluxes; the data is therefore high-dimensional, and falls under the curse of dimensionality (Sect. 2.2.3). As a result, it is essential to have a robust dimensionality reduction step in the classification pipeline.

(iii) **Large sample size:** the classification method is meant to be applied to large surveys such as the Sloan Digital Sky Survey (SDSS) or the VIMOS Public Extragalactic Redshift Survey (VIPERS). The chosen method must be able to handle large sample size in addition to high-dimensional data.

The algorithm Fisher-EM is unsupervised and uses a GMM, which is particularly well suited for large sample sizes. In addition, it has a built-in dimension reduction step, and optimizes both the GMM and the dimension reduction at the same time, thus meeting all the requirements for this project.

2.5.2 Model

Fisher-EM can be described as a modified version of the EM-GMM algorithm that was upgraded to solve the issue of high dimensionality. Just like the EM-GMM algorithm, it is unsupervised and models the data distribution as a mixture of K multivariate Gaussian probability density functions (see eq. 2.3). The difference is, however, that the data distribution is not modelled in the observed p-dimensional space, but in a latent subspace of dimension $d = K - 1$, with $d < p$. It lies on the assumption that the discriminative information in the data lives in a latent subspace of dimension lower than p, and that $K - 1$ dimensions, if chosen properly, are enough to discriminate K clusters.

If $\{y_1, ..., y_n\}$ denote the n p-dimensional observations, and $\{x_1, ..., x_n\}$ the corresponding observations but described in the latent discriminative subspace, then we can assume that $\{y_1, ..., y_n\}$ are independent realisations of a multivariate random variable Y, and that

$\{x_1, ..., x_n\}$ are independent realisations of another multivariate random variable X. Fisher-EM then presumes that the two random variables Y and X are linked by a linear transformation (eq. 2.10).

$$Y = UX + \epsilon, \tag{2.10}$$

with U being the $p \times d$ projection matrix, common to all clusters, and ϵ being a noise term that is specific to each cluster. This noise term is assumed to follow a normal distribution centred around zero with a covariance matrix ψ_k for class $k \in \{1, ..., K\}$. So if $\{z_1, ..., z_n\}$ denote the class in which the n observations belong, and if they are assumed to be independent realisations of a random variable $Z \in \{1, ..., K\}$, we have (eq. 2.11):

$$\epsilon_{|Z=k} \sim \mathcal{N}(0, \psi_k) \tag{2.11}$$

As explained above, each cluster is assumed to follow a Gaussian distribution within the latent subspace. Of course, each cluster has its own mean μ_k and covariance matrix Σ_k. We thus have (eq. 2.12):

$$X_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{2.12}$$

And because Y is linearly linked to X (eq. 2.10), Y also follows a normal distribution:

$$-Y_{|X,Z=k} \sim \mathcal{N}(U\mu_k, U\Sigma_k U^\top + \psi_k). \tag{2.13}$$

This way, the data in the observed space is also modelled by a mixture of K normal distributions (eq. 2.14):

$$f(K, \theta) = \sum_{k=1}^{K} \pi_k \Phi(U\mu_k, U\Sigma_k U^\top + \psi_k), \tag{2.14}$$

with f the probability density function, π_k the mixture proportions, Φ the multivariate Gaussian probability density function, and θ the model parameters.

This is called the discriminative latent mixture (DLM) model. There are 12 variations of this model that are obtained by adding different constraints on the model parameters. The one presented above has the most degrees of freedom, and is called `DkBk`. The `DkB` model, for instance, assumes that the noise parameter ψ_k is common to all clusters, while `DBk` presumes that the covariance matrix Σ_k is common to all clusters. The 12 DLM models and their associated constraints are summarised in Tab. 2.1.

2.5.3 Algorithm

The DLM model is similar to the GMM, but still differs as it incorporates a latent subspace (inspired from the LDA ; Sect. 2.3.1), which has to be optimised at the same time as the other model parameters. As such, in order to adjust the DLM to the data, Fisher-EM uses a

Tab. 2.1.: The 12 types DLM models provided by Fisher-EM. There are four constraints, denoted as (i) noise parameter ψ_k is common to all clusters, (ii) covariance matrix Σ_k is common to all clusters, (iii) covariance matrix Σ_k is diagonal and (iv) variance is isotropic for each cluster.

Model	(i)	(ii)	(iii)	(iv)
`DkBk`				
`DkB`	✓			
`DBk`		✓		
`DB`	✓	✓		
`AkjBk`			✓	
`AkjB`	✓		✓	
`AkBk`			✓	✓
`AkB`	✓		✓	✓
`AjBk`		✓	✓	
`AjB`	✓	✓	✓	
`ABk`		✓	✓	✓
`AB`	✓	✓	✓	✓

modified version of the EM algorithm which incorporates a third step dedicated to computing the projection matrix U introduced in eq. 2.10. It can be summarised as follows:

- E-step: the membership probabilities of each observation are computed based on the current model parameter estimation (Sect. 2.4.3)
- F-step: the $p \times d$ projection matrix U is computed such that it best discriminates the clusters. This is achieved by maximising the Fisher criterion, i.e. the ratio of between-class variance to within-class variance. This way, in the projected subspace, the clusters are compact (low within-class variance) and distinct from one another (high between-class variance).
- M-step: the estimation of the model parameters are updated given the new membership probabilities (Sect. 2.4.3).

The E-step and M-step serve the same purpose as in the EM-GMM algorithm, Due to the added complexity of the DLM model, the solutions are mathematically heavier, and are not specified here. Should the reader be interested, their full description can be found in Sect. 4 of Bouveyron and Brunet, 2012.

This optimisation loop finds an estimation of the best model parameters for a given number of clusters. The best number of clusters can then be found by exploring the K-space and looking for the value of K that maximises the likelihood.

The DLM model type and its parameters are optimized by maximizing the integrated completed likelihood (ICL), a penalised variation of the regular likelihood that is more suited for clustering (Biernacki et al., 2000). Similarly, the number of clusters is chosen by looking for an optimum in the ICL vs. K curve.

2.5.4 Usage

Fisher-EM is a fine tool that is well suited for this project, but it cannot be used mindlessly; it was essential to have a good understanding of how it works in order to use it properly. In other words, one cannot simply run the algorithm without careful and thoughtful planning and expect it to return a scientifically relevant solution. Although the steps may differ slightly from one dataset to another, the standard approach to successfully using Fisher-EM for galaxy spectra classification starts with some data preparation.

First and foremost, the spectra must be transformed and prepared to remove any unwanted features. Although some data processing is sample specific (e.g. adding artificial noise or removing sky-lines residuals), I established three common and fundamental steps:

- **Redshift correction** – Correcting the spectra from redshift is crucial to re-align the spectral features within the sample (e.g. emission and absorption lines, breaks). If omitted, the resulting classification would simply create classes of galaxies sharing similar redshift. We want the classification to segregate galaxies based on intrinsic physical properties, and it was therefore necessary for me to remove any extrinsic information within the spectra.
- **Normalisation** – Normalising the spectra is also necessary to allow Fisher-EM to focus on the spectral features rather than the amplitude level. In fact, similarly to redshift, should Fisher-EM be run on non-normalised spectra, the resulting classes would essentially just segregate galaxy masses. To avoid that, the spectra are normalised, typically by selecting a wavelength range without any significant features, and dividing each spectrum by its average value in that range.
- **Resampling** – Algorithmically, it is necessary that all the observations share the same sampling, so that each of the p features can be compared from one spectrum to another. However, after correcting for redshift, the spectra may not sample the same wavelength. As a result, it is usually unavoidable to resample the dataset to correct this issue.

Once the data has been prepared and is ready to be classified, then it is necessary to choose the appropriate Fisher-EM settings for the current dataset. It includes:

- **Initialisation** – There are four different ways to initialise Fisher-EM. By default, the K-means algorithm (Sect. 2.4.1) is run N times, and the result associated with the highest likelihood is used as the initial classification state. Similarly, a set of N random initialisations can be drawn, and the one associated with the highest likelihood is kept. The third option consists in using the hierarchical clustering algorithm (Sect. 2.4.2), but a single time, since this algorithm is deterministic. Finally, the last option allows for a personalised initialisation. Most of the time, I found that

the default initialisation with K-means worked well. However, in certain cases, I found that a random initialisation turned out to converge towards a better solution and avoid a local minimum. But there is no way to know in advance which option to chose: the choice is done by trial and error.

- **Model selection** – The DLM model and the number of clusters to explore must be chosen. My approach consists in running the algorithm for all 12 DLM model with a rough sampling of the number of clusters. Based on the best-performing models, and the shape of the ICL vs. number of cluster curve, I run Fisher-EM a second time and fine-tune the parameters to get the best solution.
- **Fitting method** – There are three methods available for the fitting of the projection matrix U: `gs`, `svd` and `reg`. The `gs` method uses the Gram-Schmidt process, `svd` refers to a method based on single value decomposition, and in the `reg`, the Fisher criterion is treated as a regression problem. In my case, since I work with large dataset, Bouveyron and Brunet (2012) advises the `svd` method which is the fastest one and thus the most appropriate.
- **Kernel** – Sometimes, the sample size n happens to be lower than the number of dimension p. For most generative models, the estimation of the model parameters becomes impossible (Bouveyron and Brunet, 2012). Fisher-EM, however, can overcome this issue with a mathematical trick detailed in Sect. 4.6 of Bouveyron and Brunet (2012). It requires the use of a kernel, which can be chosen among three options: linear, sigmoid or radial basis function. When necessary, I found that the linear kernel was performing best.

Finally, I found that it was sometimes necessary to classify the data in two steps. In my understanding, when the diversity in the data is too large, it may be difficult for Fisher-EM to find a projection matrix U that properly encompasses all the discriminant features that lie in the data, and mostly focuses on the dominating one (essentially the continuum). As a result, it produces a classification where the dispersion in the classes remains rather high. But this issue is solved by isolating the classes obtained and classifying them individually a second time. This way, the variance of the most prominent features remains low enough for Fisher-EM to detect more subtle patterns and produce a more refined classification.

During my thesis, I mostly worked on two samples of galaxy spectra: one being simulated data using the SED modelling code CIGALE (Chapt. 3), and the second being observations from the VIMOS Public Extragalactic Redshift Survey (Chapt. 4). I found that the two-step classification procedure was necessary for the latter, but not the former.

3 Unsupervised classification of a CIGALE-simulated sample of galaxy spectra

3.1 Introduction

The first step of my thesis consisted of using Fisher-EM to classify a sample of simulated galaxy spectra, motivated by the need to provide a 'proof of method'. The simulated nature of the data implies that the physical characteristics of the galaxies are known, thus simplifying the analysis of the physical relevance of the classification results. This part of my work led to a publication (Dubois et al., 2022).

First, in Sect. 3.2 I present the code that was used to simulate the galaxy spectra, and the resulting mock catalogue is described in Sect. 3.3. Then, the results obtained on the noiseless catalogue are presented in Sect. 3.4, and those obtained with added noise are shown in Sect. 3.5. Finally, I discuss these results in Sect. 3.6 and conclude on this portion of my work in Sect. 3.7.

3.2 The Code Investigating GALaxy Emission

3.2.1 Purpose of the code

Nowadays, thanks to numerous observation infrastructures, cutting edge instruments, and powerful data processing tools and archives, we are lucky to have access to multiple large scale spectroscopic, imaging and photometric surveys of galaxies spanning a wide range of wavelengths and redshifts. In an effort to interpret these observations and deduce intrinsic physical properties of galaxies, Boquien et al. (2019) have developed a software called CIGALE, short for Code Investigating GALaxy Emission. This software consists of several modules allowing for complex modelling of a galaxy's spectral energy distribution (SED). By matching the observations with the model, estimations of physical properties (e.g. star formation rate, age, metallicity) can thus be inferred.

Conveniently, CIGALE can be used the other way around to construct mock catalogues. Instead of fitting the model parameters to match observations, one can cover a wide range of parameter values to generate a large sample of mock spectra of galaxies. This is how CIGALE was used in this part of my work. In this case, the model's physical parameters are inputs for the user to construct the mock spectra of their choice. Among the parameters available as inputs, the user can play with e.g. the age of the galaxy, its history of star formation, its metallicity, its redshift, and much more.

The next subsection will be dedicated to giving the reader an exhaustive overview of CIGALE's modules. The necessary knowledge about the code will be given to understand how the mock sample was generated.

3.2.2 The model

CIGALE builds its models by successively running a series of modules that compute independent physical processes (Tab. 3.1); step by step, a complex model is thus constructed. First, the star formation history is defined (Sect. 3.2.2.1). Combined with single stellar population models, a composite spectrum of the stellar contribution is computed (Sect. 3.2.2.2). Then, emission from the ionised gas around young stars is processed and added to the stellar

Tab. 3.1.: Computing steps of CIGALE and the modules available. One module has to be chosen for each category. The steps are listed in the same order they are called by the program.

Step	Modules
Star formation history	`sfh2exp`
	`sfhdelayed`
	`sfhdelayedbq`
	`sfhperiodic`
	`sfh_buat08`
Stellar populations	`bc03`
	`m2005`
Nebular emission	`nebular`
Dust attenuation	`dustatt_modified_CF00`
	`dustatt_modified_starburst`
Dust emission	`dale2014`
	`dl2007`
	`dl2014`
	`casey2012`
Synchrotron radio emission	`synchrotron`
Active galactic nuclei	`fritz2006`
Redshifting & IGM absorption	`redshifting`

emission (Sect. 3.2.2.3). Afterwards, the effect of dust is taken into account both through attenuation and emission (Sect. 3.2.2.4 and Sect. 3.2.2.5). Lastly, the contributions of the synchrotron emission (Sect. 3.2.2.6), active galactic nuclei (Sect. 3.2.2.7), redshifting and intergalactic attenuation (Sect. 3.2.2.8) are computed to obtain the final spectrum.

In this subsection, I will provide a description of these modules, how they work, and the physical parameters involved. The description of the modules will be made from the perspective of a user wanting to generate mock spectra with CIGALE—because that is how it was used for this work—, but keep in mind that it is also possible to computationally adjust the input parameters to obtain the best-fitting model for a given observed galaxy. Should the reader be interested in learning more about the code, a more extensive description of CIGALE can be found in the original papers Burgarella et al. (2005), Noll et al. (2009) and Boquien et al. (2019).

3.2.2.1 Star formation history

Throughout the life of a galaxy, the star formation rate (SFR) varies greatly, from quiescent times to periods of intense star formation. As a result, accurately modelling the evolution of the SFR over time—which we shall call the star formation history (SFH)— is tough. One can opt to use complex simulations to obtain realistic SFHs, but this approach is obviously very computation-heavy, and while it can prove to be useful when studying individual galaxies, it is not necessarily the best approach when working with numerous sources. An alternative, which was broadly used in the literature, consists of using rather simple

analytical models that encompass the general trend of the evolution of the SFR. CIGALE allows both approaches. In fact, the SFH module has five simple analytical models available, but it is also possible to input a personalized SFH that the user may have obtained from an independent simulation. For the purpose of this study, the analytical approach was taken. The choice must be made between the following analytical models.

`sfh2exp` – In this model, the SFH is described as the sum of two decaying exponentials. Usually, one exponential is used to describe a main population of stars that formed throughout the whole life of the galaxy, i.e. with a timescale of the order of several Gyr, while the second exponential can simulate sudden sporadic burst of star formation, i.e. with a much smaller timescale of the order of several Myr. The SFR can thus be expressed analytically as a function of time by the following formula (Eq. 3.1).

$$\mathrm{SFR}(t) \propto \begin{cases} \exp(-t/\tau_{main}) & \text{if } t < T_{main} - T_{burst}, \\ \exp(-t/\tau_{main}) + k \times \exp(-t/\tau_{burst}) & \text{if } t \geq T_{main} - T_{burst} \end{cases} \tag{3.1}$$

Where τ_{main} and τ_{burst} are the time constants of the main stellar population and of the starburst, k the amplitude of the starburst exponential relative to the main exponential, and T_{main} and T_{burst} the age of the first stars formed in each of the two stellar populations. The relative amplitude k can be written analytically as a function of a more physical quantity: the mass fraction of stars from the starburst population, that we shall call f_{burst}. As a result, the sfh2exp model has 5 physical parameters : τ_{main}, τ_{burst}, T_{main}, T_{burst} and f_{burst}. Note that by default, the total amplitude of the SFR is not parametrized, because it is normalized such that the total mass of stars formed over 13 Gyr is 1 $M_\odot$. It can be scaled up to a given mass to fit the user's needs, but it was not necessary for the purpose of this work. This also applies to the four other models.

`sfhdelayed` – This model is a slight variation of the previous one, designed to smooth out the SFR at the start of star formation, where a simple decaying exponential might be too abrupt. As such, the SFR of the main stellar population is weighted by a factor $\frac{t}{\tau^2}$, but the starburst remains unchanged, resulting in Eq. 3.2.

$$\mathrm{SFR}(t) \propto \begin{cases} \frac{t}{\tau_{main}^2} \exp(-t/\tau_{main}) & \text{if } t < T_{main} - T_{burst} \\ \frac{t}{\tau_{main}^2} \exp(-t/\tau_{main}) + k \times \exp(-t/\tau_{burst}) & \text{if } t \geq T_{main} - T_{burst} \end{cases} \tag{3.2}$$

This model is parametrized like `sfh2exp` with the same 5 inputs: τ_{main}, τ_{burst}, T_{main}, T_{burst} and f_{burst}.

`sfhdelayedbq` – The two previous models sometimes are not sufficient, as they cannot properly describe the recent quenching of star formation that is sometimes observed. This

model addresses this issue by setting the SFR to a constant value after a given time (Eq. 3.3).

$$\mathrm{SFR}(t) \propto \begin{cases} \frac{t}{\tau_{main}^2} \exp(-t/\tau_{main}) & \text{if } t < T_{bq} \\ r_{\mathrm{SFR}} \times \mathrm{SFR}(t = T_{bq}) & \text{if } t \geq T_{bq} \end{cases} \tag{3.3}$$

Like the previous models, τ_{main} and T_{main} are the time constants and the age of the main stellar population. Then, T_{bq} is the time when a quick burst or quenching of star formation happens, and r_{SFR} is the ratio of the SFR after and before the event. Thus, `sfhdelayedbq` has 4 inputs: τ_{main}, T_{main}, T_{bq} and r_{SFR}.

`sfhperiodic` – As its name suggests, this model is designed to produce a periodic SFH, and can take three shapes: exponential, delayed or rectangular. However, this model doesn't allow a second stellar population within the periodic signal, unlike `sfh2exp` and `sfhdelayed`. It is used to simulate repetitive and identical bursts of star formation. Whether the user choses an exponential, delayed or rectangular shape, the inputs remain the same: *type* for the shape of the periodic signal, δ for its period, T for the age of the oldest star, and τ for the time constant of each burst—equivalent to τ_{main} and τ_{burst} for `sfh2exp` and `sfhdelayed`.

`sfh_buat08` – The last option is based on the work of Buat et al. (2008) which provides an empirical relation between the SFR and the rotational velocity of the galaxy (Eq. 3.4).

$$\mathrm{SFR}(t) \propto 10^{a+b\times\log(t)+t^{1/2}}, \tag{3.4}$$

with a, b and c being constants defined by the rotational velocity. This model thus has two input parameters: v the velocity and T the age of the oldest star of the galaxy.

3.2.2.2 Stellar populations

Once the SFH has been chosen, the composite spectrum of the galaxy can be computed. The first step in this process is to model the emission of single stellar populations (SSP). Several libraries of SSPs are available in the literature, and CIGALE allows the user to choose from the library of Bruzual and Charlot (2003) or that of Maraston (2005). Once the SSP library is chosen, the composite stellar spectrum is obtained by computing the dot product of the SFH (a vector of SFR at different time steps) with a matrix that contains the SSP spectra throughout the corresponding stages of evolution.

`bc03` – The SSP library of Bruzual and Charlot (2003) is available for metallicities of 0.0001, 0.0004, 0.004, 0.008, 0.02 and 0.05 and for the IMFs of Salpeter (1955) and Chabrier (2003).

m2005 – The SSP library of Maraston (2005) is available for metallicities of 0.001, 0.01, 0.02 and 0.04, and for the IMFs of Salpeter (1955) and Kroupa (2001).

In both cases, the user has three parameters to chose: the initial mass function (IMF), the metallicity Z, and the age of the separation between the young and the old star populations T_{sep}. The latter is used to divide the population into young stars that have not yet accreted or ejected their surrounding gas, and older stars that did. This is necessary for the next module to compute the nebular emission around young stars.

3.2.2.3 Nebular emission

As introduced in the previous module, young stars can ionize the remaining gas surrounding them, inducing re-emission both in the form of emission lines and thermal continuum. CIGALE computes both of these contributions with the templates provided by Inoue (2010) and Inoue (2011), which have been improved on several aspects (parameter sampling and abundance values). First of all, the **nebular** module is parameterised by the ratio of ionizing photon density over hydrogen density, called the ionisation parameter U. The metallicity also plays a role in the process, but it is assumed to be the same as previously defined in the stellar populations module. To account for the effect of ionising photons either escaping the galaxy or being absorbed by dust, the module offers two other parameters defining the fraction of escaping photons $f_{\rm esc}$ and of dust-absorbed photons $f_{\rm dust}$. The nebular emission is thus rescaled by the corresponding factor following the prescriptions of Inoue (2011). Finally, the user may define the line width to account for the motion of the gas. In the end, the nebular emission module thus has four input parameters: the ionization parameter, the fraction of escaping photons and dust-absorbed photons, and the line width.

3.2.2.4 Attenuation laws

Dust grains effectively absorb UV to NIR emissions and re-emit it thermally in the mid to far-IR. This module aims to model dust attenuation, and it does so following empirical attenuation laws that the user may choose between the models of Charlot and Fall (2000), or a combination of Calzetti et al. (2000) and Leitherer et al. (2002).

dustatt_modified_CF00 – This model follows the prescriptions of Calzetti et al. (2000), which assumes that the attenuation has two components: attenuation of the interstellar medium (ISM) and an additional attenuation for young stars still embedded in gas and dust. So in effect, the light from young stars is assumed to be attenuated first by their birth cloud, and then by the ISM, while old stars are only affected by the ISM. Both attenuations are assumed to follow a power law. The model is thus parameterised by their respective slope, the V-band attenuation for the ISM, and the ratio of attenuations for old and young stars.

The line between young and old stars is set by the T_{sep} parameter defined in the stellar populations module (Sect. 3.2.2.2).

dustatt_modified_starbust – Instead of modelling the physical processes, this module provides an alternative approach, which consists of taking attenuation curves derived from observations (from Calzetti et al., 2000 and Leitherer et al., 2002) and extending them to allow for more flexibility and generalisation. Namely, the module provides the possibility to (i) multiply the attenuation curve by a power law, and (ii) to add a UV bump as a Drude profile. The resulting attenuation is thus (Eq. 3.5):

$$k_\lambda = \left(k_\lambda^{ref}\left(\frac{\lambda}{550}\right)^\delta + D_\lambda\right) \times \frac{E(B-V)_{\delta=0}}{E(B-V)_\delta}, \tag{3.5}$$

with k_λ^{ref} the attenuation curve from Calzetti et al. (2000) and Leitherer et al. (2002), λ the wavelength in nm, γ the power law slope, and D_λ the Drude profile. The last term is there to normalize the attenuation in case $\delta \neq 0$ so that it matches the wanted extinction E(B-V). Finally, the module applies a correction factor to process the extinction of the emission lines, which can be different from that of the continuum.

3.2.2.5 Dust emission

The previous module models the absorption of the dust, but this energy absorbed is re-emitted in the infrared through thermal emission. In CIGALE, the re-emission is modelled in this dust emission module, after computing the absorption in the previous module. There are four different emission modules available, based on different sets of models from the literature.

casey2012 – This module uses the models of Casey (2012) which considers a two-components emission: (i) thermal black body emission in the far-IR from the heated dust around young stars and (ii) a power law in the mid-IR to approximate dust emission around AGN. This module is parameterised by three parameters: dust temperature, emissivity index of the dust, and mid-IR power law coefficient.

dale2014 – This one is based on Dale et al. (2014) and is the simplest out of the four available options. It models the dust emission with a single parameter α which defines the mass of heated dust (eq. 3.6)

$$dM_{dust} \propto U^{-\alpha} dU, \tag{3.6}$$

with M_{dust} the mass of the heated dust, and U the radiation field intensity. The module also has an optional parameter which provides the possibility to add an AGN component. AGN emission is computed using templates from Dale et al. (2014), and the input AGN to dust luminosity fraction.

dl2007 – This module uses the models presented in Draine and Li (2007), and is the most complex and flexible dust emission model available in CIGALE (with its updated version **dl2014**). The emission is based on two components: (i) emission of the dust heated by the overall stellar population and (ii) a specific component for heated dust in star-forming regions. The former is parametrised by a unique radiation field U_{min}, while the latter considers a gradient from U_{min} to U_{max} following a power law with a fixed coefficient. The relative proportion of these two component is set by a parameter γ. In addition, this module models the contribution of polycyclic aromatic hydrocarbon (PAH), parameterised by the PAH mass fraction q_{PAH}.

dl2014 – The work of Draine and Li (2007) used in the **dl2007** module was updated in Draine et al. (2014), and the modifications were adapted in this module for CIGALE. Aside from a refinement of the templates, two changes were made which affect the input parameters: (i) U_{max} is no longer a free parameter and is set to 10^7, and (ii) the power law coefficient defining the radiation distribution between U_{min} and U_{max} is now a free parameter.

3.2.2.6 Synchrotron emission

Synchrotron emission designates the non-thermal emission produced by electrons travelling at a relativistic speed through a magnetic field. This type of phenomenon is typically observed around compact objects but also on a larger scale due to super-novae spreading relativistic electrons which may encounter local magnetic fields on their path. To account for this phenomenon, CIGALE uses the known correlation between the non-thermal radio and thermal far-infrared fluxes of galaxies from Helou et al. (1985). This relation is parameterised by the correlation factor $q_{\rm IR}$, and allows the inference of the 21cm flux based on the FIR flux. The rest of the radio spectrum is assumed to follow a power-law parameterised by a coefficient α, and scaled to match the expected 21cm flux. The **synchrotron** module thus only takes two input parameters, $q_{\rm IR}$ and α, to model this complex phenomenon based on empirical laws.

3.2.2.7 Active galactic nuclei

The emission from an active galactic nucleus can be modelled in a simple way by the **casey2012** and **dale2014** modules from the dust emission section (See Sect. 3.2.2.5). A second option is to opt for the **fritz2006** module which provides more complex modelling based on the radiation transfer work of Fritz et al. (2006). The geometry is assumed to consist of the central source surrounded by a torus of dust. The total emission is the sum of three components: the emission of the source, the scattering of the dust, and the thermal re-emission of the dust. It is parameterised by four geometrical parameters: the radius of the torus r, its opening angle θ, the ratio of the maximum to minimum radii of the torus

r_{radii}, and the angle between the line of sight and the axis of the AGN ψ. In addition, the dust density within the torus is modelled by the following relation (Eq. 3.7):

$$\rho \propto r^{\beta} \exp^{-\gamma|cos\theta|}, \tag{3.7}$$

and is thus parameterised by γ and β. And finally, the last parameter is the optical depth at 9.7 μm τ.

3.2.2.8 Redshifting & IGM attenuation

The final spectrum is obtained after (i) shifting it according to its redshift and (ii) taking into account the attenuation of the intergalactic medium (IGM). The **`redshifting`** module hence takes a single parameter, the redshift z, and multiplies the wavelengths by $1 + z$ to shift the spectrum, and divides the fluxes by $1 + z$ to dim it down. Afterward, the absorption of the IGM in the line of sight is computed following the estimates of Meiksin (2006).

3.3 The mock catalogue

3.3.1 Module and parameter selection

The mock sample I used for this part of my work was originally created by Dr. P. Sharma in the context of an internship LAM[1] under the supervision of Pr. D. Burgarella and Pr. D. Fraix-Burnet, and contained a total of 1 000 000 spectra covering the entire spectral range that CIGALE provides, i.e. from FUV to radio. I did not contribute to generating the simulations per se, as this was done a year prior to the start of my PhD. However, it was necessary for me to dig into how the CIGALE code functions in order to properly understand the data I was working with. In fact, I had to spend a significant amount of time re-organising the simulation outputs, and retrieving the corresponding simulation parameters using the various initialisation files used to generate the data. In this section, I summarise the characteristics I gathered about this mock catalogue.

The modules that were chosen for the simulation are listed in Tab. 3.2. The SFH was modelled using **`sfhdelayed`** which provides the possibility to simulate a bimodal population: a "main" stellar population i.e. stars that form throughout the entire life of the galaxy; and a "burst" stellar population i.e. a sudden onset of star formation (starburst). The stellar populations were simulated using the **`bc03`** module with an IMF from Chabrier (2003). For the dust attenuation, the **`dustatt_modified_starburst`** module was used, and for dust emission, **`dl2014`**. Finally, no AGN contribution was added.

[1] Laboratoire d'Astrophysique de Marseille, Aix-Marseille University

Tab. 3.2.: List of the modules used to simulate the mock catalogue, and their corresponding physical step.

Step	Module
Star formation history	`sfhdelayed`
Stellar populations	`bc03`
Nebular emission	`nebular`
Dust attenuation	`dustatt_modified_starburst`
Dust emission	-
Synchrotron radio emission	-
Active galactic nuclei	-
Redshifting & IGM absorption	`redshifting`

These modules, as they model a various number of physical processes that occur in galaxies, may influence different spectral regions of the spectra. For instance, the synchrotron module affects the radio fluxes, while the dust attenuation modules have an impact from the UV to NIR. However, in the work that I conducted, the spectra were restricted to the optical wavelengths (496 fluxes between 380.66-737.00 nm). This choice was made to stand on the same ground as previous studies of classifications that were done on large-scale optical surveys such as the Sloan Digital Sky Survey, SDSS (Fraix-Burnet et al., 2021). As a notable side effect, this reduced the number of unique spectra in the mock catalogue to 11 475. This is explained by the fact that, in CIGALE, the optical part of the spectrum is not affected by all the modules. Namely, the dust emission, synchrotron emission and AGN modules have no effect at these wavelengths. Therefore, two spectra can be issued from different simulations, but as long as their differences only lie in the modules that do not affect the optical range, they end up being effectively identical in this spectral region. As a result, a great number of spectra in the original catalogue were duplicates in the optical, and these redundancies were removed, hence significantly decreasing the sample size. Nonetheless, this mock sample was sufficient to meet the expected goals.

In order to choose realistic combinations of simulation parameters, three reference galaxies were used: M105, NGC2976 and M82. They correspond to three archetypes of galaxies that we observe in the universe. First, "dead galaxies" that are mostly red, elliptical and old. They do not form new stars, and as a result, have essentially no emission lines in their spectra. Then, "starburst galaxies", which, at the time of observation, showcase a particularly high burst of star formation, hence inducing blue spectra and with intense emission lines. And finally "Milky-Way type galaxies", which are sort of a middle ground. They are still forming new stars but at a more moderate rate than starburst galaxies, and thus have less extreme features. CIGALE was used to fit SED models on observed photometric data from these 3 reference galaxies. This led to 3 sets of parameters that were used as reference values. To generate the catalogue, the parameters were varied around these references. All the combinations of parameter values are listed in Tab. 3.3, and their distribution shown in Fig. 3.1. Bellow, I make a summary of the modules and their corresponding relevant

Tab. 3.3.: CIGALE parameters used to generate the data

T_{main} (Myr)	τ_{main} (Myr)	f_{burst}	T_{burst} (Myr)	τ_{burst} (Myr)	Metallicity	E_BV_lines	E_BV_factor	redshift
9500 10500 11500	500 1000 3000				0.02 0.05	0.0005 0.01 0.025 0.05 0.075 0.1	0.25	0.001 0.002 0.003 0.004 0.005
10000 11000 12000	2000 5000	0	-	-				0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
6000 10000	5000							0.1 0.2 0.3
5000 7500	5500							0.4 0.5 0.6
4000 6500	4500							0.7 0.8
13000	5000							0
2000 5200	3000	0.001 0.01 0.1	20 100	4500	0.02	0.005 0.077 0.148 0.220 0.292 0.363 0.435 0.507 0.578 0.650	0.44	0.8 0.9 1
2000 7000	4000			6000				0.7
2500 7500				7000				0.6
3000 8000	5000			7000				0.5
4000 9000				8000				0.4
5000 10000	5000 10500			50000				0.1 0.2 0.3
10000 11000 12000	6000 9000			10000				0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
6000	5000 8000	0.1 0.25 0.5	5 20 50	11000	0.008 0.02	0.005 0.116 0.226 0.337 0.447 0.558 0.668 0.779 0.889 1.000		0.4 0.5 0.6
4000 7200	5000			6500				0.1 0.2 0.3
5000 10000	6000			9000				

parameters, i.e. parameters that have an effect on the optical part of the spectra and that of interest for this work.

1. **Star formation history** – `sfhdelayed`
   Relevant parameters: T_{main}, τ_{main}, T_{burst}, τ_{burst}, f_{burst}

   T_{main} is the age (in Myr) of the main stellar population in the galaxy. It is varied from 2 000 to 13 000 Myr to simulate a range of young and older galaxies. τ_{main} is the e-folding time (in Myr) of the main stellar population and characterises the SFR of the main population together with T_{main}. In the sample, τ_{main} varies from 500 to 10 500 Myr. In effect, this means that some galaxies would have older stellar populations, while some others have a more spread out star formation history. f_{burst} is the mass fraction of stars produced during a burst of star formation, and ranges from 0 to 0.5. When $f_{burst} = 0.5$, it means that half of the galaxies' stellar mass comes from this burst of star formation. This is an extreme case, of course, but smaller values were also explored. T_{burst} is the age (in Myr) of the burst of star formation and ranges from 5 to 100 Myr, meaning that only recent bursts of star formation were simulated. τ_{burst} is the e-folding time (in Myr) of the burst of star formation, it characterises the SFR of the burst event together with T_{burst}, and varies from 4 500 to 50 000 Myr. These values are of the same order as $T_{main}-T_{burst}$ and much higher than T_{burst}, so that the SFR is essentially constant during the second burst of star formation. When $f_{burst} = 0$, no burst is considered in the history of the galaxy, so T_{burst} and τ_{burst} are therefore set to 0.

2. **Stellar population** – `bc03`
   Relevant parameters: Metallicity

   Metallicity is assumed to be identical for the main and burst stellar populations, and takes three possible values: 0.008, 0.02, and 0.05. For the IMF, we use that of Chabrier (2003). The `bc03` module has a third parameter (see Sect. 3.2.2.2): the age threshold between young and old stars, but this parameter was left at its default value of 10 Myr for all simulations.

3. **Nebular emission** – `nebular`
   Relevant parameters: Metallicity

   The nebular metallicity is assumed to be identical to the stellar metallicity, and thus takes the same values as the previous module. The other parameters (see Sect. 3.2.2.3 of this module were kept at default values for all the simulations.

4. **Attenuation laws** – `dustatt_modified_starburst`
   Relevant parameters: $E(B\text{-}V)_{cont}$

   $E(B\text{-}V)_{cont}$ is the reddening of the continuum that intervenes in the computation of the attenuation due to the dust inside the galaxy. The attenuation of the emission lines is different but proportional to the continuum extinction. As such, we will

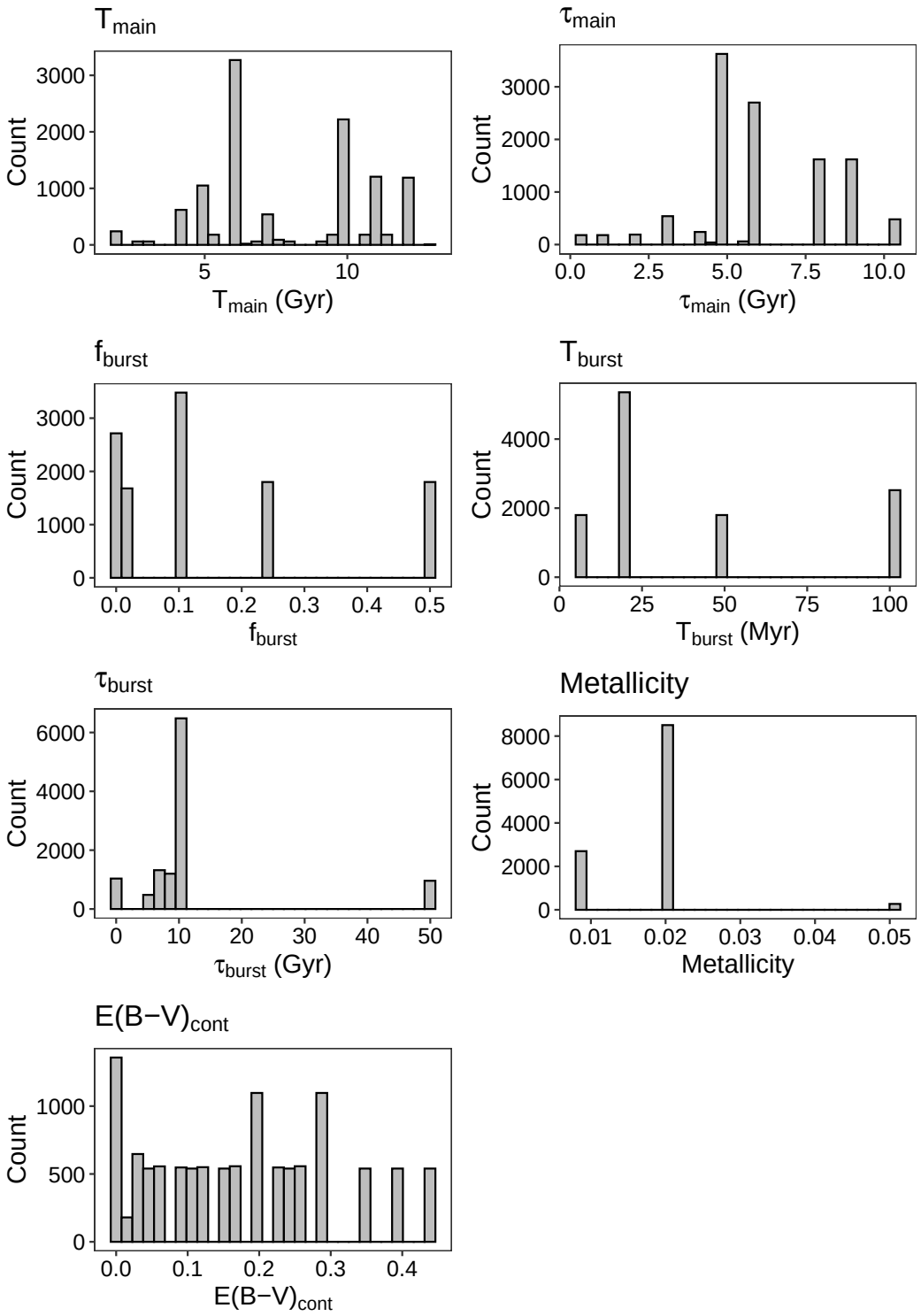


Fig. 3.1.: Histograms of CIGALE parameters input values in the study sample. This sample was not created to fit some specific parameter distribution, but rather covers the parameter space as much as CIGALE allowed it while keeping the spectra realistic.

Tab. 3.4.: Parameter linear correlation coefficients. The parameters are not intrinsically correlated in CIGALE, but the combinations of values used to generate the sample for this study may show underlying involuntary correlations between some parameters.

	T_{main}	τ_{main}	f_{burst}	T_{burst}	τ_{burst}	Metallicity
τ_{main}	0.17					
f_{burst}	−0.42	0.01				
T_{burst}	0.21	0.18	−0.28			
τ_{burst}	−0.06	0.38	−0.11	0.20		
Metal.	0.33	−0.20	−0.39	0.11	−0.02	
E(B-V)$_{cont}$	−0.22	0.10	0.27	−0.10	−0.01	−0.28

only discuss E(B-V)$_{cont}$ throughout this chapter. In the simulated spectra, E(B-V)$_{cont}$ takes a wide variety of values from 0.000125 to 0.44. The other parameters of the `dustatt_modified_starburst` module (see Sect. 3.2.2.4) have not been varied in the sample, and are therefore ignored in this work.

The rest of the modules (dust emission, synchrotron emission, and redshifting & IGM absorption) have virtually no effect on the optical spectra, and can therefore be ignored. Originally, the redshift z was still varied from 0 to 1 to include the effect of the intergalactic medium (IGM) on the spectra. But because this effect is very small on the optical range, and since we shifted the spectra back to their rest-frame for the analysis, the redshift ends up being irrelevant to the study. The remaining 11 475 spectra are therefore entirely parameterised by the star formation, stellar populations, nebular emission and attenuation laws modules. The other modules have strictly no impact on this work and will therefore no longer be mentioned.

As defined previously, the T_{main}, τ_{main}, f_{burst}, T_{burst}, τ_{burst} and Metallicity parameters are the building blocks of the stellar populations composing a galaxy. It is thus natural that varying the values of these parameters directly influences the continuum and the absorption lines of a galaxy spectrum (e.g. a high value of T_{main} reddens the spectrum and deepens molecular lines). Moreover, E(B-V)$_{cont}$ translates the presence of dust into the ISM by reddening the spectrum. On the other hand, the emission lines in a galaxy spectrum are the signature of a recent star-forming event in the history of that galaxy. Therefore, the same parameters T_{main}, τ_{main}, f_{burst}, and T_{burst} are also responsible for the presence and the height of these lines. Moreover, since the interstellar medium in CIGALE has the same metallicity as the stars, the metallicity parameter will also affect the emission lines in addition to the stellar component. Finally, because the chosen values of τ_{burst} are much higher than those of T_{burst} (see Table 3.3), τ_{burst} only weakly affects the spectra.

Since some modules of CIGALE are built around templates, a significant portion of the parameters only take discrete and pre-determined values, making it impossible to fully cover

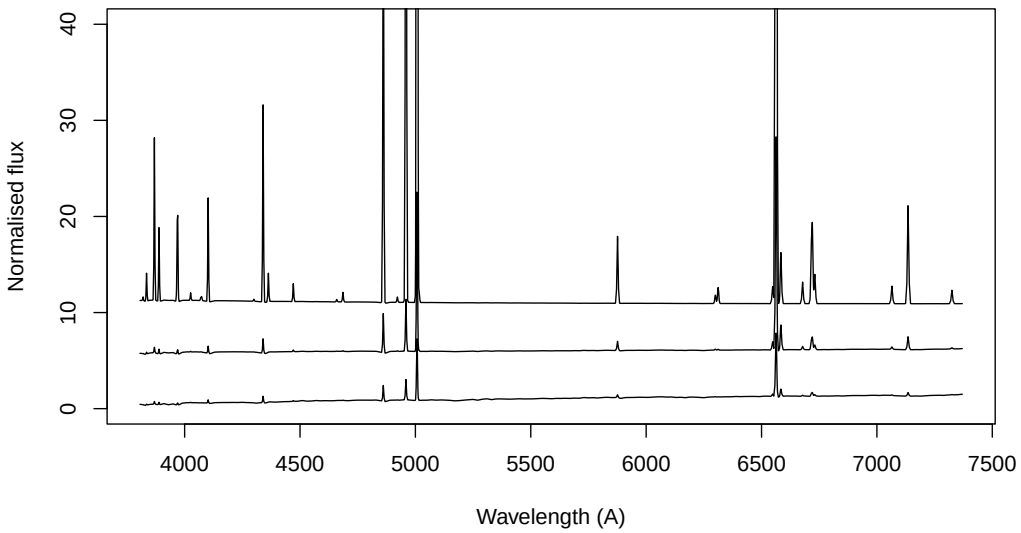


Fig. 3.2.: A few examples of spectra from the CIGALE simulated sample.

the parameter space. Some parameters such as E(B-V)$_{cont}$ allow for a decent sampling, while some others such as Metallicity only take very few values. The sampling density of a parameter may affect its discriminant power slightly, but in a very indirect way since the clustering is performed with the spectra and not with the input parameters.

Moreover, while the parameters are not intrinsically correlated in CIGALE, we may observe some unwanted correlations in our catalogue due to the parameter sampling. Such correlations have to be kept in mind when analysing the discriminative properties of the classification method, as one parameter may deceptively appear well discriminated due to its correlation with another discriminated parameter. However, for our sample, the Pearson correlation coefficients remain small (see Table 3.4) and reaches -0.42 (f$_{burst}$ versus T$_{main}$) at most.

3.3.2 Data preparation

As presented in the previous section, a mock catalogue of 11 475 optical spectra of galaxies was created using CIGALE. To properly prepare the data before using the classification algorithm, two pre-processing steps were used: (i) redshift correction and (ii) normalisation (see Chapter 2 Sect. 2.5). For the CIGALE mock catalogue, resampling was not necessary because the spectra were generated in such a way that they happen to be sampled identically once shifted back to their rest-frame. The normalisation was done by dividing a spectrum's fluxes by its mean flux between 505 and 581 nm, a region where the spectra have no emission lines or significant spectral features.

Additionally, in this work, we were interested in investigating the effects of noise on the classification results. The simulated spectra are originally noiseless, so to assess the impact of different noise levels, I created duplicates of the original mock catalogues with added noise of signal-to-noise ratio (S/N) ranging from 1 to 500. Noisy sample sets were generated by adding a Gaussian noise of matching S/N to each monochromatic flux of the noiseless data. I then studied successively the noiseless sample, and the noisy samples.

But before diving in showing the classification results, which are presented in the next section, I illustrate in Fig. 3.2 a few examples of spectra from the mock catalogue. The spectra are depicted after applying the pre-processing steps, so in other words, this is the data that was fed to the classification algorithm.

3.4 Analysis of the noiseless spectra

3.4.1 Optimal model and number of classes

The algorithm fits the data distribution with a DLM model which has many free parameters that have to be optimised in order to find the best classification possible for a given dataset (see Chapter 2 Sect 2.5). This optimisation is done using the ICL, a statistical criterion which is a variation of the regular likelihood. The best DLM model was found to be $A_{kj}B_k$ in all the cases studied in this work (including the noisy data). This model is such that in each group, the covariance matrix Σ_k is assumed to be diagonal: $\Sigma_k = diag(\alpha_{k_1}, ..., \alpha_{k_d})$.

Similarly, the optimum number of clusters K is also given by the maximum ICL value, and one has to look at the ICL vs. K curve to find the best K-value. The goal of this work on a mock catalogue was not only to obtain a classification, but also to assess the stability of the algorithm and the effect of noise. The latter is addressed in Sect. 3.5, and the former is done by computing several classifications for each number of clusters K, yielding a distribution of ICL values for each.

Let us note that Fisher-EM sometimes finds an empty cluster. This results in an undefined log-likelihood that stops the algorithm. This is what we denote as the non-convergence of the algorithm. To investigate the stability of the algorithm and its convergence rate, I repeated the computations a few dozens times, which was found to be a good compromise between computation time and sufficient data. In Fig. 3.3, I show the convergence rate of the algorithm as per this definition, as well as the ICL vs K curve. We see that there is a remarkable jump in ICL values at $K = 14$. In addition, the algorithm does not converge for $K = 13$, while there is a 100% convergence rate at $K = 12$ and $K = 14$. Henceforth, $K = 13$ appears as a frontier between two very distinct regimes, which I both studied and compared. For $K < 13$, the algorithm always converges, and the ICL is maximum at $K = 12$.

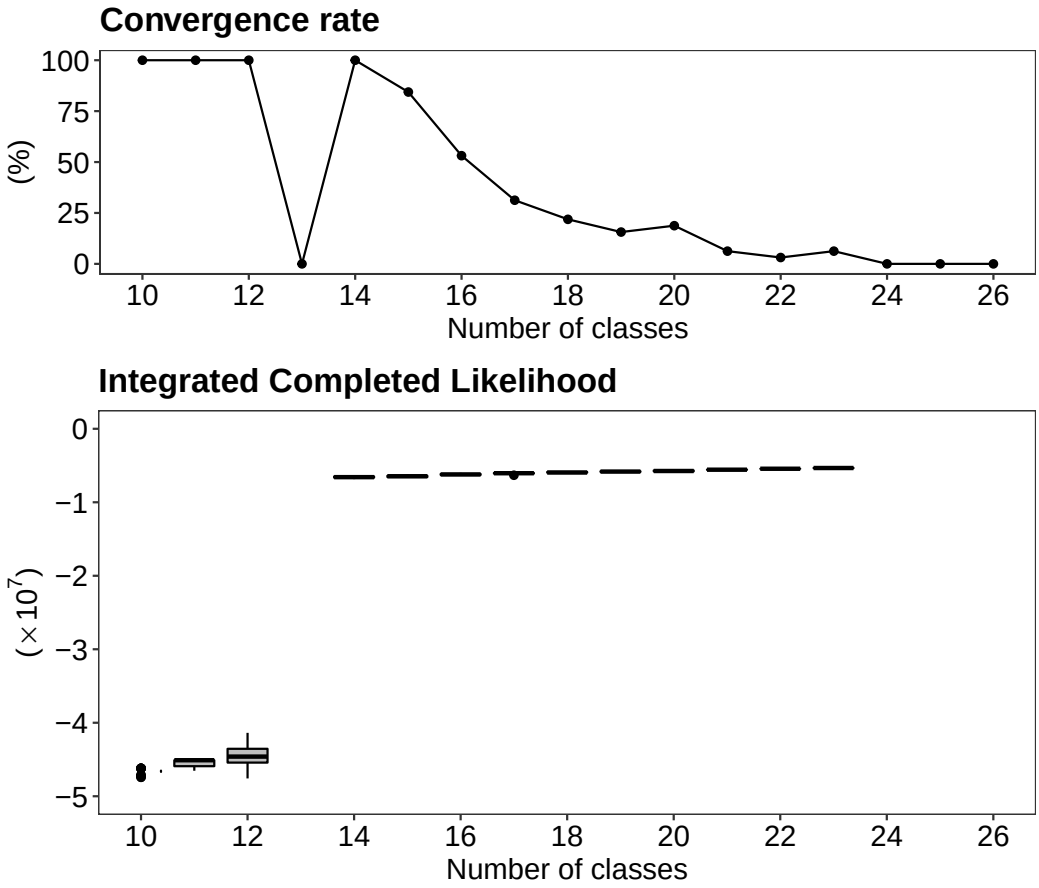
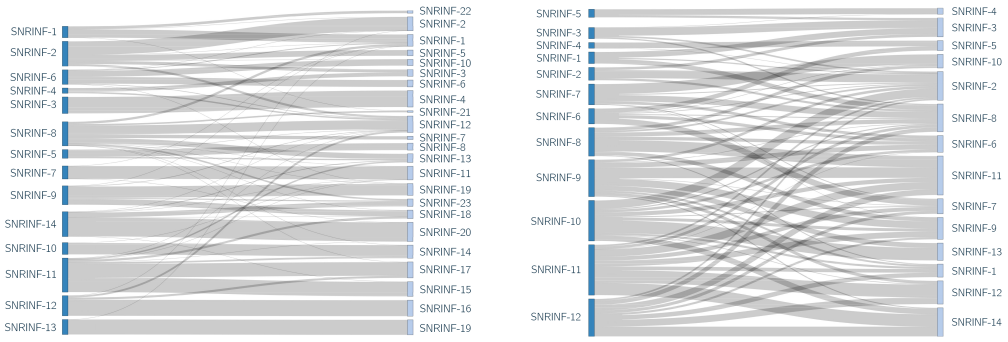


Fig. 3.3.: Clustering analysis of noiseless spectra with Fisher-EM. *Top:* Convergence rate as a function of K. For every K value considered, 32 classifications were calculated. *Bottom:* Boxplots of the ICL are a function of the number of clusters K. The horizontal bars show the median value, the boxes represent the two quartile values, the whiskers extend to points that lie within 1.5 times the interquartile range of the lower and upper quartile, and data beyond are shown individually with dots.



(a) *Left:*K=14. *Right:* K=23

(b) *Left:*K=12. *Right:* K=14

Fig. 3.4.: Comparisons between three classifications of the noiseless spectra "SNRINF": K=12, K=14, and K=23. In panel (a), the composition of the classes of K=14 (left) and K=23 (right) are compared, and K=12 (left) and K=14 (right) are compared in panel (b). The colour boxes represent the classes, the grey lines represent galaxies that are shared by the two classes they link, and the height of the colour boxes is proportional to the number of galaxies in a given class.

In the $K > 13$ regime, the convergence rate decreases as the number of clusters increases, and no convergence is obtained for any $K > 23$. Strictly speaking, the ICL is the greatest for $K = 23$, and from a statistical standpoint, $K = 23$ is therefore the best result, but has a low convergence rate. In addition, the classification at $K = 14$ has no convergence issue and matches the classification at $K = 23$ very well in that they share many classes containing the same galaxies (i.e. they have a similar class composition, see Fig. 3.4a). Undeniably, the K=23 classification is more refined through its 9 additional clusters and a slightly better ICL. Nonetheless, the $K = 14$ classification seems to be a good compromise between reproducibility and goodness-of-fit, and it was therefore chosen to represent the $K > 13$ regime. In addition, the dispersion of the ICL, as shown with boxplots in the ICL-versus-K figure, is found to be very low. This means that Fisher-EM provides consistent classification results on this dataset.

The galaxies are found to be distributed differently in the classifications at $K = 14$ and $K = 12$ (Fig. 3.4b), showing that the two regimes in fact correspond to two distinct classifications, hence the need to study them both.

3.4.2 Spectral classifications

The mean spectra and dispersion of each class of the retained classifications (i.e. $K = 14$ and $K = 12$) are shown in Fig. 3.5 and Fig. 3.6 (the vertical scale is arbitrary, but it is the same for all plots of spectra). Despite its better ICL, the $K = 14$ classification shows a higher dispersion than $K = 12$ for most of the classes, except for a few exceptions (classes 4, 10, and 14), and in some cases, blue and red continua are even mixed (classes 2, 3, 6, and 9). On the other hand, the dispersion within most classes of the $K = 12$ classification is rather small, indicating a decent homogeneity of the classes.

The distribution of the spectra among the 14 and 12 classes is relatively well-balanced (upper-left panels Figs. 3.7 and 3.8), although this is arguably more the case even for the 14 classes. It varies from about 200 to 2000 spectra with an average of circa 500.

Although the classification is done on the basis of the spectroscopic data, one of the major pros of working with simulations is that the results can be analysed and interpreted through the prism of the simulation parameters. This is what I present in the next section.

3.4.3 Parameter distribution among classes

As I explain in section 3.3, the simulated spectra are associated each with a set of parameter values. I explained that the relevant parameters are T_{main}, T_{burst}, τ_{main}, τ_{burst}, f_{burst}, Metallicity, and $E(B\text{-}V)_{cont}$.

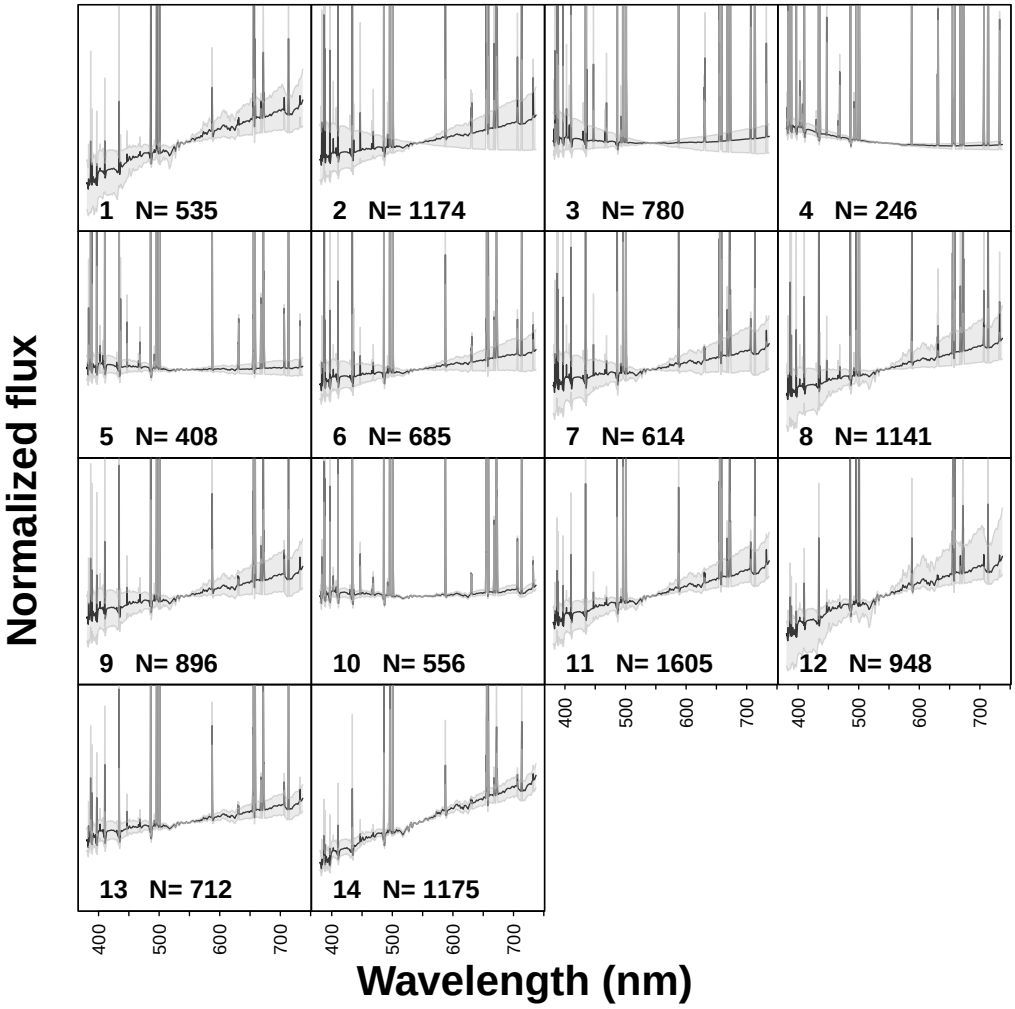


Fig. 3.5.: Fourteen-cluster classification of the noiseless spectra, with the mean spectra (in black) and their dispersion (in grey) for every class. N is the number of members in each class. All the spectra were normalised by their mean values between 505 and 581 nm, a region where the spectra have no emission lines. The scale is the same for all panels and all other figures of spectra throughout this chapter. The dispersion corresponds to the 10% and 90% quantiles for each monochromatic flux. The classes are sorted by ascending average T_{main}.

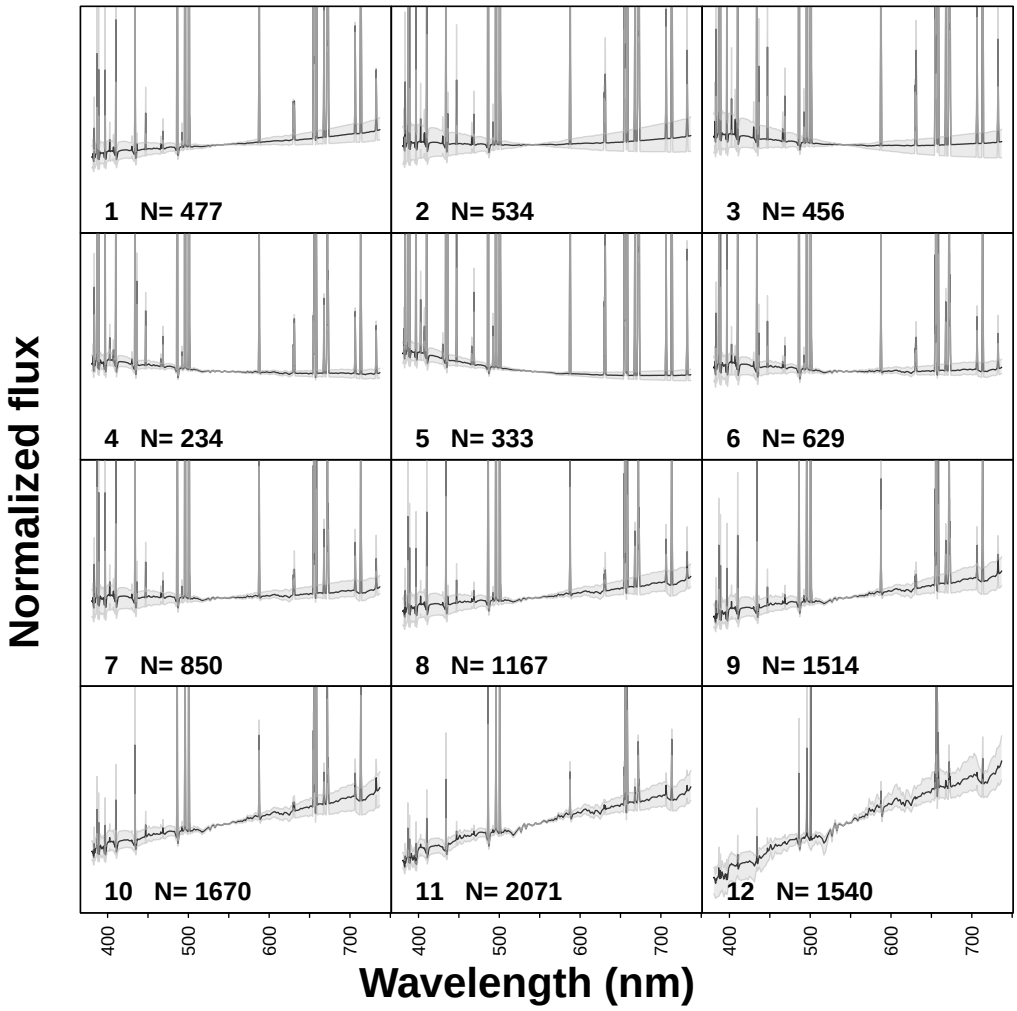


Fig. 3.6.: Twelve-cluster classification of the noiseless spectra (see Fig. 3.5).

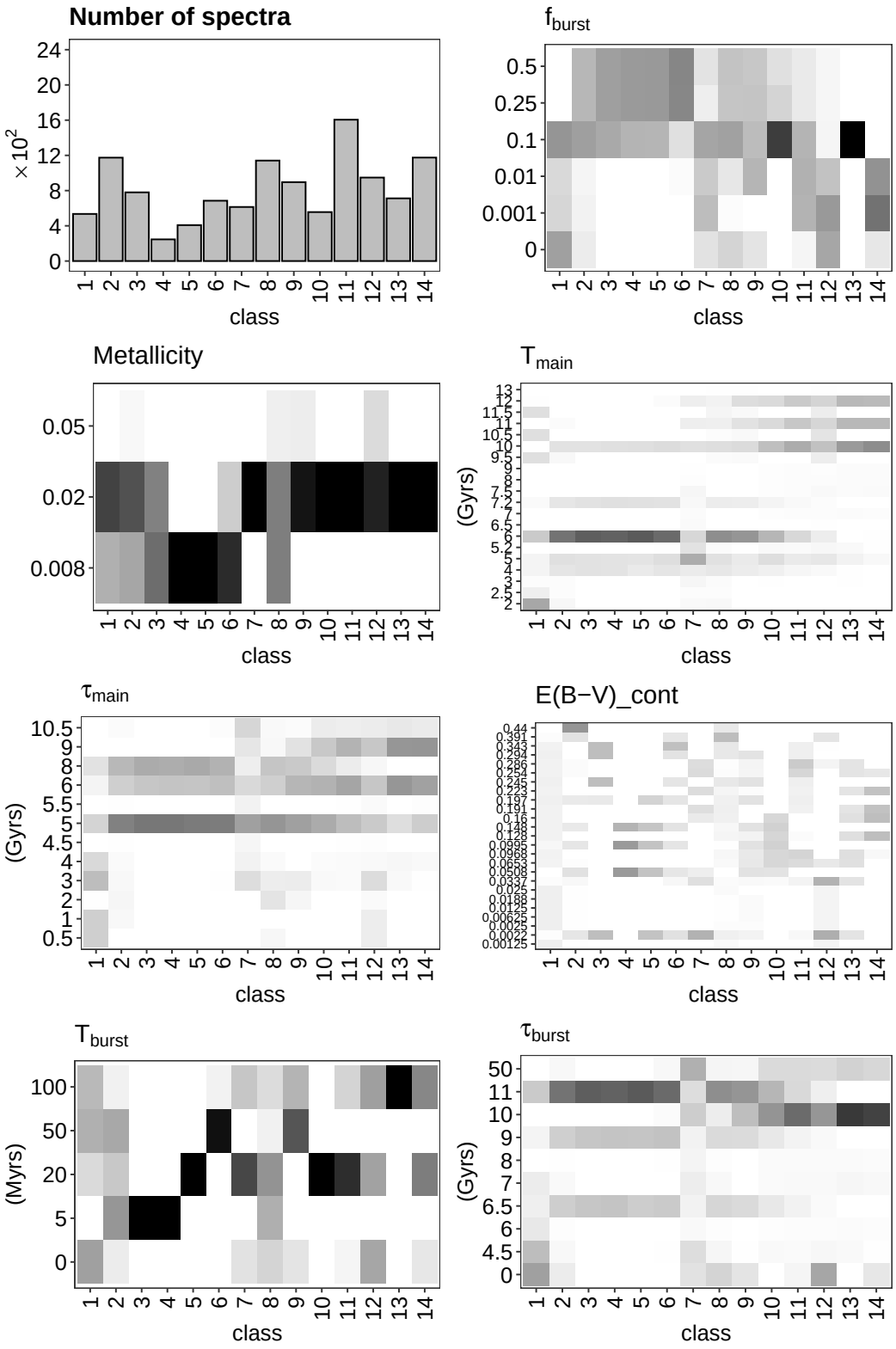


Fig. 3.7.: Fourteen-cluster classification of the noiseless spectra. *Top left:* number of spectra contained in each class. *All others:* heatmaps of the relevant CIGALE input parameters among the 14 classes on noiseless spectra. All possible parameter values (see Table 3.3) are represented on the y-axis, and the class index on the x-axis. The within-class densities of the parameter values are illustrated in the form of a heatmap, where a dark square equates to a density of 1, and white of 0. The classes are sorted by ascending average T_{main}.

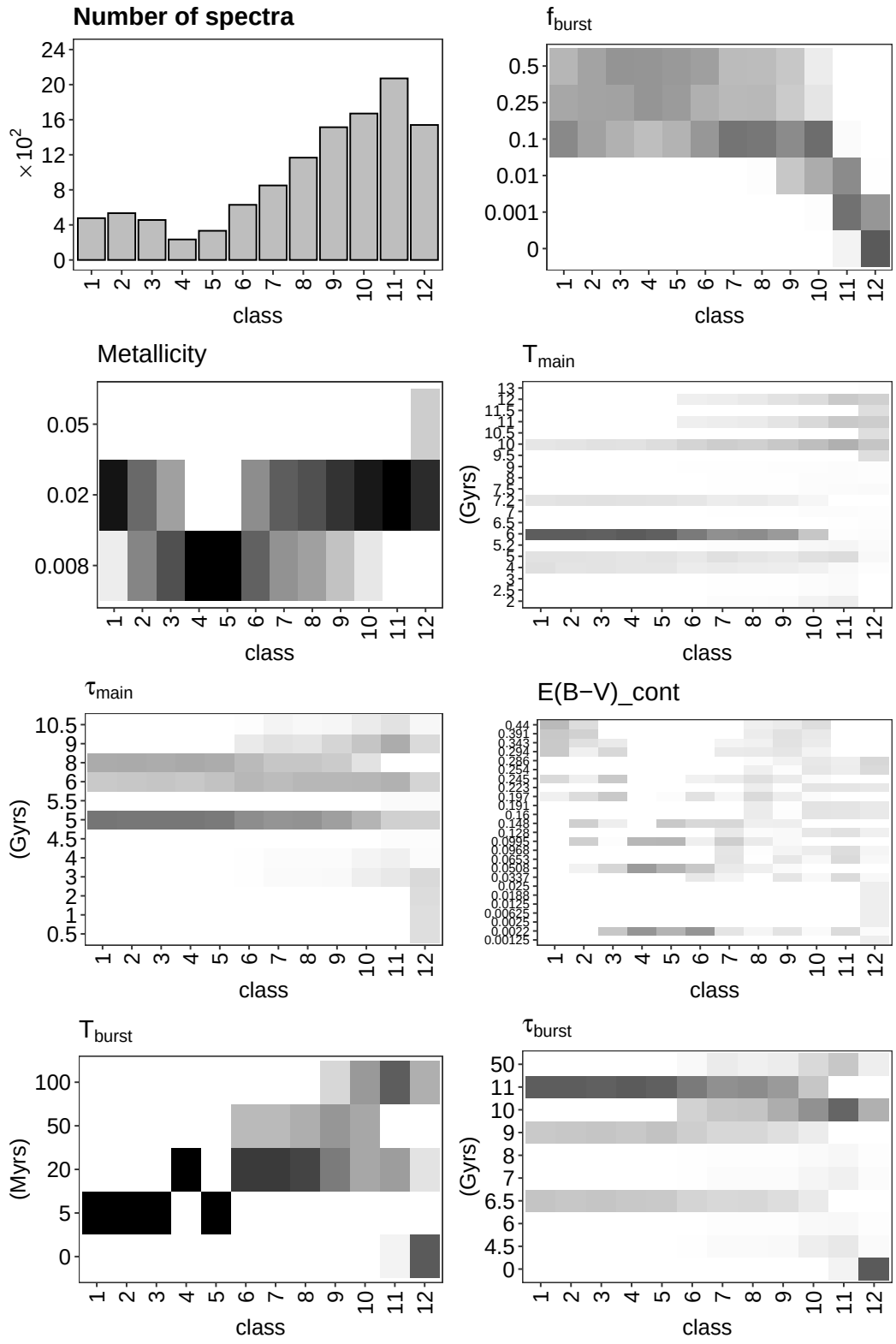


Fig. 3.8.: Twelve-cluster classification of the noiseless spectra (see Fig. 3.7).

A parameter-by-parameter analysis and class-by-class analysis of the classifications is made possible by visualising the parameter distributions inside the $K = 12$ and $K = 14$ classes (Fig. 3.7 and Fig. 3.8). Below, I describe and analyse the results on the $K = 14$ and $K = 12$ classifications successively.

3.4.3.1 Classification at K=14

We observe in Fig. 3.7 that f_{burst} is mostly separated into two categories in this classification: higher values (classes 2, 3, 4, 5, 6, 10, and 13) and lower values (classes 1, 7, 8, 9, 11, 12, and 14). The separation is not very sharp, however, especially for some lower f_{burst} classes. The metallicity is not well sampled as it only takes three discrete values (0.008, 0.02, and 0.05), but is rather well segregated in the classes. Half of the classes gather one single metallicity value, two of them being the lower value (classes 4 and 5), and five being the medium value (classes 7, 10, 11, 13, and 14). The other half contain a mixture of usually two values: lower and medium (classes 1, 3, and 6) or medium and higher (classes 10 and 12), with the exception of classes 2 and 8, which mix all three values. The higher metallicity value is not as well separated as the medium and lower values. The classes were sorted by ascending age T_{main}. A clear distinction is made in the classification between younger and older galaxies. Three categories can be drawn: classes containing the youngest galaxies (1 to 6), classes containing an equivalent mixture of young and old galaxies (7 to 12), and classes of older galaxies (13 and 14). τ_{main} has a high dispersion in the classes, although the lower values happen to be concentrated in the classes with lower f_{burst}. The T_{burst} values are well isolated among the classes with high f_{burst}, with the exception of class 2, which also contains a fraction of low f_{burst} galaxies. In the other classes, where the burst of star formation is less prominent, no distinction is made between T_{burst} values. τ_{burst} is quite dispersed in most of the classes with lower f_{burst} classes. This is slightly less the case for higher f_{burst} classes, but τ_{burst} remains a poorly discriminated parameter in the classification. $E(B\text{-}V)_{cont}$ does not appear to be well separated. Except for a few classes (4, 10, and 14), high and low $E(B\text{-}V)_{cont}$ values are mixed in this classification. All things considered, the classification at $K = 14$ is essentially explained by four parameters (f_{burst}, T_{main}, T_{burst}, and Metallicity), while the other three (τ_{main}, τ_{burst}, and $E(B\text{-}V)_{cont}$) show similar distributions for most of the classes, with a few exceptions.

The properties of the classes can be summarized as follows. Classes 1 to 6 are made of galaxies of younger ages (T_{main}) and a rather significant burst of star formation (f_{burst}). Except for classes 1 and 2, these classes nearly only have one value of T_{burst}: Classes 3 and 4 contain galaxies whose bursts occurred very recently in the star formation history (5 Myr), class 6 contains galaxies of older bursts (50 to 100 Myr), and class 5 medium age bursts (20 Myr). Metallicity values are also fairly well separated in these classes. Galaxies of classes

2, 3, and 6 have metallicities of 0.008 and 0.02, while classes 4 and 5 only contain galaxies with a metallicity of 0.008.

Classes 7 to 12 mostly contain two populations of galaxies: older galaxies with a significant burst of star formation, and younger galaxies. Except for class 8, they all gather galaxies of medium to high metallicity. In this category, class 10 stands out as the galaxies it is made of all have a 20 Myr old prominent burst of star formation, while the other classes do not differentiate T_{burst}.

Finally, classes 13 and 14 gather old galaxies with a faint burst of star formation. More precisely, galaxies of class 14 have had little to no burst in their star formation history, while galaxies of class 13 did, but a long time prior to observation (old T_{burst}).

Overall, each class shows its own specificity in regard to the physics of the galaxies it contains. Three groups of classes can be distinguished (1-6, 7-12, and 13-14) which essentially categorise the galaxies as young and star-forming, less star-forming, and passive and old.

3.4.3.2 Classification at K=12

As shown in Fig. 3.4b, the classification at $K = 12$ is significantly different from the classification at $K = 14$ in terms of spectrum distribution. However, from a physical standpoint, they show very similar characteristics (Fig. 3.8).

They are both mostly driven by f_{burst}, Metallicity, T_{main}, and T_{burst}. Nonetheless, their distribution among the classes shows a significantly lower dispersion for $K = 12$ despite the lower ICL. This is specifically striking for lower values of f_{burst}, which are scattered around many classes and are mixed with higher values for $K = 14$, while they are extremely well separated for $K = 12$.

Class-wise, a similar categorisation as in $K = 14$ can be made, with classes 1-5 corresponding to the young star-forming galaxies, classes 6-10 to a mixture of old and younger galaxies with burst of star formation that occurred 20-100 Myr ago, and classes 11-12 corresponding to the old and mostly passive galaxies. In addition, class 12 gathers almost all galaxies that did not undergo an additional period of star formation, while that information was not retrieved at $K = 14$.

As a whole, the classification at K=12 is more discriminative that the classification at K=14, but they are both sensitive to the same physical parameters. Despite the statistical superiority of K=14, K=12 therefore appears to be a better version of K=14 in terms of their physical discriminative properties.

3.4.4 Linear discriminant analysis

An LDA analysis (see Sect. 2.3.1 from Chapter 2) was applied to the classified data in order to identify the influence of each parameter in the classification process. While the classification was made based on the information of the spectra, LDA analysis uses the information of the parameters. The LDA analysis returns a set of components that are essentially the projection vectors that best separate the classes. Therefore, it highlights the links between them and the classification. The results are shown in Fig. 3.9 and Fig. 3.10, and described below for $K = 14$ and $K = 12$ successively.

3.4.4.1 Classification at K=14

The analysis at $K = 14$ resulted in seven components labelled LD1 to LD7 (Fig. 3.9). The first component explains almost half of the data variance, and adding the next three brings it up to almost 90%.

An overall weight attributed to each parameter by the LDA is obtained by summing the seven components weighted by their relative variance explained. This overall weight quantifies how discriminated the parameters are by the classification. The results agree well with the conclusion obtained in Sect. 3.4.3, namely, the parameters that are best discriminated are T_{burst}, Metallicity, T_{main}, and f_{burst}. The e-folding time of the burst of star formation is by far the least discriminated parameter, followed by the reddening and the e-folding time of the main stellar population.

The first component allocates most of its weight equally to T_{burst}, Metallicity, and T_{main}. The second component is dominated by the age of the burst of star formation (T_{burst}) and, to a lesser extent, by the stellar mass fraction of the burst (f_{burst}). The third component equally distributes most of its weight to three parameters, namely T_{main}, τ_{main} , and f_{burst}. The fourth component is completely dominated by the reddening ($E(B\text{-}V)_{cont}$), which was mostly ignored by the previous component. The final three components are mostly dominated by parameters that are already greatly taken into account in the first three components, but they also allocate some of their weight to the e-folding times of the SFR of the main stellar population and the burst event (τ_{main}, τ_{burst}), which were entirely insignificant in the previous components.

3.4.4.2 Classification at K=12

As expected given their similarities, the LDA analysis at $K = 12$ shows similar results than $K = 14$ (Fig. 3.10). There is a significant difference between LD4 and LD6, however. Because the explained variance of those components is so small, however, they have little to no impact on the overall weight of the parameters.

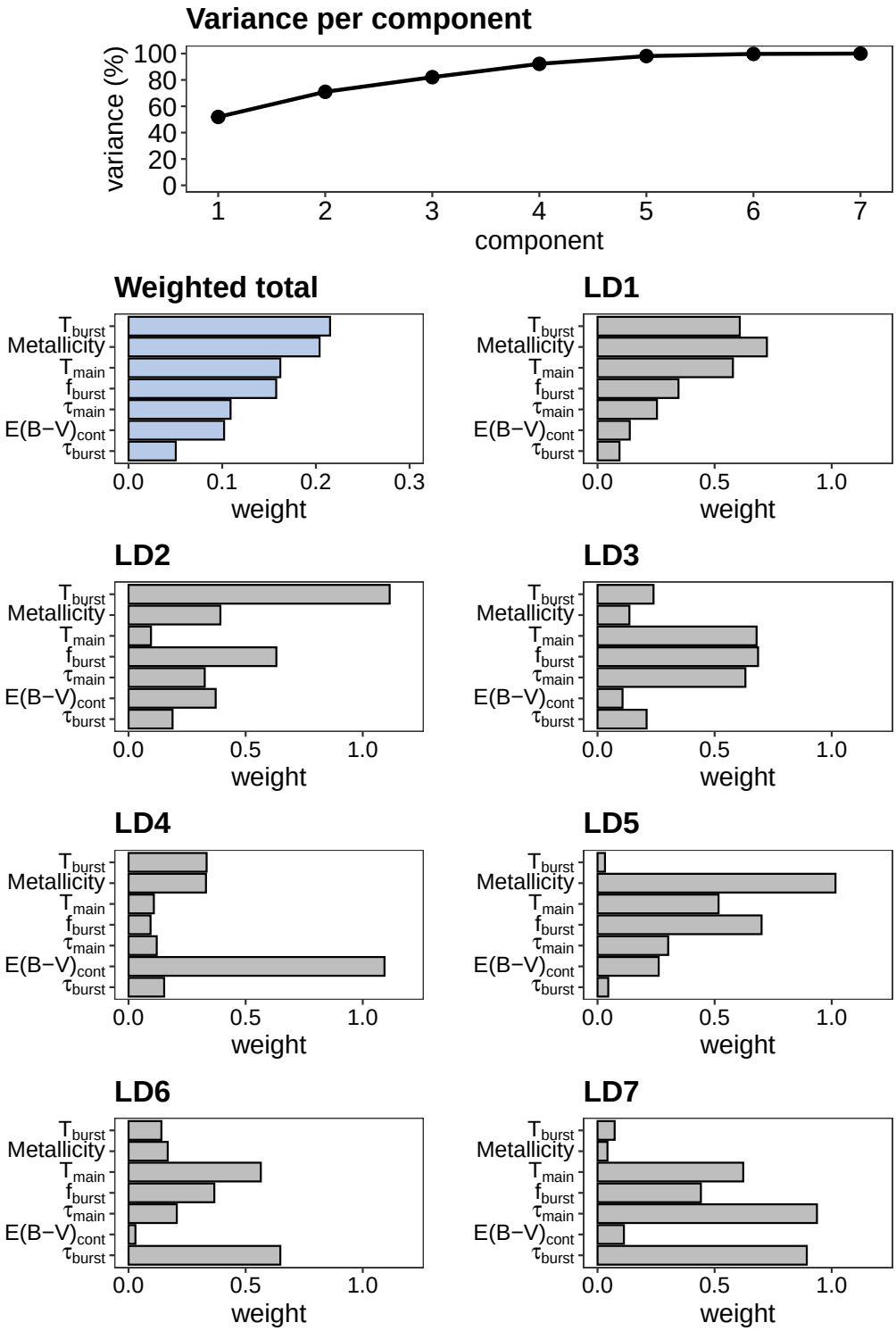


Fig. 3.9.: Linear discriminant analysis on the classification of noiseless spectra at $K = 14$. *Top:* Cumulative data variance described by the linear discriminant analysis components. *LD1 to LD7:* Weight of each parameter for components 1 to 7 of the linear discriminant analysis. *Weighted total:* Cumulative weight of each parameter among the seven components, weighted by the percentage of data variance described by each component.

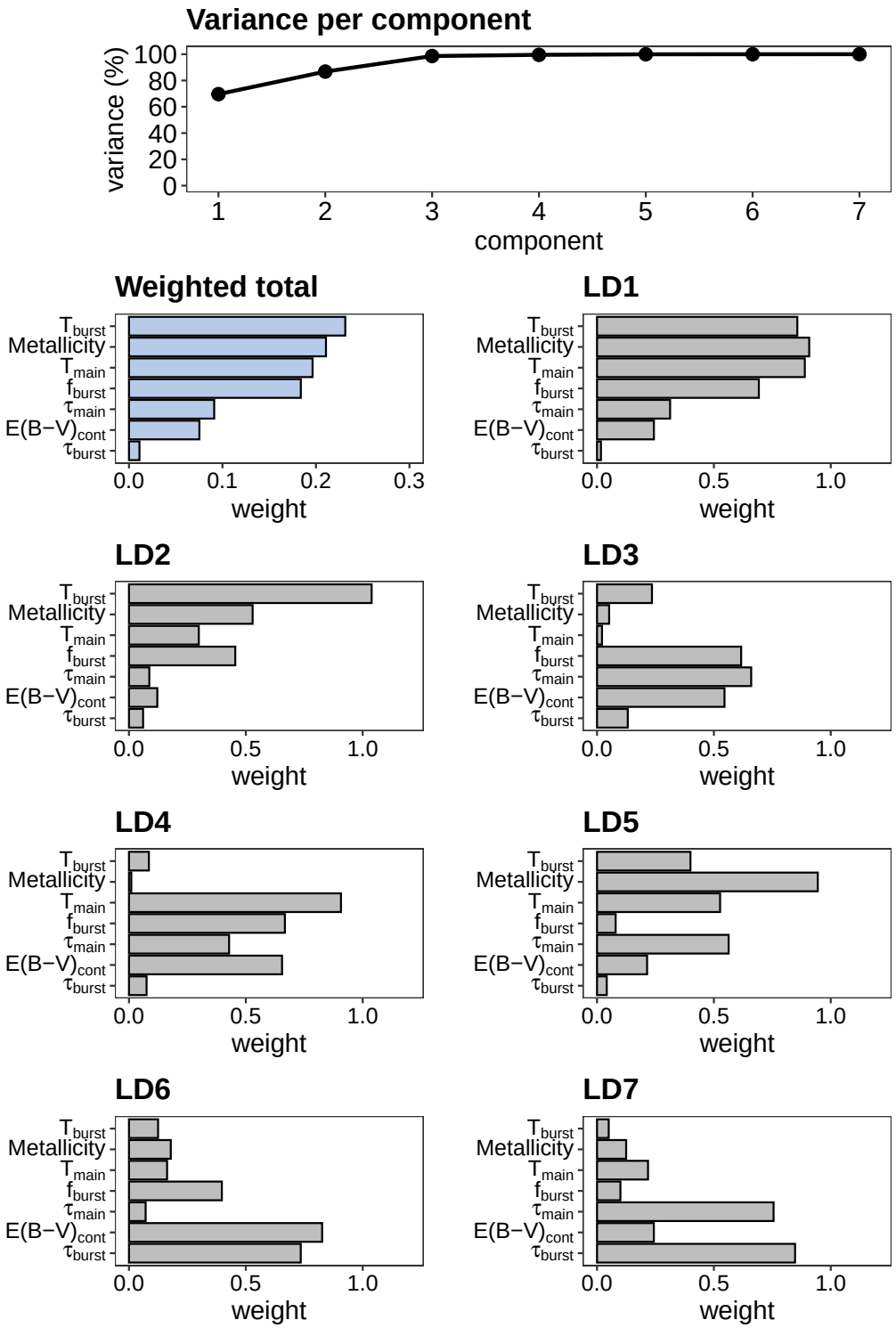


Fig. 3.10.: Linear discriminant analysis on the classification of noiseless spectra at $K = 12$ (see Fig. 3.9).

3.5 Analysis of the noisy spectra

In this section, we study the effect of noise on the classification of the spectra for different values of S/N. To do this, a Gaussian noise of constant S/N was added to the 11 475 spectra. We used seven of values for the S/N: 1, 3, 5, 10, 20, 100, and 500.

3.5.1 Optimal number of clusters

The very characteristic break of the ICL curve observed at K=13 for the noiseless spectra disappears as soon as noise as low as S/N=500 is added. The ICL curves obtained have an optimum independently of the noise level (Fig. 3.11), as opposed to the ever-increasing ICL on noiseless spectra.

For S/N≤100, the ICL reaches its maximum for K=11 to 13. In addition, convergence is reached every single time for any K smaller than 30-40, depending on the noise level.

Our study of the noise shows that the ICL curves at different S/N differ from that of the noiseless case, but they all agree and yield the same optimal number of clusters around K=12. Furthermore, we show that the optimal classifications on the spectra with added noise closely resembles the $K = 12$ one on noiseless spectra, whether it be based on the composition of the classes (Fig. 3.12) or their median spectrum (Fig. 3.13).

At S/N=500, the ICL sometimes shows another behaviour that fully depends on the random generation of the noise. Therefore, while for most generated noise the ICL showed an optimum, one particular noise vector led to an ever-increasing ICL curve (until loss of convergence) that resembles that of the noiseless spectra. Thus, we suspect that at S/N≥ 500 the noise may be insufficient to blur out the data sparsity, which is likely responsible for the ICL break and convergence issues on noiseless spectra.

In the rest of this chapter, the classification at S/N=20 and K=12 is taken as a reference for the data with added noise; all the plots will focus on this specific S/N, but the results for the other S/N are available in Appendix B.

3.5.2 Parameter distribution among classes

The optimal classification at S/N=20 is essentially identical to that on the noiseless spectra presented in Sect. 3.4.3.2 in regard to the parameter distribution in the classes (Fig. 3.14). Slight differences appear nonetheless, highlighting the loss of information induced by the additional noise. For example, there is no longer a class isolating 20 Myr old star bursts. On the other hand, some specificities in the spectra appear to be retrieved more accurately

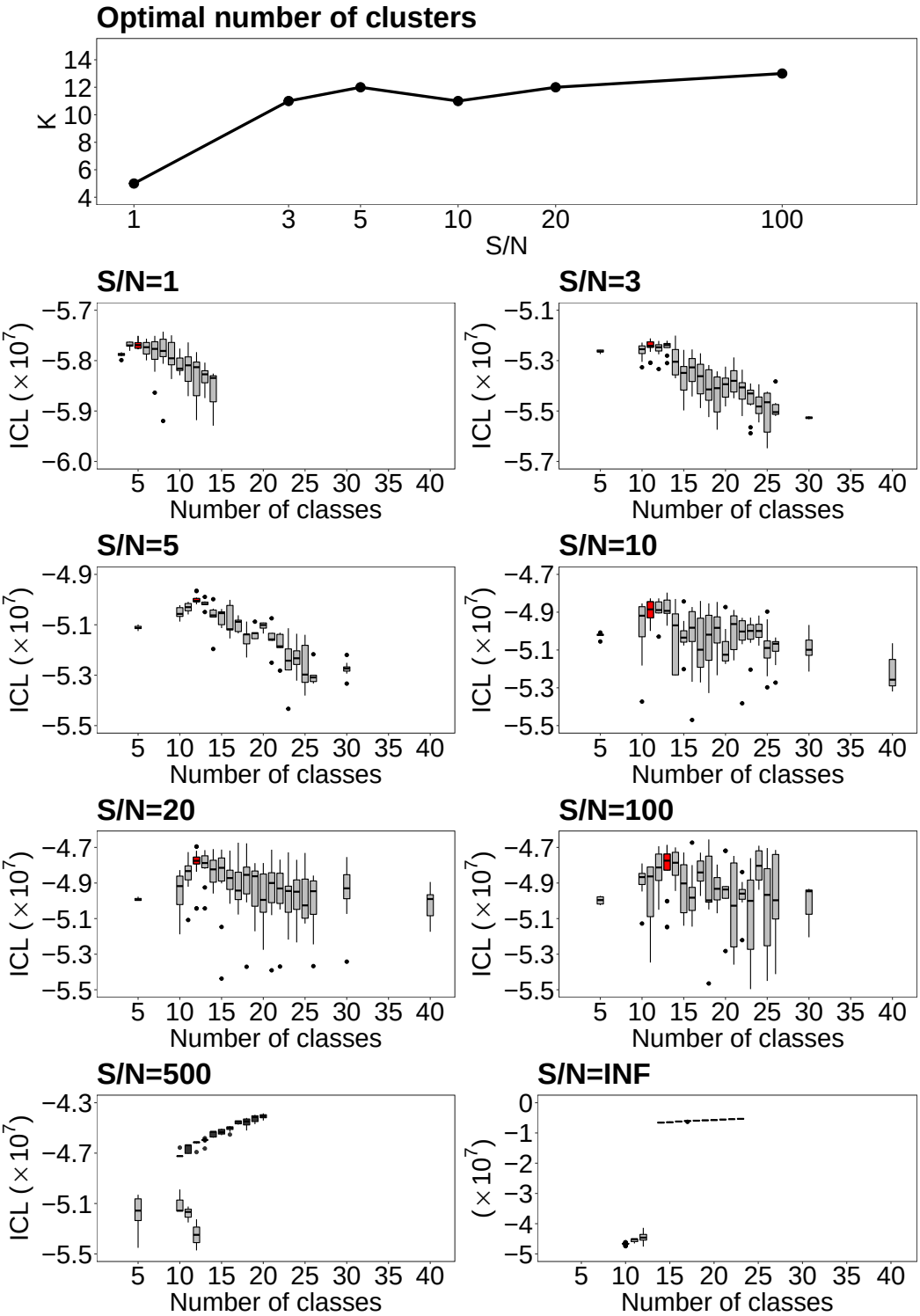


Fig. 3.11.: Summary of ICLs and optimal number of clusters as a function of S/N. *Top:* Optimal number of clusters across the different noise levels. *All others:* ICL as a function of K for different noise levels. In each of the panels, the red boxplot highlights the maximum median value of the ICL i.e. the associated best-fit K value. For the S/N of 500, two behaviours were observed depending on the randomly generated noise. They are both illustrated by the two sets of boxplots (black and grey) in the corresponding panel.

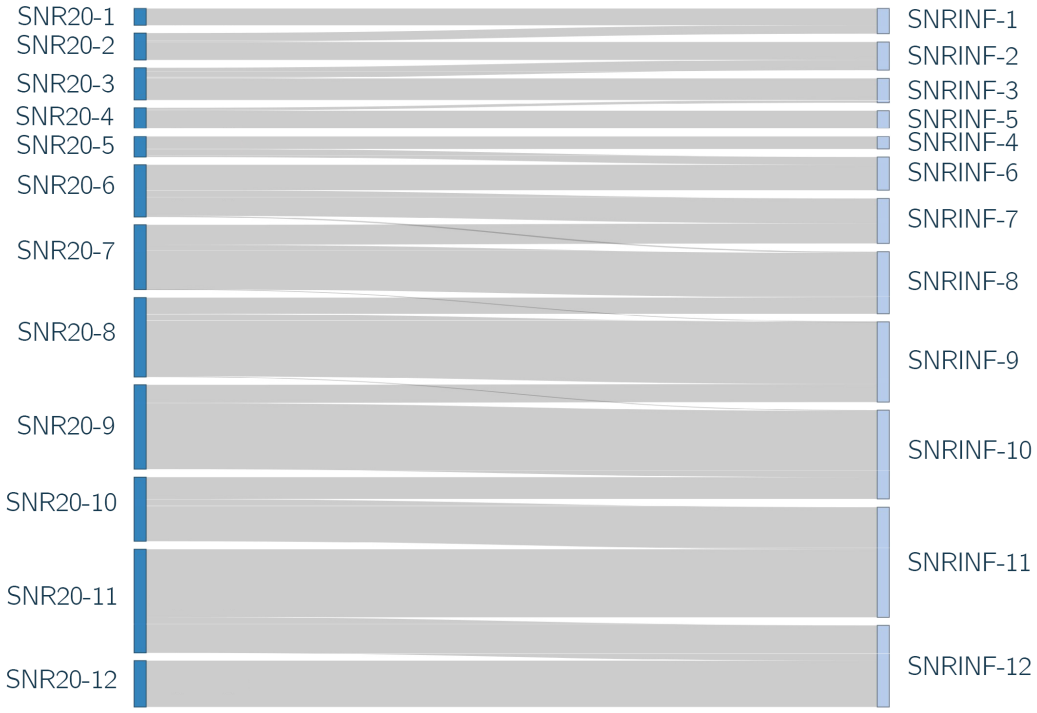


Fig. 3.12.: Same as Fig. 3.4 between the $K = 12$ classification on noiseless spectra (right) and on spectra with an added noise of S/N=20 (left).

with the addition of noise. For instance, five classes contain a unique value of metallicity, as opposed to three classes on noiseless spectra. Lower values of f_{burst} are also more sharply separated.

At greater noise (Appendix B), this enhanced ability to discriminate some parameters fades out, and the dispersion in the classes increases for all parameters. At S/N=3, the method is still capable of distinguishing burstless galaxies from burst-heavy ones, but lower non-zero values of f_{burst} become more erratically distributed around the classes. Metallicity is not as well discriminated either, as only a single class of unique value remains. T_{main}, and T_{burst} to a lesser extent, is still separated in a similar fashion despite the significant amount of noise. At S/N=1, the ICL optimum shifts from 12 to 5 classes, but the method is still capable of approximately separating old and passive galaxies from young and star-forming ones.

3.5.3 Linear discriminant analysis

The LDA applied on the noisy spectra with S/N=20 shows that the first three components completely dominate the analysis (60%, 20%, and 10%), whereas five components were significant for the noiseless spectra. The first component is similar to that of the noiseless spectra, with a more heavily weighted f_{burst} parameter, nonetheless. The second and third components are distinctively different, however. It appears that the weight of the parameters was shifted from one component to another: T_{burst}, f_{burst} , and τ_{main} are less significant in

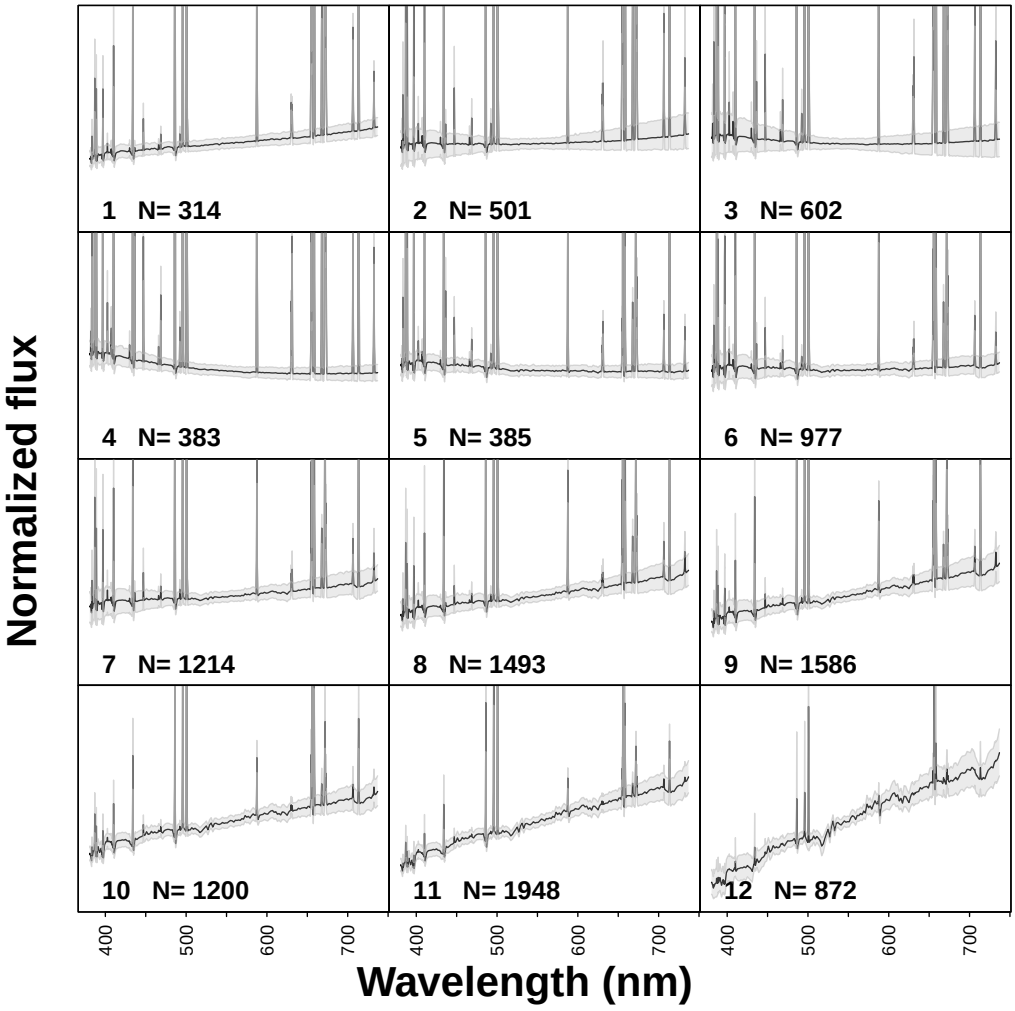


Fig. 3.13.: Twelve-cluster classification obtained on the spectra with added noise of S/N=20 (see Fig. 3.5).

the second component, but more important in the third component. Likewise, T_{main}, τ_{main} , and $E(B\text{-}V)_{cont}$ have higher weights in the second component and lower weight in the third component. While some components are indeed different, the overall relevance of each parameter in regard to the classification remains almost unchanged.

3.6 Discussion

3.6.1 Origin of the K≥13 regime

The analysis of the noiseless spectra revealed an odd behaviour at K≥13: a lack of convergence at K=13, and a high plateau of the ICL for K from 14 to 23 (Fig. 3.3). In an ideal world, the ICL curve as a function of the number K of clusters should show a clear peak because this criterion penalises higher number of classes to avoid 'overfitting' (otherwise the 'best' way to classify the sample would be to assign each spectrum to its own class). The ICL curve does not always have this ideal behaviour, however, especially in high dimensions (e.g. Fraix-Burnet et al., 2021). In particular, it often shows a plateau that ends when the algorithm fails to converge (i.e. encounters an empty cluster, see Sect. 3.4.1).

Because the classifications for K>13 differ from the classifications obtained for K=12 with and without noise, we suspected that the reason might be that the data are simulated with a necessarily limited coverage of the parameter space. To test this hypothesis, we devised a simple toy model at a lower dimension to reproduce this behaviour. After trials and errors, we found a dataset that is described in Appendix C. With this dataset, we observe a similar phenomenon: the ICL curves show an optimum at K=4, with solutions at K=2 but no convergence at K=3. Remarkably, if we add a very small amount of noise to a part of one of the five variables, solutions are found in all cases, thus reproducing what was observed with the CIGALE data.

Even if this toy model cannot be considered as a proof, we conclude that the behaviour found in the noiseless data may be due to some peculiar distribution of the data in the parameter space, that is, too small a dispersion and probably a significant level of sparsity. In addition, Jouvin et al. (2021) reported a poor performance of the Fisher-EM results in the case of very little noise, a behaviour that they were unable to explain but hypothesised to be related to insufficient constraints brought by the dataset. Such problems are known to occur in EM-GMM-based clustering (e.g. Kasa and Rajan, 2020).

The case of K≥13 is therefore thought to be an artefact resulting from the simulated nature of the spectra, and is dismissed in the rest of the discussion. Instead, K=12 is considered as the representative classification of the noiseless spectra. We stress that the noiseless situation cannot be encountered in reality.

3.6.2 Physical discrimination capacity of unsupervised classification

The classification at K=12 shows classes of spectra that are very homogeneous with little dispersion, demonstrating the ability of Fisher-EM to find structures in a high-dimensional data space. This has been noted before for the much larger SDSS sample (Fraix-Burnet et al., 2021).

The analysis of the distribution of the parameters used in the CIGALE simulations shows that this discriminative power among the spectra is also visible in the physical properties of the galaxies. Four of the seven parameters are clearly well discriminated (T_{main}, T_{burst}, f_{burst}, and Metallicity), and to a lesser extent, τ_{main} and $E(B\text{-}V)_{cont}$. This important result shows that Fisher-EM is capable of picking up the expected relevant physical parameters. The LDA analysis confirms these most influential parameters. The fact that the weights of T_{burst} and Metallicity appear stronger than that of T_{main} could be due to their small sampling density (three and four values, respectively). However, f_{burst} has only six values and has a similar weight as T_{main}. Moreover, the latter has a higher weight than τ_{main} despite having a similar distribution. Lastly, the LDA analysis shows that τ_{burst} has a weak impact on the classification, but this is probably not due to its small sampling density since it was expected from the physics itself (Sect. 3.3).

As a conclusion, we have shown that each class not only has a specific spectral shape, but its members also have specific physical properties. Hence, in a real dataset, a detailed analysis of the mean spectra of the classes should reveal these properties and transform an unsupervised classification into an objective and physical atlas of galaxy spectra.

3.6.3 Effect of the noise

The addition of noise raises two questions that we address in this section: i) whether it changes the classification itself, and, ii) whether it changes the physical interpretation.

3.6.3.1 Effect on the classification

Adding noise to our spectra strongly modifies the ICL curve by revealing a clear maximum around K=12 (for S/N≥3) with a quasi-identical classification, the noiseless one included. At S/N=1, the optimum is K=5 so that a higher level of noise tends to smear out the classes, and as expected, lessens the discriminative capability of the analysis.

The presence of noise in the data also tends to facilitate the convergence of Fisher-EM. At the S/N we considered, the convergence issues that were encountered in the noiseless case

were non-existent. Lack of convergence was still observed in the noisy case when a high number of classes was chosen, but this behaviour is usual and was seen on real data as well (Fraix-Burnet et al., 2021).

3.6.3.2 Effect on the physical meaning of the classes

Our study shows that our unsupervised classification and its physical meaning are essentially unchanged between S/N=3 and S/N=500 for our simulated data, demonstrating its robustness. At S/N=1, most of the physical discriminative capacity of the method is lost, and only five classes are found. Nonetheless, these five classes are not meaningless, and remarkably separate star-forming from passive galaxies rather well. Jouvin et al. (2021) showed that in some cases, Fisher-EM was even capable of accurately classifying data with a S/N as low as -1dB.

3.7 Conclusion

This study shows that the unsupervised classification algorithm Fisher-EM applied on thousands of CIGALE galaxy spectra yields a classification that is both robust against the initialisation of the algorithm and against the noise. Very importantly, the classification is very discriminating with respect to the physical properties of the galaxies.

Unsupervised classification in astrophysics is still in its infancy, and the first robust classification of spectra of galaxies have been published very recently (Fraix-Burnet et al., 2021). The aim of such an objective classification is to produce an atlas that is entirely data driven and that could be used later with supervised learning in large surveys. Even though the preliminary interpretation of the classes found in Fraix-Burnet et al. (2021) has shown their physical relevance, we here confirmed that unsupervised machine learning is able to yield not only a robust statistical classification, but also a physical classification of the properties of the galaxies from the spectra.

The main advantage of the unsupervised classification is that we do not add any a priori physical information into the classification process, but rely on the ability of the algorithm to detect the structures that are statistically relevant, not those we would wish it to detect. We are thus not limited by the representativeness of the training set, and we consequently avoid all the associated biases. In addition, as shown in Fraix-Burnet et al. (2021), the classification is characterised by its DLM model (see Chapter 2 Sect. 2.5) which can be retrieved and used to quickly classify new data. This process performed through the E-step only, which is extremely fast. In a sense, this means that the Fisher-EM algorithm has effectively built its own training set (Fig. 3.6) that happens to be physically well characterised (Fig. 3.8).

Our result is a strong encouragement to analyse the atlas proposed in Fraix-Burnet et al. (2021) in more depth and extend it to larger samples. The exciting perspective is to include galaxies at higher redshifts in order to study the evolution of the classification with time through a fully data-driven procedure.

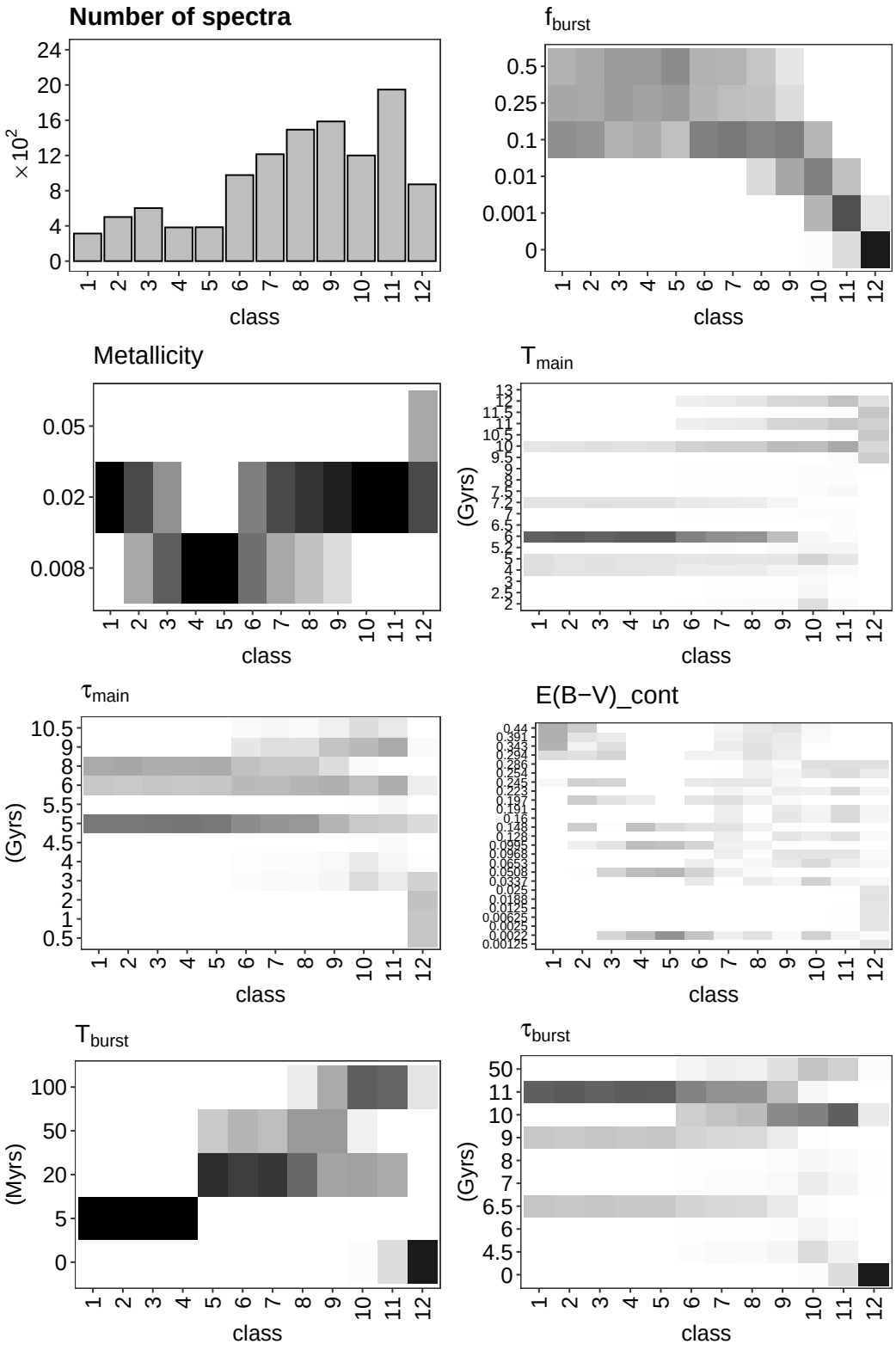


Fig. 3.14.: Twelve-cluster classification on the spectra with an added noise of S/N=20 (see Fig. 3.7).

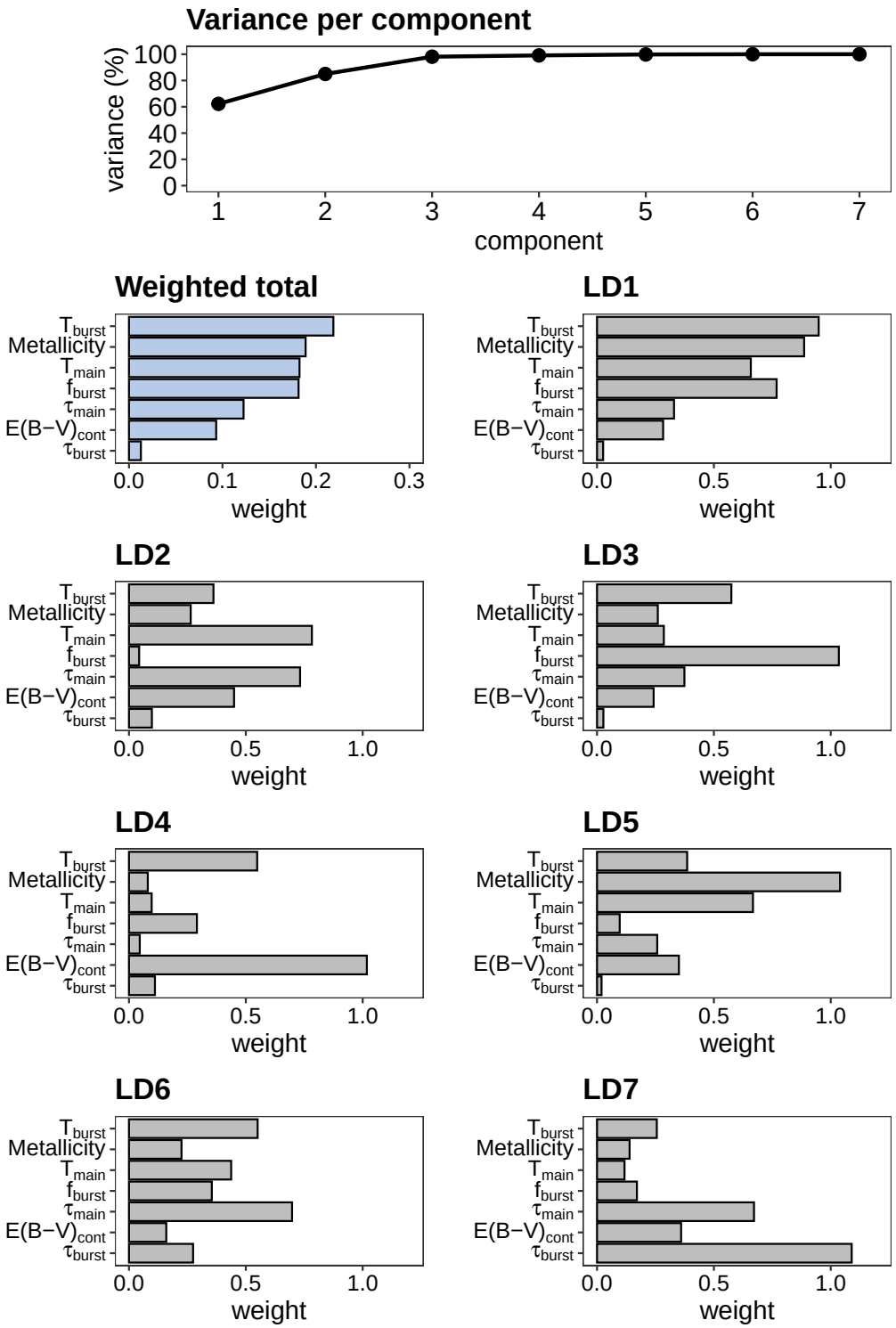


Fig. 3.15.: Linear discriminant analysis on the classification of spectra with added noise of S/N=20 (see Fig. 3.9).

4

Unsupervised classification of 0.4 < z < 1.2 galaxies with the VIMOS Public Extragalactic Redshift Survey

4.1 Introduction

In the previous chapters, I showed the need to work towards a new spectroscopic classification of galaxies (Chapter 1). I highlighted how machine-learning, and particularly, Fisher-EM, is a powerful tool that is well suited for this difficult task (Chapter 2). I showed, using simulated spectra, that this algorithm was able to provide physical classifications of

galaxies on the basis of monochromatic fluxes (Chapter 3), and that it was consistent and robust despite the presence of noise.

This chapter focuses on a part of my thesis that is a direct continuity of the work of Fraix-Burnet et al. (2021), who successfully led a data-driven spectroscopic classification of galaxies using Fisher-EM. They conducted this study using a sample of 700 000 galaxies in the local universe within $z < 0.25$ from the SDSS. They found a total of 86 classes, 37 of which contain 99% of the galaxies. The promising results from this work encouraged us to extend the study to higher redshifts to observe how the spectroscopic classification would change throughout cosmic time, and this is precisely what I did. This work is described in an article under preparation (Dubois et al., 2023).

This project was conducted using data from the VIMOS Public Extragalactic Redshift Survey (VIPERS), from which I extracted a sample of about 80 000 galaxy spectra contained within a redshift of $0.4 < z < 1.2$. Similarly to what I presented in Chapter 3 or what was done in Fraix-Burnet et al. (2021), this work is conducted on optical spectra. This piece of work was in many ways more challenging than what I did on simulations. Obviously, real observations are a lot more intricate than simulations. The "real" data is intertwined with e.g. instrumental effects, noise, contamination by the atmosphere, which all have to be addressed carefully to obtain the best results.

The chapter is organised as follows. In Sect. 4.2, I present the data and its challenges. In Sect. 4.3 I discuss the different steps that were led to prepare the data before the classification. I will show that this procedure is slightly more complicated than what was done in the previous chapter, hence the dedicated section. In Sect. 4.4, the tuning process of Fisher-EM and the difficulties I have encountered in that regard. In Sect. 4.5, I show the different results that were obtained, and discuss their interpretation and limitations in Sect. 4.6. Lastly, I conclude on this work in Sect. 4.7.

4.2 Data

The classification was done on the sole basis of the spectroscopic data from the VIPERS survey, which I present in Sect. 4.2.1. However, additional data was necessary to understand and interpret the results. This includes photometric observables, inferred physical parameters, and measurements of spectral features, which I all describe in Sect. 4.2.2. Finally, in Sect. 4.2.3 I explain specificities of the data that made it challenging to classify.

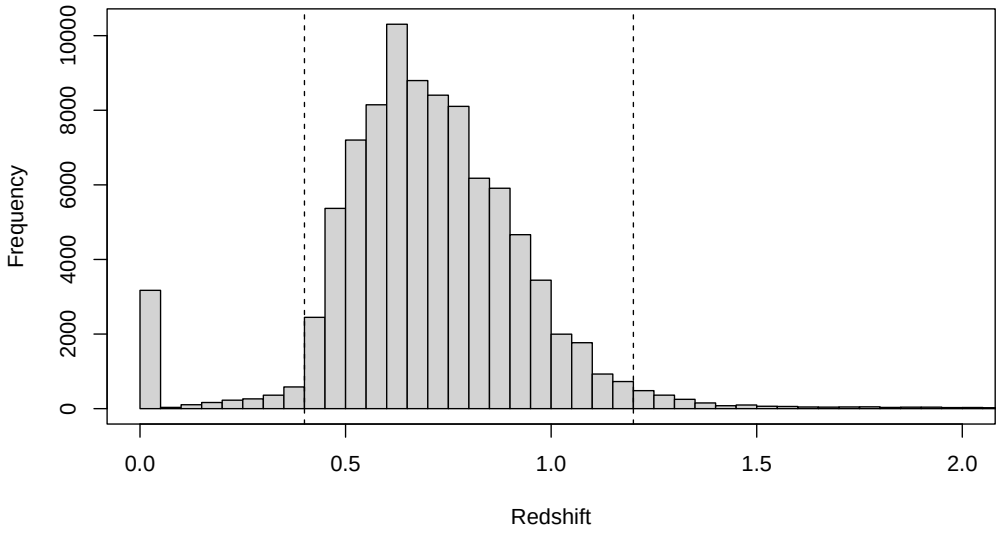


Fig. 4.1.: Histogram of redshift distribution within the VIPERS-DR2. The vertical dashed lines show the upper and lower limit of the subset selection, respectively at $z = 0.4$ and $z = 1.2$.

4.2.1 VIPERS survey

This study uses the public data release 2 (PDR-2) of VIPERS, a spectroscopic survey carried at the ESO 8m Very Large Telescope targetting galaxies from the CFHTLS Wide Photometric survey (Scodeggio et al., 2018). The sample contains 86 775 spectra of objects brighter than 22.5 mag in the z~1 vicinity. The spectra cover wavelengths from 5500 Å to 9500 Å, and were observed at a resolution of R=220 within an overall sky area of 24 deg^2 (in the so-called the W1 and W4 VIPERS fields). For further details on the survey, refer to Garilli et al. (2014), Guzzo et al. (2014) and Marchetti et al. (2017).

For this work, I used a subset of the VIPERS-DR2 that I made based on several criteria. First, I used the target selection flag provided in the VIPERS database to remove unwanted objects, which include stellar-like objects, saturated and/or objects brighter than 17.5 mag, and those dimmer than 22.5 mag (based on the VIPERS main galaxy targets criteria). Then, I limited the subset to galaxies of redshifts ranging from 0.4 to 1.2. The vast majority of the sample is contained within that scope (Fig. 4.1), and removing the tails of the distribution is somewhat unavoidable due to the strategy chosen to analyse the data that I will explain in more depth in Sect. 4.3. To summarize it quickly, I chose to divide the selected dataset into several subsamples of similar redshifts. As such, if I were to keep the tails of the distribution, it would lead to subsamples at lower and higher ends which could not be used for classification purposes due to their low sample sizes. Finally, the last criterion for choosing the subset was the quality of the redshift measurement. I excluded galaxies whose spectroscopic redshift measurement are not judged as secure as per the redshift flag provided in the database. The

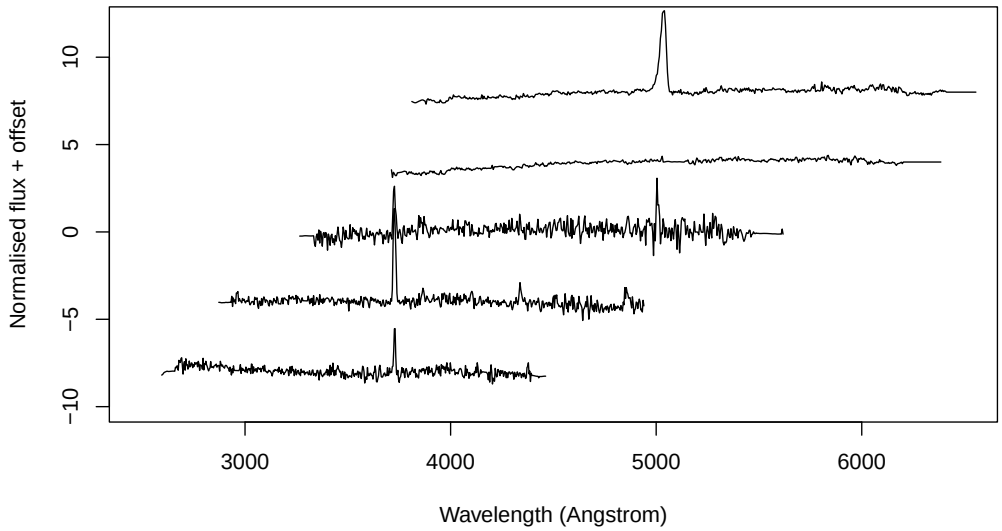


Fig. 4.2.: A few examples of spectra from the selected subset. For visualisation purposes, the spectra are normalised and offset by an arbitrary value. From top to bottom, the spectra correspond to galaxies of redshift 0.4, 0.5, 0.7, 0.9 and 1.1.

selected galaxies either have a confidence level of 99% on the spectroscopic redshift, or a 90% confidence level but within at least 2σ of the photometric redshift measurement.

The resulting subset contains 79 224 galaxies and AGNs with a median redshift of 0.7. Some example spectra are shown in Fig. 4.2 to highlight the diversity of the sample. We of course observe a diverse range of continuum shapes and spectral features, e.g. several emission and absorption lines, and also line broadening. The most prominent features are the Balmer break (hereafter D4000), and the emission lines [OII], [OIII], H_α and H_β. Since the goal of the survey was to probe spectroscopic redshifts, the observations were not necessarily obtained with long exposures. As a result, the spectra are overall pretty noisy, and smaller spectral features are often hidden by the noise in individual spectra. On a side note, we can see on the example shown that there is a significant variance in the noise from one spectrum to another. Additionally, the wide scope of redshift implies that the rest-frame wavelengths are spread from the end of the optical region to the near-UV, and depending on the redshift of the source, some spectral features will not necessarily be probed. For instance, in the two bottom spectra of the example, the [OIII] lines (around 5000 Å) are out of range (Fig. 4.2).

4.2.2 Physical parameters

While the classification procedure was led using nothing more than the spectroscopic data, I also used a set of inferred physical parameter (Tab. 4.1) to interpret classification results.

Tab. 4.1.: The set of inferred parameters associated with the VIPERS-DR2 dataset. They include spectral features, photometric magnitudes, and parameters obtained from SED-fitting.

Spectral feature	**Wavelength** (Å)
D4000	3850 – 4100
OII	3727, 3729
H_β	4861
OIII	4931, 4959, 5007
H_α	6563
Magnitude	**Wavelength** (Å)
M_{FUV}	1350 – 1750
M_{NUV}	1750 – 2800
M_U	3104 – 3972
M_B	3980 – 4920
M_G	3960 – 5480
M_V	5070 – 5950
M_R	5664 – 7144
M_I	6987 – 8541
M_Z	8487 – 10025
M_Y	9600 – 10800
M_J	11135 – 13265
M_H	14765 – 17835
M_K	19950 – 23850
Parameter	**Description**
SFR	Estimation of the star formation rate
T	Estimation of the age of the galaxy
M_s	Estimation of the stellar mass

This set of parameters was retrieved by the VIPERS team, and was used e.g. in Siudek et al. (2018a), who kindly shared it with me.

The data include 5 spectral features, being the Balmer break and the four most prominent lines. All four line measurements include the integrated flux, the equivalent width, the full width at half maximum. The Balmer break is computed as the ratio of the mean flux measured in 4000-4100 Å and in 3850-3950 Å. Magnitudes from the VIPERS Multi-Lambda Survey (Moutard et al., 2016a) are also included, as well as stellar masses, SFRs and age derived from SED-fitting. This was done using the code "Le Phare" (Arnouts et al., 1999; Ilbert et al., 2006; Moutard et al., 2016b), a similar algorithm to CIGALE that was described in the previous chapter (see Sect. 3.2 of Chapter 3). Estimations of the uncertainty for all the above are also provided.

4.2.3 Challenges

While more exciting to study, real observations come with their load of issues and challenges. Contrary to simulated data, observations result from an actual instrument which brings its own features to the data it collects (e.g. fringing, calibration issues). Additionally, observing from the ground implies the collected data will contain sky features. Since the classification method is fully data-driven, it looks for discriminative patterns within the spectra regardless of their physical or non-physical nature. It is thus important to make sure the results will reflect that actual intrinsic properties of the galaxies rather than instrumental effects or any other irrelevant information.

The VIPERS-DR2 provides cleaned spectra, from which sky lines and bright secondary sources have been subtracted. However, this sky subtraction is not perfect, and residuals remain in the cleared data. Similarly, bright secondary sources are removed, but some residual features may still be found in the data. As explained in Marchetti et al. (2017), thanks to a PCA analysis, they were able to retrieve the regions where the cleaning residuals are too significant, i.e. above an empirically chosen threshold of $1.2\sigma - 1.8\sigma$. An example is shown in Fig. 4.3. Once identified, these regions are masked and some of them were reconstructed artificially. The spectra reconstruction was mostly done using a method based on PCA analysis (for more details, see Marchetti et al., 2017), although a small fraction

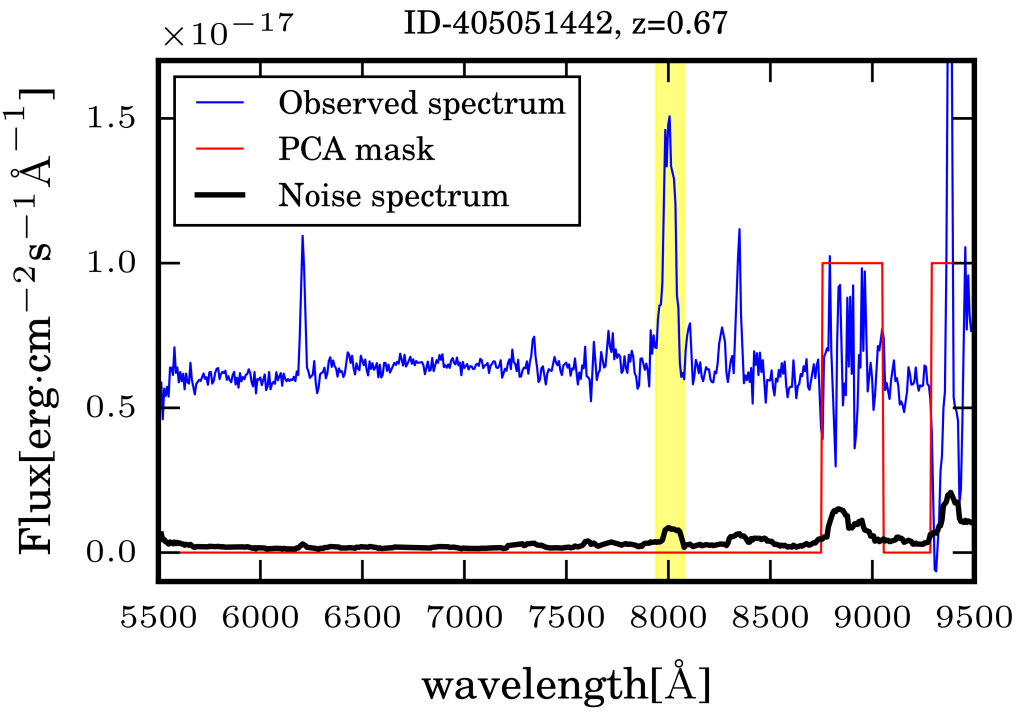


Fig. 4.3.: Figure from Marchetti et al. (2017) showing an example of residuals on a VIPERS spectrum. The yellow highlighted part shows a region with residuals from secondary source removal, and the PCA-masked regions shown with the red line correspond to sky residuals.

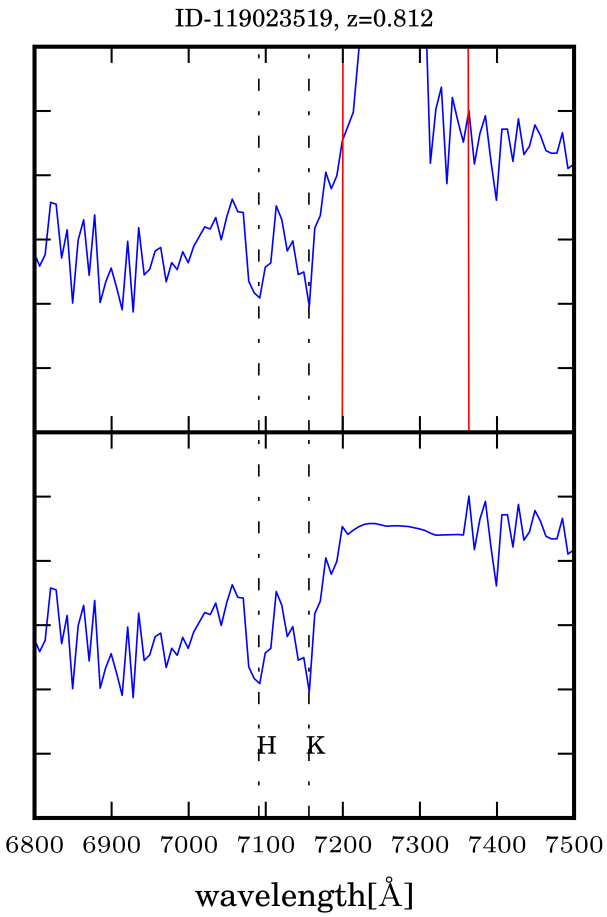


Fig. 4.4.: Figure from Marchetti et al. (2017) showing an example of spectrum reconstruction. The upper panel shows a VIPERS spectrum with strong residuals in the regions delimited by the red lines, and the lower panel shows the PCA-reconstructed spectrum.

was reconstructed with a manual interpolation. In any case, the reconstructed spectra are not recommended for scientific analysis, as the method is not capable of accurately reconstructing spectral features other than continuum. We can also note that the reconstructed chunks visually stand out from the rest of the spectra by their lack of noise, and the sporadic presence of some peculiar features (Fig. 4.4). Such particularities could be interpreted by the classification algorithm as discriminative features, although they are non-physical and irrelevant to the study. To prevent this, there is no other choice but to mask these regions and not account for them in the analysis. And while this is a significant issue on its own, it is amplified by the fact that the dataset probes a wide range of redshift. In fact, the sky lines, since they come from Earth's atmosphere, are independent of the source's redshift, and are thus always going to be observed at the same wavelengths. The spectral features of the sources, on the other hand, are redshifted, meaning that the regions affected by sky residuals in the sources' rest-frame spectra change with redshift. However, the classifier we use requires that all spectra must be sampled identically, meaning that if one spectrum has a masked portion, this mask has to be transposed to all other spectra too, leading to an enormous loss of information when extended to the entire dataset.

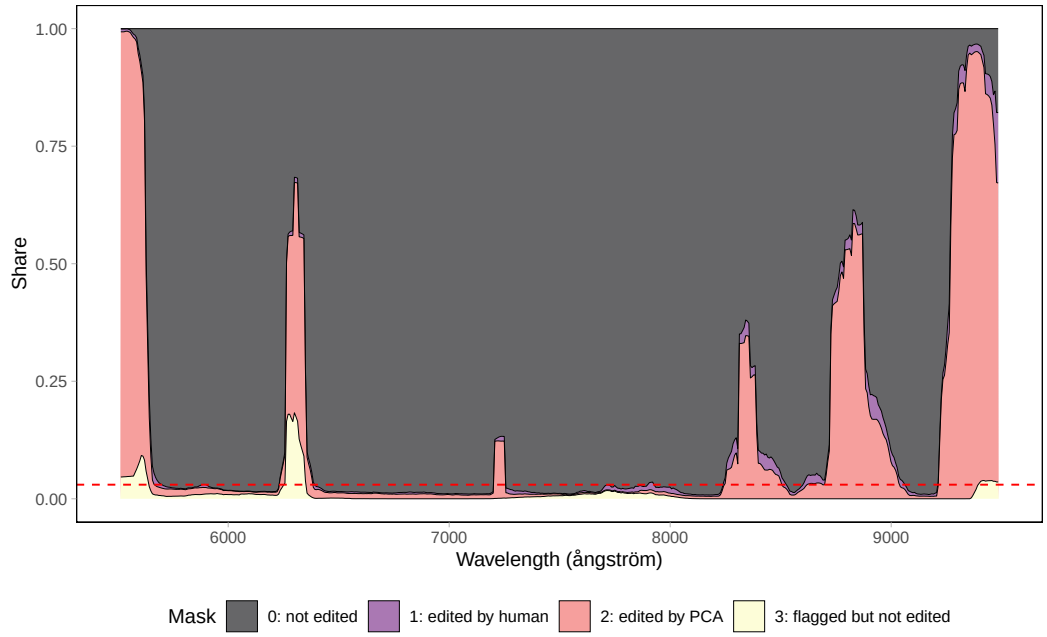


Fig. 4.5.: Fraction of masked and reconstructed spectra per monochromatic flux in the observed frame within the selected VIPERS dataset. The red dotted line represents the threshold of 5% above which a monochromatic flux is to be masked for the whole sample.

Unfortunately, the constraints brought by the masks are far from negligible, as highlighted in Fig. 4.5. In fact, the lower and upper ends of the spectra are almost constantly masked due to instrumental fringing, and significant sky residuals remain in 4 main regions for 10 to 70% spectra. Additionally, on this figure we can see that even outside significant sky contamination regions, the share of masked spectra is never really null. By setting a masking threshold at around 5% (red dotted line), we end up with 37% of the monochromatic fluxes masked in the observed frame. And when shifting the spectra back to their rest-frame, the masks end up covering virtually the entire spectral range due to the wide redshift spread.

Additionally, the wide range of redshifts the sample covers brings another, more obvious, issue to the table: some features are visible in certain sources, but redshifted out of the observed spectral range in some others. The instrument gathered spectroscopic data from about 5500 Å to 9500 Å in the observed frame, which translates to a range of 3900-6800 Å in rest-frame for the closest sources ($z = 0.4$), and 2500-4300 Å for the furthest ones ($z = 1.2$). As such, the OIIλ3727 line, for instance, is out of scope for galaxies of redshift $z < 0.5$. Similarly, the H_β line is out of scope for galaxies of redshift $z > 0.95$. In other words, all the spectra within the sample do not share the same spectral features depending on their redshift. As such, if we were to classify the whole sample, the resulting classes would essentially separate the galaxies by groups of redshift, which is somewhat pointless. In addition, since the classifier requires all spectra to sample the same monochromatic fluxes, if we were to work on the entire sample all at once, we would have no other choice but to limit the

data to the common rest-frame spectral range for all observations, i.e. 3900-4300 Å, hence throwing out most of the data.

Finally, as shown in Fig. 4.2, the S/N can vary a lot from one observation to another. Even though I show in Chapter 3 that Fisher-EM is still capable of providing good results despite the presence of high levels of noise, mixing data with different levels of noise can be problematic. For example, small features may be visible in certain spectra, and hidden by the noise in others, hence potentially skewing the discriminative capacity of the algorithm.

All these specificities in the data impose unavoidable constraints on the strategy to be adopted for the classification process. The sample cannot realistically be classified as a whole, and a different approach has to be taken. Alongside the regular data preparation similar to that presented in the two previous chapters, I discuss in the following section the solutions that were found to deal with these challenges.

4.3 Data preparation

As explained in Sect. 2.5 of Chapter 2, pre-processing the data is a crucial and mandatory step before using the classification algorithm Fisher-EM. Here, I applied the regular data preparation procedure detailed in Chapter 2, which I summarize in Sect 4.3.1. However, as I introduced in the previous section, the dataset is challenging in several aspects, and requires additional manipulations that I detail in Sect. 4.3.2 and Sect. 4.3.3.

4.3.1 Regular procedure

The spectra were first and foremost shifted back to their rest-frame using the spectroscopic redshift measurements provided in the VIPERS-DR2. This ensures the spectral features are aligned from one spectrum to another, which is necessary for Fisher-EM to provide relevant classifications. However, unlike the simulated dataset presented in Chapter 3, the VIPERS rest-frame spectra do not exactly sample the same monochromatic fluxes. Hence, it was necessary to resample the spectra. This was done using the same method as in De et al. (2016) and Fraix-Burnet et al. (2021) which consists in doubling the sampling rate beforehand to preserve spectral features. Finally, the rest-frame spectra were normalised by their mean value between 4150 Å and 4250 Å, a spectral region with no prominent features and that is common to the entire sample.

4.3.2 Masks and subsamples

As shown in Sect. 4.2.3, the VIPERS dataset cannot realistically be classified as a whole. The cleaning residuals, instrumental optical distortions or fringing are too significant in certain part of the spectra to be kept in the analysis, and must therefore be masked. These masked regions differ from one spectrum to another due to redshift, but the masks have to be applied to all the spectra due to constraints from Fisher-EM, which can result in a significant loss of information. To minimize the share of masked wavelengths, it is thus necessary to split the sample into smaller bins of redshift.

I chose to define the subsamples such that the cosmic-time step between two consecutive bins is constant. I originally also considered simply splitting the data into bins of constant z-width, but it leads to less evenly distributed sample size due to the asymmetry of the redshift distribution. Additionally, a constant cosmic-time step makes more sense from an evolutive perspective, as we thus progress linearly back in time through the history of the universe from bin to bin. The smaller the cosmic-time step is, the less information is lost from masking. But decreasing the time-step also decreases the number of galaxies in each bin. As such, a compromise must be found between conserving as much information as possible and having large enough bins to sample properly the diversity and proportion of the populations of galaxies. The latter is difficult to assess without a priori knowledge of the expected populations of galaxies. However, the bin size is also constrained from an algorithmic standpoint. In fact, most generative clustering methods require the sample size n to be greater than the number of sampled wavelength p (see Chap. 18 of Hastie et al. (2009) for an overview of this problem). While Fisher-EM can overcome that limit, I chose to use it as a lower bound for the sample size. Under the constraints of minimising the constant time-step between successive bins and having $n > p$ for each bin, the optimum was found to be 26 bins separated by a time-step of 163 Myr (see Table 4.2 in Sect. 4.5).

Despite the efforts to limit the loss of information due to masking, the mask-rate of the rest-frame spectra remains close to 50%, meaning that roughly half of the spectra are masked before being classified. No matter what, this number cannot be brought down lower than the mask-rate of the observed spectra, i.e. 36%. All things considered, while very high, a mask-rate of 50% is acceptable. Lastly, because the VIPERS sample is not uniformly distributed in redshift, the sample size varies by one order of magnitude between the smallest (bin 1 with 491 galaxies) and biggest bin (bin 11 with 5504 galaxies).

With this approach, the goal is not to obtain a single classification of the whole VIPERS dataset. Instead, the idea becomes to analyse each subsample separately, i.e. obtain one classification per subsample. Then, it will be possible to compare the classifications from bin to bin and look at how certain characteristics of the classes evolve over cosmic time.

4.3.3 Denoising

The VIPERS spectra can be quite noisy, as shown in Fig. 4.2. This is not an issue per se, since Fisher-EM deals rather well with noisy data, but I still explored the possibility of denoising the data using wavelet transform. While I was able to significantly improve the S/N of certain spectra, the procedure inevitably creates small line-like features in the spectra, and it is not so clear whether this would help improve the classification results in the end. Given the strategy chosen to deal with the VIPERS data revolves entirely around avoiding sky residuals, instrumental artefacts and peculiar features from PCA reconstruction, I concluded it would not be coherent to proceed with wavelet denoising without extensively and carefully investigating its interaction with Fisher-EM. And although it would be worth exploring, I chose not to spend time on this and abandon the idea.

4.4 Tuning Fisher-EM

Once the data has been pre-processed, the next step is to find the best statistical model, parameters and settings for Fisher-EM. This is an exploratory process; every dataset is different, so there is no absolute best model. Instead, one has to explore the parameter space and the settings options while looking to maximise the likelihood. In this section, I present how I tuned Fisher-EM for the VIPERS dataset and some issues that were encountered in the process. In Sect. 4.4.2, I discuss the selection of the statistical model (DLM, number of classes). In Sect. 4.4.1, I describe some convergence issues that this dataset was causing, and how this problem was overcome.

4.4.1 Initialisation and convergence issues

The first computations that I ran on the VIPERS data were unfortunately quite problematic, and Fisher-EM struggled to find solutions. In fact, most of the time, it ended up outputting a result with a log-likelihood of $-\infty$, which is usually associated with an empty cluster. Of course, these classifications could not be considered valid since Fisher-EM was not behaving as it should, and it was necessary to investigate the issue before proceeding any further. These problems were unexpected, and it took a while to figure them out, since everything worked fine prior to this, both on simulated data, and on real observations from another survey. I explored many possibilities: I tried reverting to previous R and Fisher-EM versions, running the computations on different machines, with subsets of varying sizes, played with different settings of Fisher-EM, but to no avail.

Some say that insanity is doing the same thing over and over and expecting different results, but perhaps insanity is a great scientific asset because surprisingly enough, it is precisely

how I managed to understand and solve this issue. In fact, after running several computations on the same dataset with the same settings, I observed that, while Fisher-EM converged towards an empty-cluster solution most of the time, it would find a regular solution here and there. This is a strong indication that the issue comes from the initialisation, since it is the only part of Fisher-EM that can be stochastic. There are four initialisation options available within Fisher-EM (see Sect. 2.5 of Chapter 2): personalised input, random, hierarchical, or k-means. Up until then, only the k-means initialisation was used. It is the recommended method, and it had been working well so far. It runs the k-means algorithm (that I explain and illustrate in Sect. 2.4.1), which is non-deterministic, N times, and keeps the highest ICL solution as the starting point for Fisher-EM. In most circumstances, it makes sense to use this method as it is able to provide a decent initialisation for Fisher-EM, hence allowing it to converge faster towards its final solution. However, the unwanted side effect, which is what I encountered with VIPERS data, is that it may indirectly guide Fisher-EM towards a local minimum. This hypothesis was confirmed by running the computations again with a random initialisation, which almost completely resolved the problem. The last bit of issues I still had concerned the outermost bins, i.e. the ones with the lowest sample size. Although the bin selection was made under the constraint of $n > p$, or in other words, the sample size being always greater than the number of monochromatic fluxes, the lasting computation issues come from the fact that we do not always have $n >> p$. And while this would be an issue with most clustering methods due to ill-conditioned matrix inversion (known as the high dimension and low sample size problem), Fisher-EM is capable of overcoming this issue through a mathematical trick involving kernel theory, granted the option is activated. I do not detail the mathematics behind this kernel trick in my manuscript, but the explanations can be found in Bouveyron and Brunet (2012). In the end, the combination of the use of the kernel trick with a random initialisation solves entirely the convergence problems that I initially encountered.

With these issues being solved, the last missing step to obtain the final results consists of finding the best statistical model for the given datasets. This includes the optimum number of clusters, and the best performing discriminative latent mixture model. In the next section, I detail this model selection process, and additionally show how a 2-step classification approach appears to be necessary for this dataset.

4.4.2 Model selection

To recap what I explain in Sect. 2.5 of Chapter 2, Fisher-EM models the data as a mixture of Gaussian distributions within a discriminative subspace. This model is called a discriminative latent mixture (DLM) model, and one can choose among 12 variations of it (their names and specificities of the 12 DLM variations are shown in Tab. 2.1 from Chapter 2), which stem from putting more or less constraints on the DLM's degrees of freedom. There

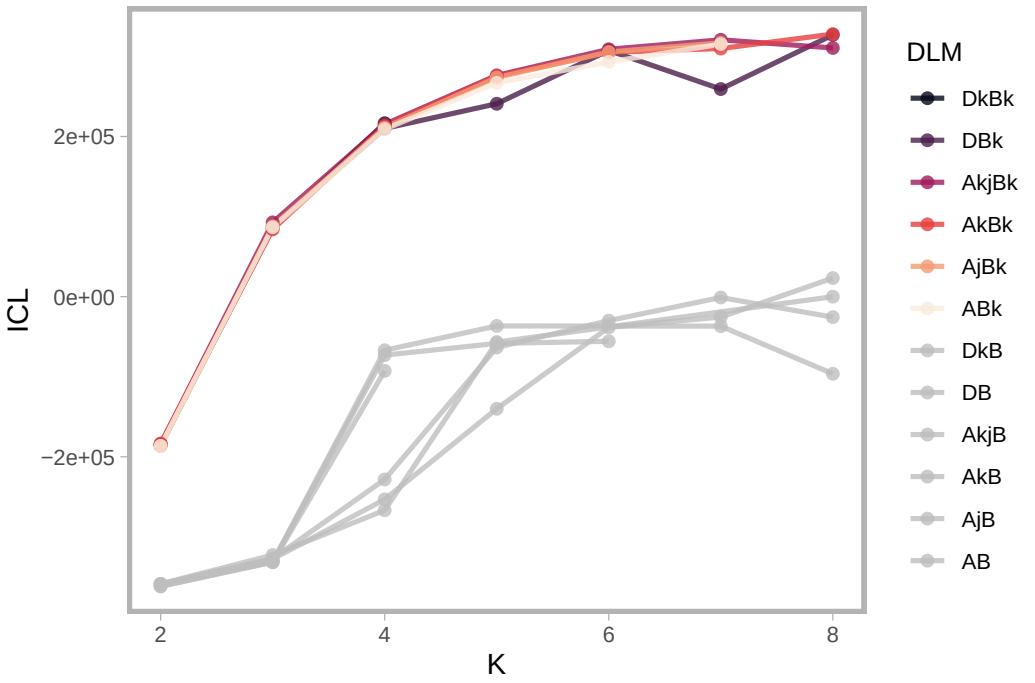


Fig. 4.6.: Graph of the ICL as a function of the number of classes K obtained on the 3rd bin ($z = 0.45$). The ICL is equivalent to a likelihood, and the higher it is, the greater the fit. The colour scheme shows the 12 DLM variations, thus revealing the best performing models. The "–B" models, which perform poorly on the VIPERS data are shown in gray, and the "–Bk" models, which perform the greatest, are shown in colour.

is no obvious rule that dictates the DLM choice, and it is thus necessary to compare the performance of the different DLM model and select the most suited one given the dataset.

In theory, it would be possible to compute Fisher-EM with the 12 DLM models on all of 26 bins subsamples, and pick the best performing model each time. However, this approach would be very computation-heavy, and would take an unnecessary amount of resource and time. Instead, I ran complete computations with all 12 DLM models on 2 extreme bins (one at $z \sim 0.4$ and the other at $z \sim 1.2$), and compared the performance of the DLM models in both cases. In fact, the noise level and spectral features contained in the spectra change gradually from one bin to another due to redshift, which may potentially affect the model performances. The conclusion is that we observe a similar behaviour on the two extreme bin, and it is thus reasonable to assume it would also be observed on all 26 subsamples. Two groups of DLM models can very clearly be distinguished based on their performance: the "–Bk" models, and the "–B" models (Fig. 4.6), which refers to how the noise is modelled within the DLM. As I show in Chapter 2, half of the DLM variations have a class-specific modelling of the noise (the "–Bk"), while the other half have a common modelling of the noise for all the classes (the "–B"). Here, it is very clear that a class-specific representation of the noise fits much better with the VIPERS data, and the "–Bk" models thus largely outperform the other models. This is most likely a consequence of the disparate S/N within the VIPERS spectra that I highlight in Fig. 4.2; the freedom of adjusting the modelling of

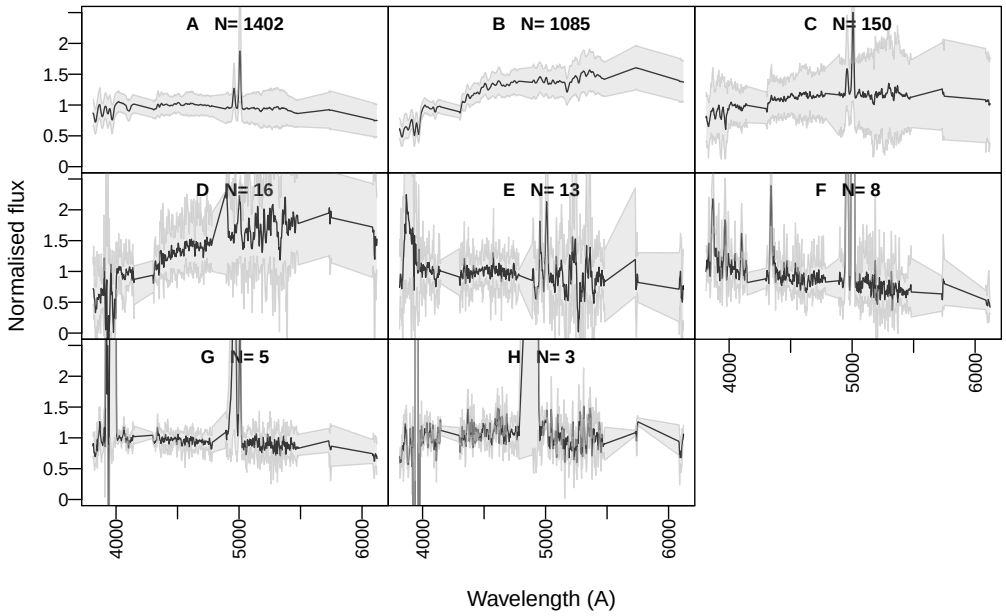
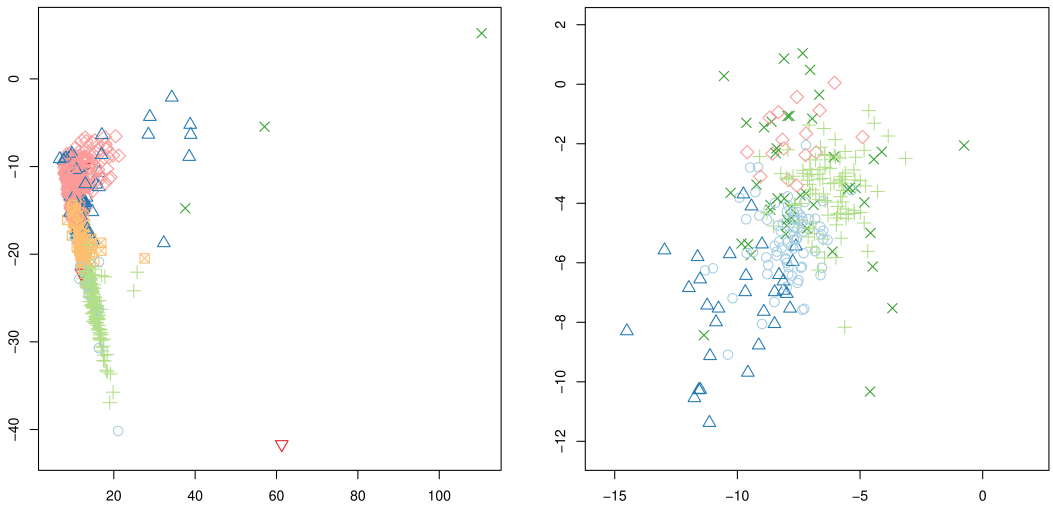


Fig. 4.7.: Representation of the mean spectra (in black) and their dispersion (in grey) for the 8 classes obtained on the $z = 0.45$ bin. Within the gray region lie 90% of the data in the class. In the top region of each panel, the class name is indicated (left) as well as the size of the class (right). This example highlights how most of the spectra are contained within the first 3 classes, while the rest of the classes are outliers.

the noise for each class simply leads to a better fit. However, there is no obvious winner among the six "–Bk" models despite the several computations that I ran, and they seem to all perform equally well. Thus, I decided to choose the less constrained one, i.e. the DkBk model (see Tab 2.1), and all the results presented in this chapter were computed with this DLM.

Unlike the DLM variation, we cannot assume the optimal number of classes K remains the same from one subsample to another. In fact, this parameter is strongly dictated by the spectral features available in the data, and is thus expected to vary based on redshift and the location of the masks. Typically, if an entire emission line is masked within a subsample, we cannot expect to retrieve the same number of classes as for a subsample which does not have this issue. Similarly, $z \sim 0.4$ bins cover a significantly different spectral range than the $z \sim 1.2$ bins, and will thus not necessarily lead to an identical optimal number of classes. For this reason, this parameter is optimised for each bin, and no generalisation is made. The optimisation simply consists in exploring the K-space with the fixed DLM, to find the value that maximises the likelihood criterion, similarly to what is shown in Fig. 4.6. Ideally, we would want a clear optimum to arise, but in practice it is not always the case. Often, the ICL vs K curve hits a plateau after a certain threshold, where the ICL barely changes up until values of K that are too high for the algorithm to find solutions. This is typically the kind of behaviour that was observed on the noiseless simulated catalogue from the previous chapter



(a) Initial classification **(b)** Subclassification

Fig. 4.8.: Projection of the data on the two first dimensions of the Fisher discriminative subspace. The classes are highlighted with data point shapes and colours. The left panel shows the data projection for the initial classification of the $z = 0.44$ bin. The right panel shows the data projection of one of the main classes after being subclassified.

(see Fig. 3.3). In such cases, we thus consider the plateau threshold to be the optimum. On the 26 VIPERS subsamples, I find an optimum number of classes ranging from 5 to 10.

However, these classes are not really satisfactory in the sense that they are not as discriminative as we could expect, as shown by the high dispersion of flux values within the classes (Fig. 4.7). While this phenomenon was not observed with the simulated data in the previous chapter, it is not totally unexpected, since it was also somewhat observed on SDSS spectra in Fraix-Burnet et al. (2021). This is also highlighted by the size of the classes, as shown in the example of Fig. 4.7, where 98% of the galaxies lie in only 3 out of the 8 classes. My interpretation of this behaviour is that it stems from the presence of outliers in the dataset. In fact, the discriminative capability of Fisher-EM lies on its ability to find a subspace that most separates the classes, but looking at the projection of the data from one of the redshift bin sample into the two first dimensions of the discriminative subspace reveals that it discriminates very well some outlier classes, but the vast majority of the data lies in a small region of this 2D-space (Fig. 4.8a). This led me to believe that Fisher-EM was struggling to find a subspace that discriminates both the outliers and the main population, hence resulting in somewhat disparate classes. To fully prove this, it would be necessary to look at all the 2D-spaces to assess the discrimination on all dimensions, but this is not realistically feasible. Nonetheless, we observed that this issue is overcome by running Fisher-EM on each one of the disparate classes individually, thus allowing it to find a proper discriminative subspace for these subsets of data. This results in a much finer end result with less within-class dispersion, and a better coverage of the 2D-space when projecting the data in the first two components of the newly found subspace (Fig. 4.8b).

This amounts to sub-classifying the main classes, similarly to what Fraix-Burnet et al. (2021) did. Although, as I pointed out, some classes gather outliers and thus do not contain many galaxies. These classes, by nature, do not need to be sub-classified. To separate them from the rest of the classes, I used an arbitrary size threshold of 5% of the subsample size. Any initial class with a sample size above that threshold gets sub-classified, while the ones below the threshold are left as is. This way, the final classifications amount to the concatenation of outlier classes and the subclasses. The subclassifications are computed with the same DLM variation as the main classifications, i.e. DkBk. This choice is justified by the fact that the computations are made on a subset of the data, so the conclusions on the performance of the models should also be valid in that case. In the end, the final classifications are made of a total number of classes ranging approximately from 20 to 40 depending on the bin. Note however that there is one bin in particular that led to a much lower number of classes, which I further discuss in the next section.

4.5 Results

After careful preparation of the data and tuning of Fisher-EM, I was finally able to obtain the classification results of the VIPERS spectroscopic dataset, which I present in this section. First, in Sect. 4.5.1 I give an overview of the classification results. However, I show that due to having split the sample into numerous subsamples, the analysis of the results becomes quite difficult. Thus, in Sect. 4.5.2, I present a solution to that problem, and the resulting final form of the VIPERS classification scheme.

4.5.1 Classifications

Since the dataset was split into 26 subsamples, the results consist of 26 individual and independent classifications. The number of classes obtained is not expected to be constant from bin to bin. In fact, it is simply the result of a log-likelihood maximisation process, and therefore fully depends on the specificity of the data sample. Each bin covers its own spectral range, has its own masks and sample size, and therefore has its own optimal number of classes. Here, the number of classes varies from 20 to 43 classes for all but one bin. The only exception is the 4th bin at $z = 0.47$ which resulted in only 8 classes (Tab. 4.2). This is an unfortunate consequence of the location of the masks for this bin, which completely cover up the H_γ and H_β lines, as well as partially the OIII line, hence significantly reducing the discriminative information available within the spectra (Fig. 4.9). As such, the first classification yielded only two classes and essentially just separated outliers from non outliers, with the first class containing 99% of the spectra. The subclassification of the first class yielded 7 classes, which mostly segregate various continuum slopes and a few additional outliers.

Tab. 4.2.: The characteristics of the subsamples: their range of redshift and epoch, their rest-frame spectral range, the masking rate, sample size, and number of classes yielded.

Bin index	1	2	3	4	5	6	7	8	9	10	11	12	13
Lower redshift	0.40	0.42	0.44	0.46	0.48	0.51	0.53	0.55	0.58	0.60	0.63	0.65	0.68
Upper redshift	0.42	0.44	0.46	0.48	0.51	0.53	0.55	0.58	0.60	0.63	0.65	0.68	0.71
Lower Epoch (Gyr)	9.24	9.07	8.91	8.75	8.59	8.43	8.26	8.10	7.94	7.77	7.61	7.44	7.28
Upper Epoch (Gyr)	9.40	9.24	9.07	8.91	8.75	8.59	8.43	8.26	8.10	7.94	7.77	7.61	7.44
Lower wavelength (Å)	3939	3884	3829	3774	3718	3663	3608	3553	3498	3443	3388	3333	3278
Upper wavelength (Å)	6680	6585	6490	6395	6301	6206	6111	6016	5922	5827	5732	5637	5543
Masking rate (%)	42	41	52	42	45	47	43	49	55	44	53	44	45
Sample size	491	1134	1579	2084	2682	3444	3119	3754	3996	4978	5504	4763	4984
Number of classes	20	28	30	8	27	31	33	34	27	29	40	40	40

Bin index	14	15	16	17	18	19	20	21	22	23	24	25	26
Lower redshift	0.71	0.74	0.77	0.80	0.84	0.87	0.91	0.94	0.98	1.02	1.06	1.11	1.15
Upper redshift	0.74	0.77	0.80	0.84	0.87	0.91	0.94	0.98	1.02	1.06	1.11	1.15	1.2
Lower Epoch (Gyr)	7.12	6.95	6.79	6.62	6.46	6.30	6.14	5.97	5.81	5.65	5.49	5.33	5.17
Upper Epoch (Gyr)	7.28	7.12	6.95	6.79	6.62	6.46	6.30	6.14	5.97	5.81	5.65	5.49	5.33
Lower wavelength (Å)	3223	3168	3112	3057	3002	2947	2892	2837	2782	2727	2672	2617	2562
Upper wavelength (Å)	5448	5353	5258	5164	5069	4974	4879	4785	4690	4595	4500	4406	4311
Masking rate (%)	51	51	54	52	52	53	57	47	47	51	48	59	55
Sample size	4277	5258	4466	3947	3546	4088	3217	2346	1775	1343	1233	663	553
Number of classes	35	43	38	28	30	26	26	27	25	27	21	20	21

Looking at Tab. 4.2, we can see that the most defining factor for the number of classes is found to be the sample size, with a Pearson correlation factor of $r = 0.75$. The larger the sample size, the more diversity in the spectra from a statistical standpoint, leading to additional classes. Increasing the sample size may, for instance, lead Fisher-EM to distinguish two classes with only slightly different continuum shapes that could have been gathered in a single class from a physical standpoint. In essence, it can be thought of it this way: if the data is initially modelled by two Gaussian distributions, and a few data points are added in between, they will not be numerous enough to justify the creation of a third Gaussian, and will be assigned to one of the existing two. However, if enough data points are added, at some point it will be statistically better to add a third Gaussian distribution to the model, i.e. increase the number of classes. As opposed to flower species (see Figure. 2.2 of Chapter 2), where the frontiers from one species to another are well-defined, galaxy spectra have more of a continuous distribution, which makes the choice of the number classes less obvious and induce this kind of effect. It is important to understand as well that classification problems rarely have one "true" solution, but rather, multiple statistically good solutions. As such, the fact that the number of classes changes from bin to bin should not be seen as a problem at all, and the main trends and separation patterns remain the same throughout the whole sample. Intuitively, we think of the spectral coverage as another defining factor for the number of classes, as it dictates the characteristic features visible in the spectra. However, because of the spectral overlap of the bins as well and the intrinsic redundancy of information within spectra, it is found not to affect the number of classes significantly ($r = -0.1$). As for the masks, while in some cases it can be problematic, typically when

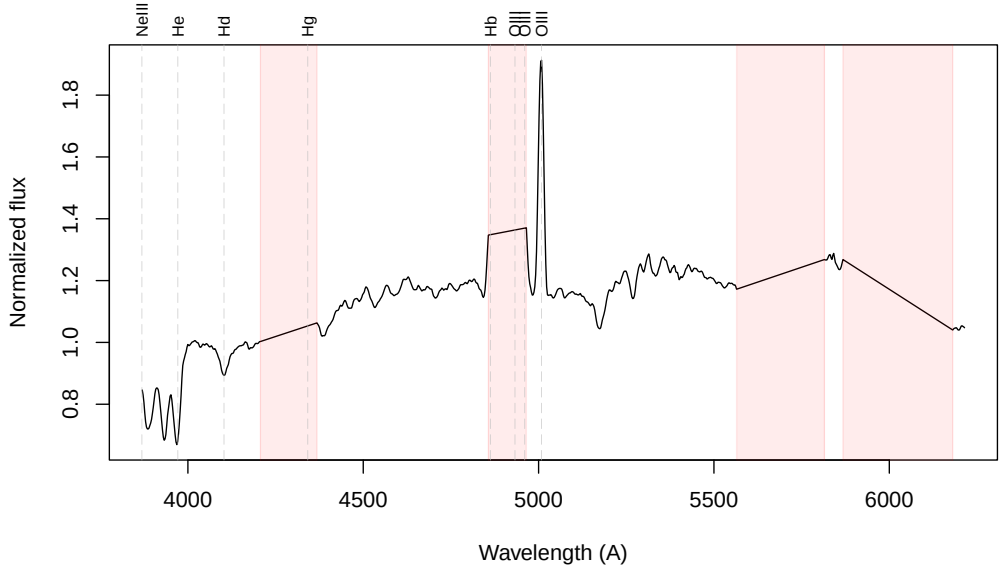


Fig. 4.9.: Stacked spectrum of the 4th bin at $z = 0.47$ (black), which is the only bin that was strongly affected by the masks (indicated in red) due to their unfortunate position that overlaps with multiple significant lines (indicated in dashed lines).

an important region of the spectra is masked such as an entire emission or absorption line, overall there are enough redundancies and correlations between data points to allow the classifications not to be too affected by it ($r = 0.05$) and to be consistent from bin to bin. As noted above, only one bin was significantly affected by the masks.

The resulting classifications can be visualised by stacking the spectra (i.e. computing the mean value of each monochromatic flux) contained within each class, which give us a first insight to the kind of galaxies that each class separates. Additionally, the dispersion within the classes can be visualised by plotting the 10% and 90% quantiles. The first bin is taken as an illustrative example in the corpus of this chapter, but the stacked spectra for all the bins can be found in Appendix D. The first bin yielded 20 classes, and their stacked spectra are shown in Fig. 4.10. Aside from a few outlier classes, the spectra are rather well distributed among the classes, and the within-class dispersions are relatively low in most cases, indicating that the classes do, in fact, efficiently gather similar spectra. In this bin, the H_β and OIII lines are present and play a significant discriminative role in the classification process. On a side note, stacking the spectra greatly improves the S/N, and reveals smaller features like H_δ, NeIII or MgII. Essentially three main types of classes emerge visually: red quiescent galaxies (e.g. class A3), blue galaxies with intense emission lines (e.g. class C3), and galaxies in between the two previous types, with visible but less prominent emission lines and flatter continuum (e.g. class B2). Within each of these three visual categories, further segregation arises, mostly based on continuum slope and line

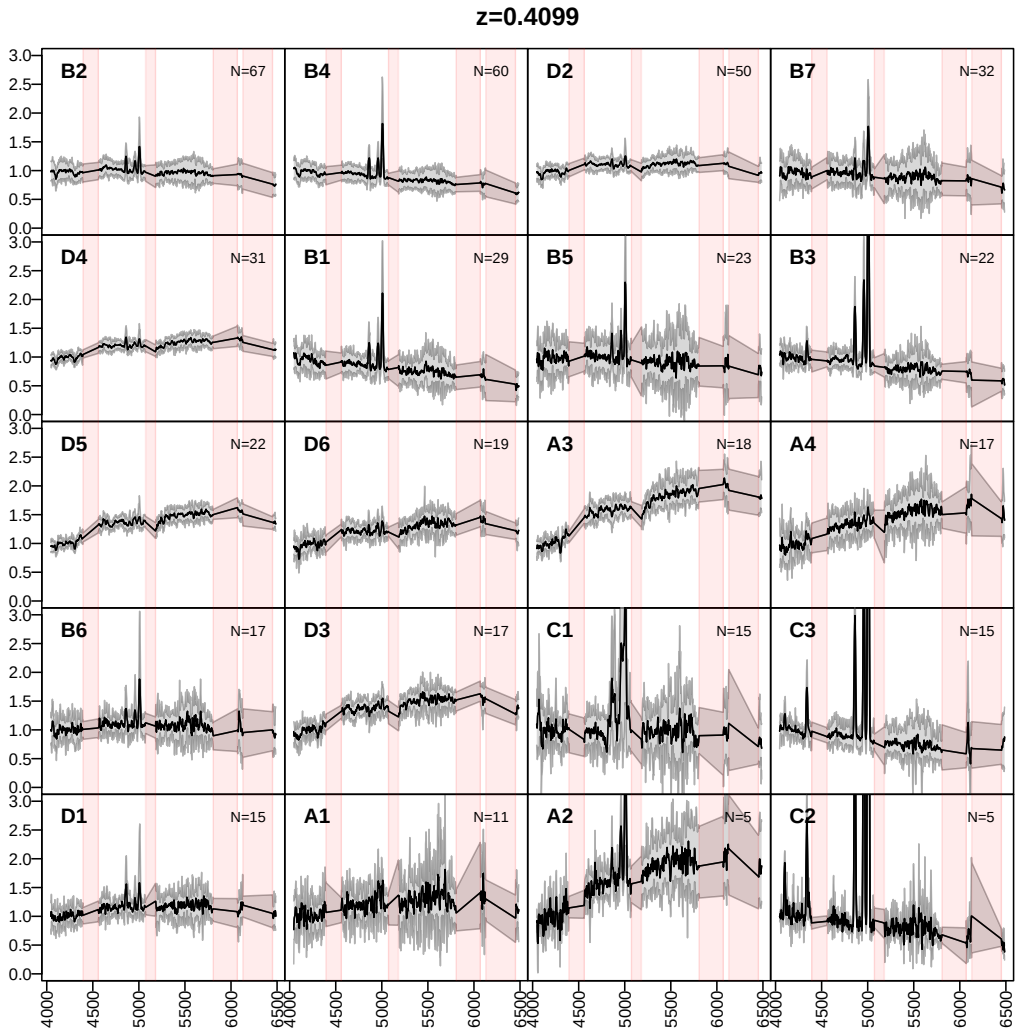


Fig. 4.10.: Stacked spectra (black) of the classes obtained on the first bin, at $z = 0.41$. The dispersion (10% and 90% quantiles) are shown in grey, and the masks in red. Each panel shows one class, whose name is displayed in the top left corner and size in the top right one. The class first letter of the class name corresponds to the main class, and the number to its corresponding subclass.

intensities. We can note, however, that the dispersion in line intensities is generally higher than in the continuum.

This kind of visual inspection, although it is not sufficient to interpret the results, confirms that the classification is successful in the sense that it appears to properly gather galaxies sharing similar spectral features with satisfying dispersion given the S/N of the data. Additionally, it allows to visually pinpoint some of the typical features that stand out and discriminate the classes from one another. The slope of the continuum is well separated, most likely due to how redundant of an information it is, in the sense that it is encompassed in all the data points, as opposed to an emission line whose information is concentrated in a narrow location within the spectra. Nonetheless, thanks to the subclassification procedure, the classes also separate the spectra based on the prominent lines and the Balmer break.

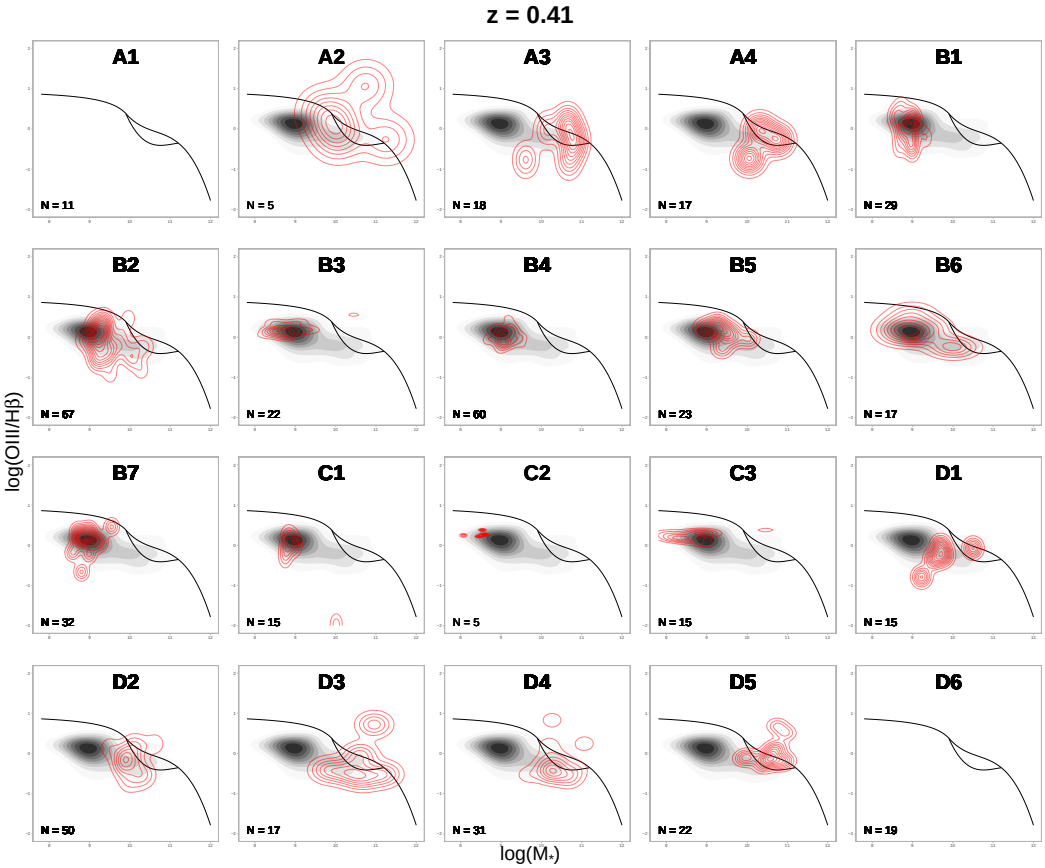


Fig. 4.11.: MEx diagrams of the 20 classes from the first bin, at $z = 0.41$. Each panel shows one class, whose name and sizes are displayed at the top and bottom left corner. The black line delimits the star-forming region (bottom), AGN region (top), and intermediate region (centre). The black contours show the distribution of the bin, and the red contours show the distribution of the class.

The conclusions drawn from the 25 other bins (Appendix D) are essentially identical, but it is important to note that the discriminative features gradually evolve as we progress through the bins due to change in the rest-frame spectral window. To go one step further in understanding the results, we can look at the distribution of various properties within the classes.

Some of these properties, as described in Sect. 4.2.2, are directly derived from the VIPERS spectra, e.g. lines equivalent width, or Balmer break. In that sense, it is understandable that these properties could be well separated in the classes due to the information being straightforwardly available within the spectroscopic data on which the classification is based. Some properties, on the other hand, originate from additional data, different spectral regions, or complex inference processes, and whether these properties could be insightful is thus less intuitive. In any case, studying the properties' distribution within the classes is the next logical step to further making sense of the classifications. Typically, spectral lines in galaxies are often used as diagnostics to identify the source of ionisation, hence separating AGNs from star forming galaxies. The BPT diagram is perhaps the most commonly used

AGN diagnostic tool (Baldwin et al., 1981), but it makes use of the NII and Hα lines, which are unfortunately out of scope within the VIPERS spectra. This is a common issue that affects higher redshift galaxies, and alternative diagnostic tools have therefore been developed. Among them, the Mass-Excitation (MEx; Juneau et al., 2011) diagram, which uses the OIII and H_β lines equivalent width, and the stellar mass. This diagram was usable with the VIPERS data up to $z = 0.9$, after which the OIII line is no longer available. As a result, I was able to plot the classification on the MEx diagram for 19 out of the 26 bins. In Fig. 4.11, I show the resulting diagrams for the first bin at $z = 0.41$. The 18 other bins are shown in Appendix E. The sample is mostly bimodal, and is distributed in the star-forming and intermediate region, with very little AGNs. We can see that most of the classes isolate either galaxies from one region or the other. However, most of the classes in this sample which end up in the intermediate region show no obvious sign of emission on their stacked spectrum. Still, this indicates a successful segregation of the stellar mass despite the fact the spectra were normalised.

Several other diagrams could be investigated (e.g. colour-colour, colour-magnitude) to assess the relevance of the classification from a physical standpoint, but we would remain severely limited by the fact that in total, more than 750 classes for the whole set of redshift bins were yielded. One of the initial motivation of classifying galaxies is to simplify the process of studying their general properties. It does not replace a thorough study of one object, but it brings a general understanding of many galaxies at once. However, so far, with a total of 750 classes, this goal is not exactly fulfilled. The sample was not classified as a whole, but we have instead 26 classifications at successive redshift bins. But it is possible to look at the links between the classes obtained throughout the different bins to 'connect' the 26 classifications; this is what the next section focuses on.

4.5.2 Evolutionary tree

The population of galaxies can be expected to be very similar within two successive bins. In fact, the difference in redshift from one bin to the next is of the order of a few 10^{-2} (i.e. 100-200 Myr), and the sample sizes are large enough, such that we may assume that similar galaxy archetypes could arise in both. The observations were made with the same instrument, the data was processed identically, and both bins cover almost the same spectral window. Thus, it is reasonable to assume that two successive bins should yield similar classifications. Not identical, of course, but we should overall see similar classes emerge. There may be some new classes appearing, some others disappearing, but significant similarities should be present. Under this assumption, it is possible to compare the classes yielded in two successive bins and try to match them with one another based on a given similarity criterion. This way, pairs of similar classes are found, and links between two bins emerge. This process can be repeated one step further down the list of bins again and again, up until we have been through them all, hence creating a chain of links between classes. This approach

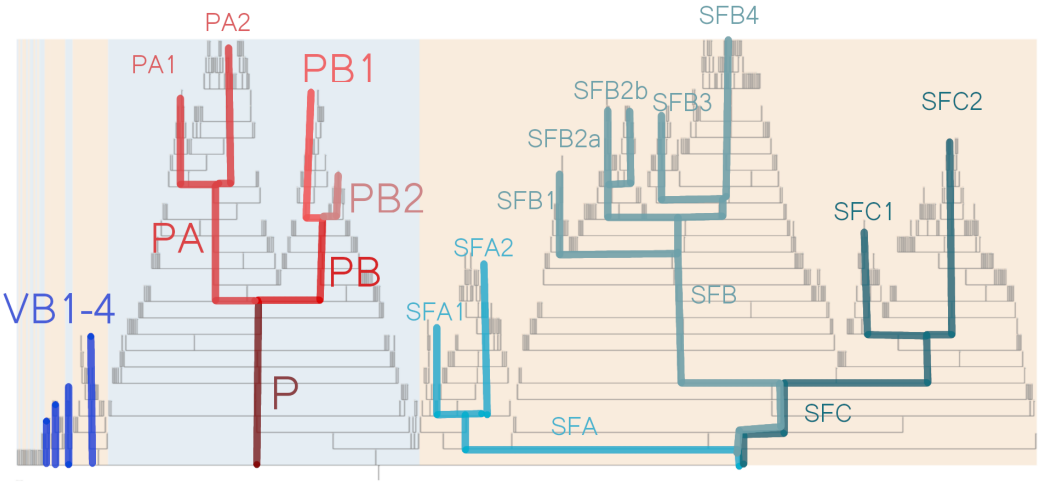


Fig. 4.12.: This tree-like structure highlights the links between the galaxy classes from a redshift of z = 1.2 down to z = 0.4. Each vertical step in the tree corresponds to a certain epoch, linearly sampled from 4 Gyr after the Big Bang (bottom of the tree) to 9 Gyr (top of the tree). Each node represents a class, and similarity links from epoch to epoch were retrieved using k-NN. The black line is the complete classification tree, and it is overlapped with a simplified representation of the three main structures that appear. The nomenclature is explained further down this section. Lastly, the background colours highlight the structures with a common ancestor.

is interesting in two aspects. First, it takes advantage of the expected similarities between classes to simplify the analysis and interpretation. Instead of having to make sense of 750 classes, we can focus on the dozens of chains of similar classes that should arise. Secondly, and perhaps most interestingly, by building these chains of classes starting from the bin of highest redshift, down to the bin of lowest redshift, the chains that emerge can be seen as evolution pathways of the classes starting from ~5 Gyr up to ~9 Gyr after the Big Bang.

The realisation of this idea required the choice of a suitable approach to measure the similarity between two spectra, and of an appropriate automated method to find the links between all the classes. As discussed in Chapter 2, the similarity between two spectra can be computed using a distance metric. There are various paradigms available, but for the sake of keeping this exploratory project simple at first, I opted for the most obvious metric, i.e. the Euclidean distance. As for the matching process, I performed it using the supervised classification algorithm k-Nearest Neighbors (k-NN) that I detail in Sect. 2.3.2. In short, for a given class of bin i, it computes the Euclidean distance between the mean spectrum of that class and all the spectra within the bin $i + 1$. Then, it looks at the classes that the k closest spectra belong to, and outputs the most prevalent one. This process is applied starting from the bins of highest redshift down to the lowest in order to possibly highlight the changes and evolution that galaxies underwent throughout the history of the Universe. This way, the classes of higher redshift are sort of regarded as ancestors from a construction standpoint.

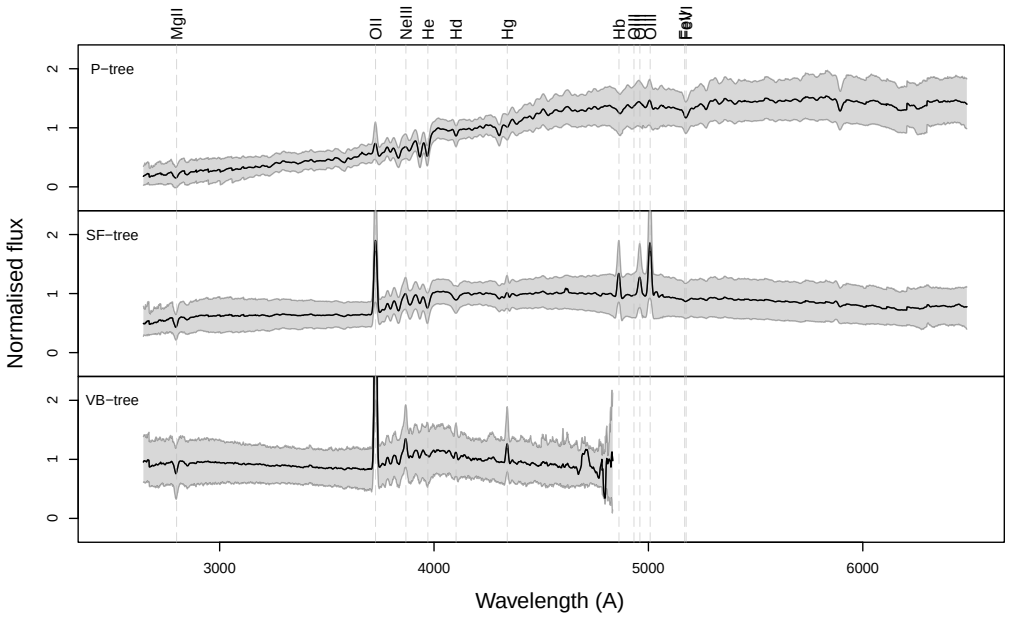


Fig. 4.13.: Stacked spectra (black) of the P-tree (top), SF-tree (middle), and VB-tree (bottom). The dispersion (10% and 90% quantiles) are shown in gray. Some observed emission and absorption lines are highlighted with vertical dashed lines, and the corresponding source is shown at the top. Note that all three panels have the same vertical scale, and that OII line in the bottom panel is cropped out to focus on the continuum and the dispersion. The stacked spectrum of the VB-tree is limited to a narrower spectral range than the two other trees, since it only includes galaxies of redshift $z > 0.9$.

Extending this process to all the classes and all the bins leads to a tree-like structure of links shown in Fig. 4.12.

Despite its complexity, it is clear that the structure can be divided into three independent sub-structures which do not share any common ancestors. At the very right and at the centre, there are two independent large structures that extend from z=1.2 (bottom of the tree) to z=0.4 (top of the tree). On the left, there are a few smaller branches that disappear around z=0.9, which we may consider as the third structure for the sake of simplicity. These three structures will be referred to as the P-tree, SF-tree and VB-tree for reasons explained below. The tree structure is quite complex; it shows multiple branches, which divide into several others at certain epochs, or stop at some others. The bottom of the tree, which, per construction, correspond to the classes of highest redshift, make up the roots, or in other words, the ancestors. As we progress upwards through the branches, we get to classes of lower and lower redshift, down to z=0.4 at the very top. Similarly to how it is possible to identify the type of galaxies gathered in a class by looking at the mean spectrum and dispersion within said class (e.g. Fig. 4.10), the spectra can be stacked along the P, SF and VB trees to highlight their spectral characteristics. Doing so reveals the archetype of the 3 structures (Fig. 4.13); the SF-tree contains galaxies with a rather blue continuum and significant emission lines, indicating these galaxies are actively forming stars. On the other

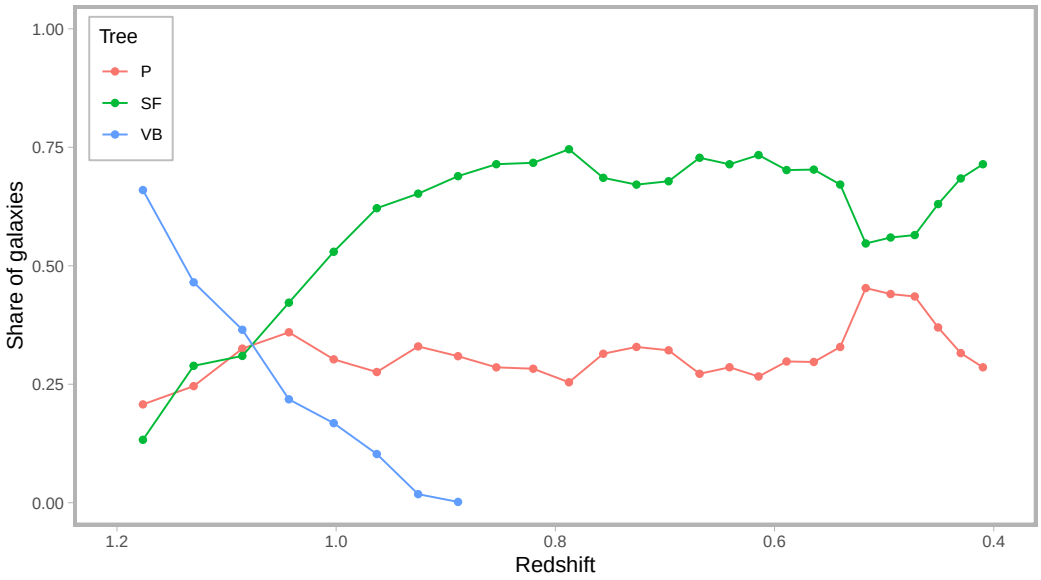


Fig. 4.14.: Share of the VIPERS galaxies among the P, SF and VB trees as a function of redshift. The value is normalised to exclude the few classes that do not belong in either of the three trees in the first two bins. These classes can be seen in the left-most part of Fig. 4.12.

hand, the stacked spectrum of the P-tree is the complete opposite of the previous one. It is essentially flat with no prominent emission lines, and a redder continuum. In this tree, we thus mostly find old quiescent galaxies which are no longer forming stars. Finally, the VB-tree, although it actually consists of 4 independent structures, gather very similar spectra. They are very blue, with extremely intense emission features. As you must have guessed, the names of the structure come from the apparent nature of the galaxies they are made of; Very-Blue (VB) galaxies, Passive (P) galaxies, and Star-Forming (SF) galaxies.

It is possible to investigate the relative population of each tree by summing the number of galaxies contained in the classes they are made of, epoch by epoch (Fig. 4.14). At higher redshifts, the VB tree is the most populated, but its size quickly diminishes down to zero at $z = 0.9$. The SF-tree, on the other hand, sees its population increase at the same pace as that of the VB-tree decreases, suggesting the population of galaxies found in VB at higher redshifts are transferred to the SF-tree at lower redshifts. Finally, the share of P galaxies remain quite constant around 25%.

Within these three sub-trees lie intricate structures from which we can make out a few main branches, as highlighted in Fig. 4.12. Structurally, the VB-tree is the simplest and can be synthesised in 4 branches that will be referred to as VB1-4. Next, the P-tree has more complex and "tree-like" structure. It takes root with a common ancestor at $z = 1.2$ and a single main branch (P) up until $z = 0.76$, where it divides into two branches (PA and PB) which then separate again into two branches each at $z = 0.54$ (PA1 and PA2) and $z = 0.64$ (PB1 and PB2). Lastly, the SF-tree is significantly more complex than the two previous ones. Three main branches (SFA, SFB and SFC) emerge very early on: at z=1.13

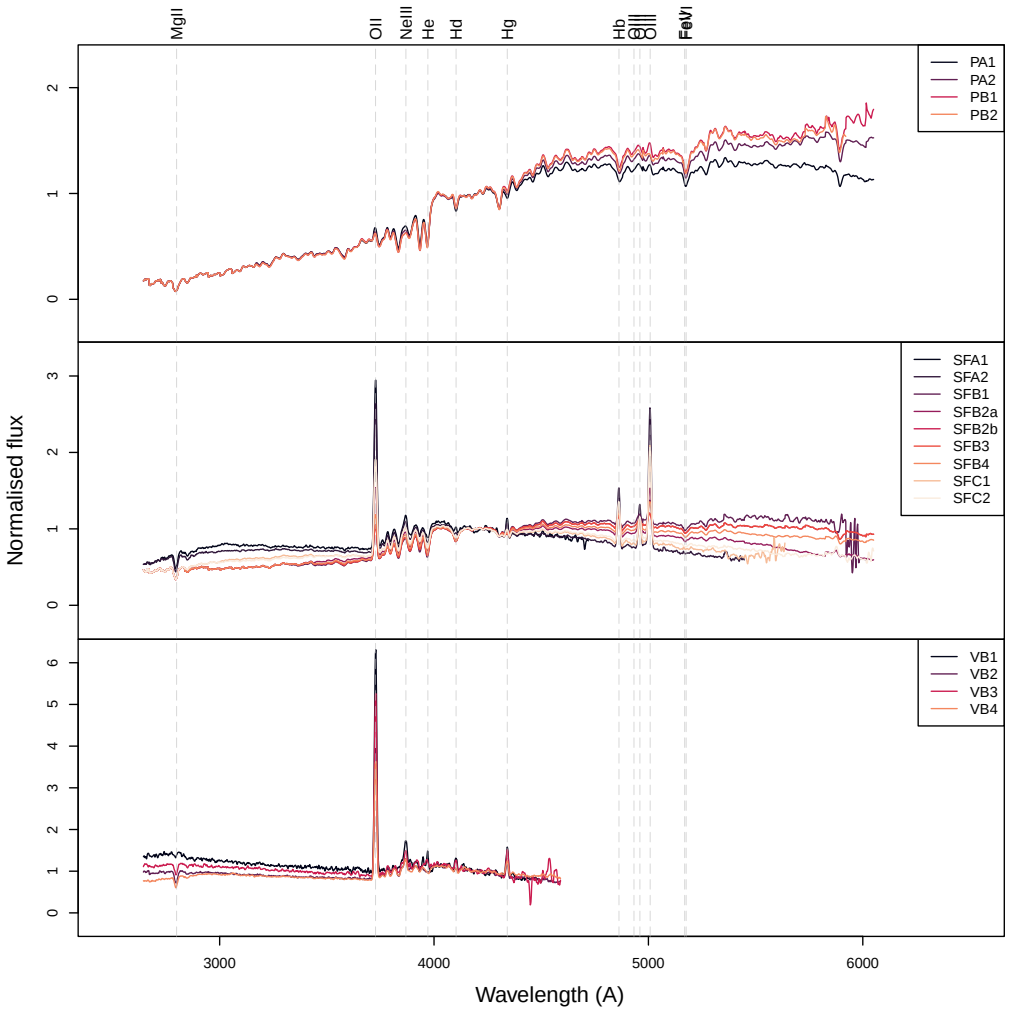


Fig. 4.15.: Stacked spectra of the branches within the P-tree (top), SF-tree (middle), and VB-tree (bottom). The branch nomenclature can be found in Fig. 4.12. The spectra were stacked from the very top of the branch down to the common root, e.g. the stacked spectrum of PA1 includes the PA and P sections. The dispersion is not shown here for the sake of clarity. Some observed emission and absorption lines are highlighted with vertical dashed lines, and the corresponding source is shown at the top. Note that the vertical scale varies from one panel to another to fully include the emission lines within the plot.

for SFA and z=0.93 for SFB and SFC. Similarly to the P-tree, each of these branches then divides into several sub-branches. SFA is the smallest branch, with only 2 sub-branches emerging from it (SFA1 and SFA2), which appear at z=1.04 and disappear respectively at z=0.82 and z=0.70. Similarly, SFC divides into 2 sub-branches (SFC1 and SFC2) at z=0.85, which die out at z=0.64 and z=0.52. Lastly, the SFB branch divides into a larger number of sub-branches (SFB1, SFB2a, SFB2b, SFB3 and SFB4) at redshifts ranging from z=0.61 to z=0.70. Once again, we may compute the stacked spectra of the branches to learn about their spectral characteristics and differences (Fig. 4.15). Doing so confirms that the branches bring a finer degree of spectral discrimination and that each branch appears to have its own specificities.

All in all, this shows that the construction of the evolutionary tree succeeded in synthesising the complex results in a form that is more intelligible. The initial 750 classes were reduced to a total of 3 trees and 14 branches, hence making the classification results more prone to analysis and interpretation in an evolutionary perspective, to which the next section is dedicated.

4.6 Discussion

With the final VIPERS classification scheme in hand, it is now possible to draw some astrophysical interpretations and conclusions. First, in Sect. 4.6.1, I provide further insights on the branches thanks to the physical parameters, and shed some light on the astrophysical meanings of the branches. Then, in Sect. 4.6.2, I discuss some elements of comparison between the VIPERS classification and the SDSS classification of local galaxies from Fraix-Burnet et al. (2021). Finally, in Sect. 4.6.3, I discuss the limitations and aspects of this work that could be improved.

4.6.1 Interpretation of the branches

I highlighted in the previous section that three categories of classes emerged from the k-NN analysis: (i) the P-tree, (ii) the SF-tree, and (iii) the VB-tree. Initially, their interpretation was limited by the large number of classes yielded, but the evolutionary tree solves that issue. In fact, it is now possible to isolate branches (i.e. chains of similar classes), and investigate the evolution of their properties to make sense of the kind of galaxies that make them up; this is what I present in this section. First, in Sect. 4.6.1.1, I focus on the P-tree and its branches. Then, I discuss the SF-tree in Sect. 4.6.1.2. And lastly, Sect. 4.6.1.3 is dedicated to the VB-tree.

4.6.1.1 Passive branches

Based on their stacked spectra, the galaxies gathered in the P-tree are characterised by their red continuum, with a strong D4000 break and mostly no emission lines. These are indicative of an older stellar population, and mostly no ongoing star-formation. Still, the evolutionary tree divides this category into 4 branches (Fig. 4.12). Visually, from their stacked spectra, the 4 branches seem to mostly split the P-tree population by different levels of redness (Fig. 4.15), which may be explained by varying degrees of dust attenuation, but further insights are given by looking at the evolution of the classes' properties along the branches.

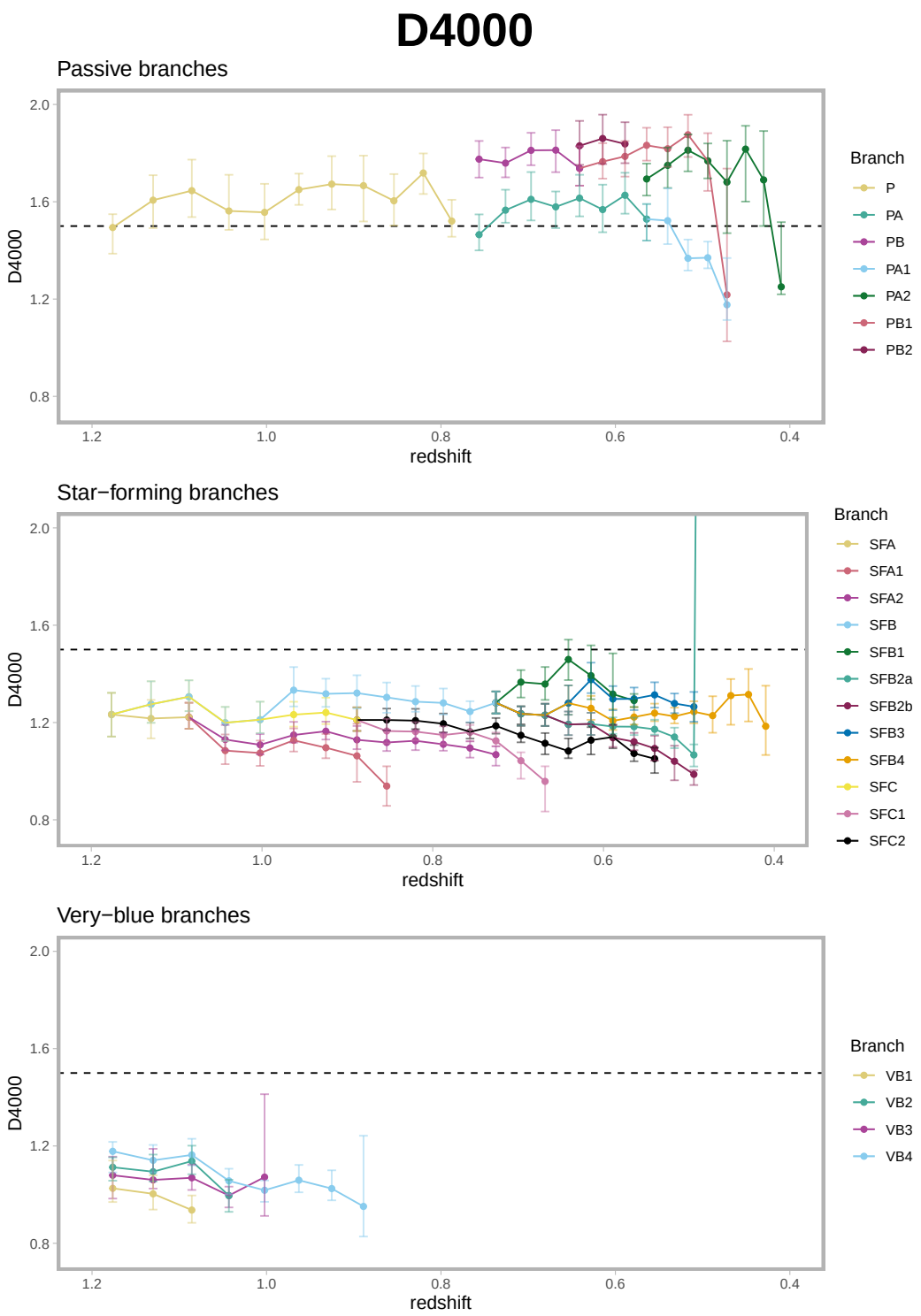


Fig. 4.16.: Balmer break (D4000) in the P-tree (upper panel), SF-tree (middle panel) and VB-tree (bottom panel) branches. The dots show the median value, and the error bars show the 25% and 75% quantiles within the class. The branches, as defined in Fig. 4.12, are indicated by the colour. The dotted line shows the separation threshold between red passive galaxies (above) and blue star-forming galaxies (below) as per Kauffmann et al. (2003)

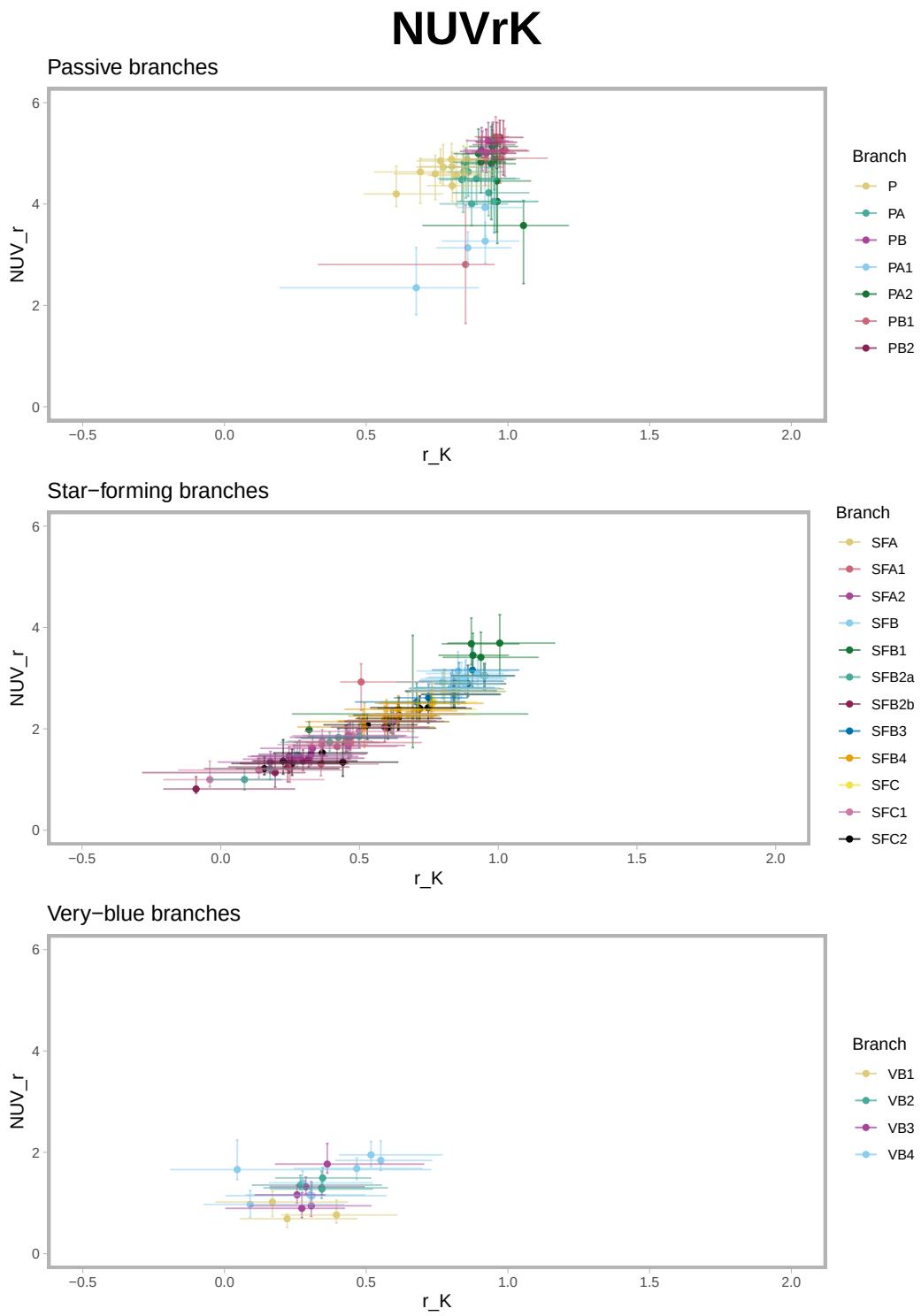


Fig. 4.17.: Position of the classes of the P-tree (upper panel), SF-tree (middle panel) and VB-tree (bottom panel) in the NUVrK diagram. The dots show the median value, and the error bars show the 25% and 75% quantiles within the class. The branches, as defined in Fig. 4.12, are indicated by the colour.

First, the D4000 break informs us about the stellar population within the galaxies. Kauffmann et al. (2003) show that there is a bimodal distribution of D4000 within the SDSS spectra, which separates galaxies with old stellar populations (D4000>1.5) from those with more

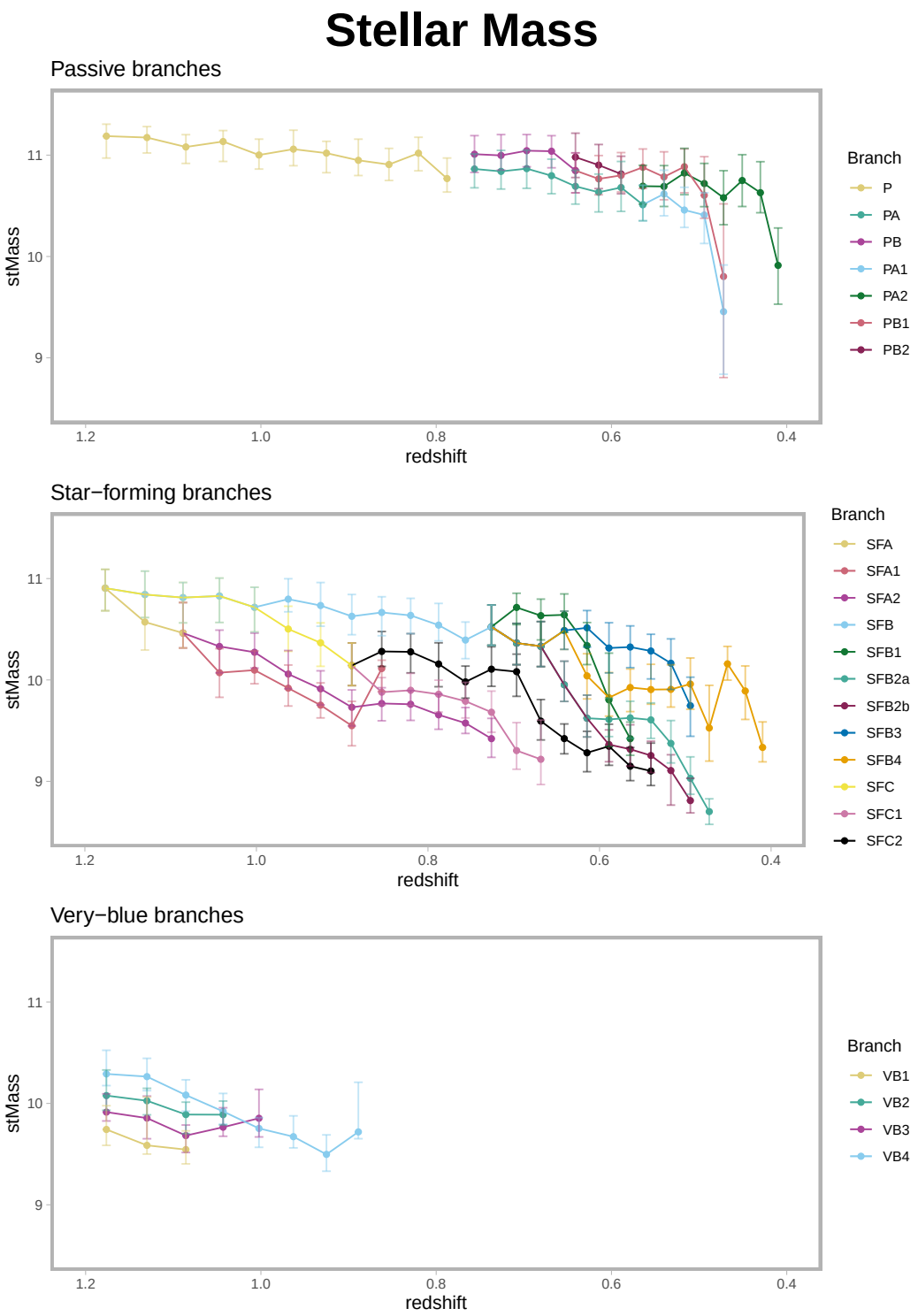


Fig. 4.18.: Log of the stellar mass (in solar masses) of the classes of the P-tree (upper panel), SF-tree (middle panel) and VB-tree (bottom panel) branches. The dots show the median value, and the error bars show the 25% and 75% quantiles within the class. The branches, as defined in Fig. 4.12, are indicated by the colour.

recent star formation (D4000<1.5). As suggested by the stacked spectra, the vast majority of classes from the P-tree have D4000 values typical of older passive galaxies (Fig. 4.16). There are a few exceptions nonetheless, out of which we can distinguish two categories. On

the one hand, there are two classes at the end of branches PA2 and PB1 which correspond to non-meaningful classes often found at the end of a branch. Such classes typically contain only a few galaxies with a lot of dispersion, and do not really correspond to any physical class. I discuss further down this chapter the possibility of implementing some type of selection to only include meaningful classes in the construction of the tree, but in its current state, all classes are included. And on the other hand, the PA1 branch shows a different behaviour, with a consistent decrease in D4000 over several successive epochs, suggesting it does carry a physical meaning.

The fork from P to PA and PB clearly separates two populations of galaxies with distinct D4000 values; in PA, we find galaxies with D4000 mostly ranging from 1.5 to 1.7, while those with a 4000 A break above 1.7 are found in PB. Interestingly, Haines et al. (2017) observes that many massive blue galaxies within the VIPERS dataset are being quenched around $z \approx 0.7$, which is possibly what is seen in the PA branch. Additionally, further down the PA branch, two paths emerge, one of which (PA1) shows features in more recent epochs that suggest signs of recently finished events of star formation. In fact, their D4000 drops significantly, they appear to have a slightly bluer continuum, and noticeable OII emission and Hδ absorption. This branch shows characteristics of post-starburst galaxies, and could represent an intermediate link. The other sub-branch, PA2, does not share these features and instead reaches D4000 levels comparable to the PB branch, as well as an overall redder spectrum. PA2, PB1 and PB2 are very similar, and mostly segregate different levels of redness, with PA1 being the less red, and PB2 the reddest. The NUVrK colour diagram, which provides information relative to the star formation (see Fig. 4.17 of Chapter 1), is in agreement with this interpretation: (i) the classes of the P-tree are located in the passive region of the diagram and (ii) the PA1 branch ends up in the intermediate zone. We also observe that the galaxies in the PA1 branch tend to be somewhat less massive than in the other branches (Fig. 4.18).

4.6.1.2 Star-forming branches

The stacked spectra of the galaxies found in the SF-tree show more complex spectral features than in the P-tree, typically found in star-forming galaxies. In addition to a much bluer continuum and a small D4000 break, we observe several significant emission lines ([OII], Hg, Hbeta, [OIII]) in all the SF-tree, with varying intensities and ratios across the branches. The SFA branches are the ones with the most intense emission lines and the bluest continuum, and both SFA1 and SFA2 have very similar features. Nonetheless, SFA2 contains slightly bluer galaxies with more intense [OII] emission than SFA1. In comparison, the galaxies in the SFC branches have less strong emission features, a slightly redder continuum, and also show strong Hδ absorption. We observe the same characteristics in SFBs but to a greater extreme; the emission lines get much fainter, and the spectra redder.

The classes of the SF-tree are, as opposed to the P-tree, characterised by a D4000 systematically smaller than 1.5 (Fig. 4.16), typical of galaxies with recent star formation events. There is nonetheless a significant amount of diversity within that tree. While the uncertainty on the estimations of star formation rate are too high to observe any significant differences between the branches, the NUVrK diagram is quite informative in that regard: the SFA and SFC branches are associated to the higher end of the star formation rate, whereas the PB branches are more so located in the intermediate region. In particular, the spectra of galaxies found in SFB1 and SFB3 resemble that of the PA1 branch, which was interpreted as a class of currently quenching galaxies.

Interestingly, we observe that overall, SF galaxies see their stellar mass decrease at smaller redshifts (Fig. 4.18). Additionally, the branches isolate distinct ranges of stellar masses, with a clear anti-correlation between star-formation rate and stellar mass. The SFA branches, which are the most star-forming, are also the less massive, while the SFB which are the least star-forming, are the most massive. This observation can likely be attributed to 'downsizing' (Cowie et al., 1996). Generally, SF galaxies are also significantly less massive than the P galaxies.

4.6.1.3 Very-blue branches

Structurally, the VB branches are different from Ps and SFs as they do not originate from the same ancestor. They are actually 4 independent branches that span through several redshifts down to $z = 0.9$, and while smaller than the P and SF trees, the VB branches still gather a significant amount of galaxies, especially at higher redshifts. It is to be noted that despite not sharing a common ancestor, the four VB branches show very similar characteristics. Spectra-wise, the four branches separate different levels of blueness of the continuum (VB1 being the bluest, and VB4 the less blue) and line intensity (in the same order).

At first glance, these branches can be seen as 'extreme' versions of the SF-tree. They are characterised by a particularly blue continuum, low D4000 values, and very intense emission lines. They are all located on the region of the NUVrK associated to the highest star formation rate. However, we note a significant NeIII emission, which is usually associated to nuclei activity. It is difficult to draw clear conclusions on that aspect, since Mg and NeV emissions would also be expected in stronger intensity than NeIII, but is not really observed here.

The trend between stellar mass and star formation rate observed for the P and SF galaxies also holds for the VB galaxies: they are the most star-forming, and also the less massive at a given redshift (Fig. 4.18). Contrary to the SF-tree, however, it is difficult to really see a trend with redshift, since the VB tree branches die out at $z \approx 0.9$.

On a last note, the fact that this tree disappears at $z = 0.9$ is more likely to be a consequence of the representativeness of the sample rather than to have an actual evolutive meaning. It is possible that there is somewhat of a selection bias which limits the amount of VB galaxies in the sample at lower redshift.

4.6.2 Comparison with local galaxies

Chronologically, I started this part of my work soon after the publication of Fraix-Burnet et al. (2021), in which they build a spectroscopic classification of galaxies from the local Universe using the SDSS survey. The initial motivation of my project was henceforth to extend this study to higher redshift. Comparing my results with those obtained in their paper is thus something I had in mind all along.

In their work, due to their sample being limited to low redshift galaxies ($z < 0.25$), there was no reason to apply the same kind of manipulations than those I opted for on the VIPERS sample. As a result, they were able to classify their sample as a whole, and ended up with a single classification. Whereas, I, on the other hand, ended up with 26 separate classifications, that I agglomerated in a structure that does not make the comparison very straightforward. There are many ideas to be explored in that regard, but as all theses go, I was limited in time and so far have only proceeded to a simple, but encouraging, comparison.

Similarly to how I decided to classify the VIPERS subsamples in a 2-step approach to improve the quality of the discriminant subspace, they initially classified the SDSS sample into 4 main classes, which were then subclassified, hence resulting in their final classification. These four main classes essentially consist of: (A) a class of red quiescent galaxies, (B) a class of galaxies with bluer continuum and some emission lines, (C) a diverse class with a lot of dispersion both in continuum shapes and line intensities, and (D) a small class of galaxies with a particularly blue continuum and very intense emission lines. These main classes are in a way reminiscent of the P, SF and VB categorisations that emerged from the construction of the evolutionary tree. Thus, I decided to compare the stacked spectra of the SDSS main classes with that of the P, SF and VB trees (Fig. 4.19). After renormalising them on the same spectral region, it is clear that the P-tree and SF-tree match remarkably well with the SDSS-A and SDSS-B classes respectively, both from a continuum a line intensity standpoint. The continuum slope of the VB-tree also matches perfectly with that of the SDSS-D class, although the latter has a lot more visible lines. Overall, this suggests that we are witnessing the same main categories of galaxy emerge from the analysis both in the z~1 vicinity and in the local Universe. Nonetheless, there is no equivalent of class SDSS-C, but it is difficult to draw any conclusions out of that due to how diverse this class is, in the sense that it spans a wide range of continuum shapes. Still, it seems to be characterised by noticeable line broadening, which we do not retrieve within the VIPERS classification.

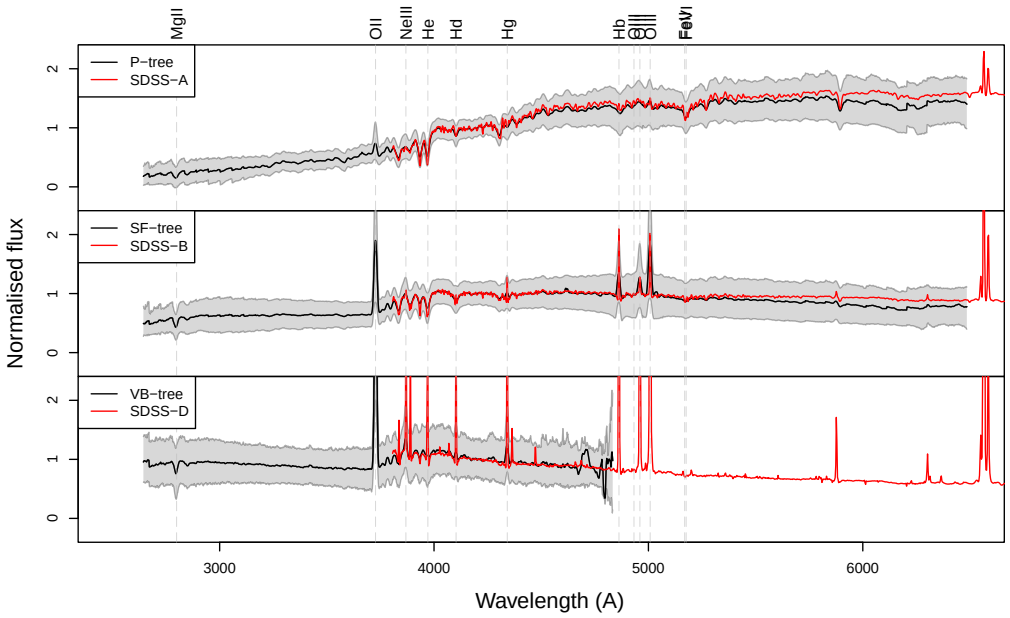


Fig. 4.19.: Same figure as in Fig. 4.13 with an additional layer showing the stacked spectra of the SDSS main classes A, B and D. The SDSS spectra were renormalised in the same spectral region than the VIPERS spectra to make the comparison possible.

This simple comparison could be improved by lowering the resolution of the SDSS stacked spectra down to the VIPERS resolution (R= 220). This can easily be done convoluting the spectra with a Gaussian function of appropriate width, and would likely slightly affect the emission and absorption features. Additionally, the SDSS spectra appear to be generally redder, which could perhaps be explained by some kind of systematic error linked to flux calibration. Still, the comparison portrayed here is good enough to conclude of a remarkable match between the SDSS main classes and the VIPERS sub-trees.

Now, of course, the next step would be to go one step further and compare the SDSS subclasses with the VIPERS main branches. This could be done, for instance, by extending the k-NN procedure from the end of the branches down to the SDSS subclasses. This is something I have not had the time to explore, but is high on the priority list of the things that would be worth doing in the near future. We have to note, nonetheless, that the SDSS spectra generally have a better S/N, and a better resolution as well. This means that the spectra typically contain more information and finer details than those observed in the VIPERS survey. As a result, we can also expect the SDSS classification to be more complex and diverse, which may translate in some difficulties to match them with the VIPERS branches. This did not affect the comparison with the main classes because they are quite broad, but it may be more significant with the subclasses.

4.6.3 Limitations

This work was built on solid foundations of published work both in the field of statistics and astrophysics. In that sense, there is not much room for improvement in the initial aspects of this project; the selection and preparation of the data were carefully led, the clustering algorithm was thoughtfully chosen and scientifically backed up, and its application and tuning properly managed. However, due to unavoidable constraints and so far unexplored practical issues, I was led to conduct a significantly more exploratory kind of work with the bin-by-bin classification, and the construction of the tree. As a result, on several occasions, I had to make choices to keep things simple and manageable, which could possibly be improved on various aspects. In this section, I discuss these matters and propose some ideas and possibilities to improve this work.

First, let us talk about the subclassification procedure. Although this is common practice to improve the spectral discrimination, it unavoidably begs the question: is it sufficient to sub-classify once? It would, in fact, absolutely be feasible to keep going and sub-classify the subclasses once, twice, and so on. And it is actually difficult to objectively decide what is best to do. Realistically, there is probably no objective answer to that question, and it should perhaps be seen instead as strongly dependent on the needs and motivations that led to classifying the sample in the first place. Here, from a purely statistical standpoint, the main classification (i.e. prior to subclassification) was a completely valid classification, but it was deemed insufficient from a physical standpoint due to the relatively high dispersion of the spectral features within the classes. I found that sub-classifying once led to "good enough" dispersion, which is somewhat subjective. In this specific case, I also happened to be limited by the low sample size, which sealed the deal and convinced me not to go any further with the subclassification procedure. However, in a hypothetical world where the VIPERS sample was larger, the decision would not have been so easy. In practice, it should also be motivated by how the spectral features and the level of detail that are observed in the spectra are well reflected and separated within the classes. It should also be noted that the classification is obtained using the discriminant features of the spectra; i.e. linear combinations of monochromatic fluxes resulting from the Fisher analysis, which can only be so discriminant, in the sense that the spectra are characterised by a mostly continuous distribution of their features. For instance, although we can make the distinction between red and blue galaxies, in practice there is a gradient of continuum slopes between the two. As such, we can settle for a simple bimodal separation, or decide to go for finer slices along the gradient. Schematically, this is pretty much the kind of dilemma that the subclassification amounts to. In this work, I made the choice to go for a rather fine separation with a one-step subclassification. But it would have been technically valid to settle for a less discriminative outcome with only the main classification, or instead, should the sample size allow it, to go for an even finer classification with an additional layer of subclassification. In the end, it all boils down to the nature of unsupervised classification; since no prior constraints are set on

the expected classes, the concept of "right" or "wrong" classification does not really make sense. A classification is a classification. It might be better or worse than another with regard to a specific goal, but that is for us humans to evaluate. Of course, the process is still largely data-driven, and the classifications remain optimised on the basis of a statistical criterion, but as with most artificial intelligence algorithms, it is only "intelligent" to a certain extent, and human supervision remains necessary in some aspects.

Next, an important aspect that I mentioned earlier is the distance metric. Here, I used Euclidean distance; it is simple, it gives a valid distance measurement between spectra, and was hence deemed sufficient for this work. However, two significant limitations arise. First, the Euclidean distance will favour the continuum over narrow features, simply because the latter are, by definition, concentrated on a few monochromatic fluxes, while the continuum spans the entire spectral range. This means that a slight variation on the continuum will have more impact on the distance metric than a strong variation on the continuum. Therefore, the tree favours linking two galaxies of very similar continuum but major differences in narrow features, rather than the other way around. Of course, continuum and narrow features are intrinsically correlated, for instance a blue galaxy is much more likely to have strong emission features than a red galaxy, but it might affect the resulting tree nonetheless. Additionally, in the presented method, the distances are computed on the full spectra, which contain several hundreds of monochromatic fluxes despite the fact that distance metrics suffer significantly from the curse of dimensionality (see Chapter 2, Sect 2.2.3). In other words, the higher the dimension, the poorer their performance. As such, it would be beneficial to apply some kind of dimension reduction before the distance computation to improve the results. Additionally, this would likely help balance out the dissymmetry between the continuum and narrow features in the distance computation. The most logical approach for this that comes to mind is simply to re-use the Fisher-EM discriminant subspace as a mean for dimension reduction. The projection matrix can easily be retrieved and used to project the spectra onto the subspace, thus greatly reducing their dimension while selecting discriminant features in the process. The issue lies in the fact that the projection matrix is specific to a classification, meaning that each classification and each subclassification has its own independent projection matrix. Which naturally begs the question: which projection matrix should be used? This idea definitely seems worthy of being explored, but the correct strategy is not so obvious, and it would require some work to conclude.

In the end, the distance is used as a metric to build the tree, but the underlying method is probably where most improvements could be made; what I presented in this manuscript is essentially the simplest version of what it could be. The k-NN algorithm seems well suited for the task, although other alternative could certainly be used. But the way the tree was built does not allow complex structures to emerge, such as branch merging, or links that skip an epoch. Making this possible, in addition to changes in the distance computation, would likely help reduce the number of dead branches, which are very numerous in the current version of the tree. The goal is not to remove all the dead branches, of course, as

some of them have an actual physical meaning, being that this type of galaxy is no longer observed at lower redshifts. But so far, many of the dead branches are, in fact, dead, as a mere consequence of the construction method rather than as a reflection of an underlying change of galaxy population. Additionally, I showed that Fisher-EM often yields small classes of outliers spectra, which are not always physical. It could be data with calibration issues, or sometimes just very noisy spectra. Although nothing of the sort was done, it could be beneficial to select the classes that we deem as relevant before proceeding to building the tree. In fact, many ideas could be explored to improve the tree. In its current state, it does not really reflect evolution paths of galaxies the way a phylogenetic tree would. Instead, it should be seen as a tool for linking together classifications at different redshifts. The links produced highlight how the classes change with look-back time, which may or may not reflect actual physical evolution pathways. From the results I obtained in this work, this method seems promising, since, despite its simplistic construction, it was nonetheless able to retrieve interesting physical changes at certain epochs. Still, in its current state, these meaningful evolutionary paths are intertwined with non-physical features and paths, but this would likely be greatly improved by implementing the ideas mentioned in this section.

On a final note, it is worth reminding that the quality of both the classifications and the tree are conditioned by the quality of the spectra. This includes the resolution of the instrument, the amount of noise, whether the observations are taken from ground-based observatories or with space telescopes, and so on. The VIPERS spectra used in this work were taken with a spectral resolution of $R \approx 220$, which is ten times lower than the resolution of the SDSS spectra. This work still led to a successful and complex multi-redshift spectral classification, which certainly demonstrates the strength of the method, but also shows its potential to provide even more refined results with higher quality spectra in the future.

4.7 Conclusion

In this work, I successfully used the latent subspace clustering algorithm Fisher-EM on a large sample of optical spectra of galaxies with redshift of $0.4 < z < 1.2$ from the VIPERS DR2. The selected sample contains 79 224 spectra, and was split into 26 subsets by bins of redshift. A classification was obtained on each individual subset, following a two-step classification procedure. First, main classes were yielded, and each of these classes were then sub-classified to reach a more refined end result. Similarity links between the classes of the 26 subsets were found using the k-Nearest Neighbour algorithm, thus resulting in a tree-like structure that highlights evolution paths of the classes.

The result constitutes the first proper automated spectral classification of galaxies in the $z \sim 1$ vicinity. Although the combination of sky and instrumental residuals with large redshift span added significant constraints on the analysis, they were overcome thanks to

adaptive masks, and both supervised and unsupervised classification methods. As a result, the spectral diversity of galaxies of redshift $0.4 < z < 1.2$ was successfully mapped with a fully data-driven approach.

My prior work on simulated spectra showed that automated spectral classifications from Fisher-EM were discriminant with respect to the physical characteristics. The same conclusion is drawn from the study of the VIPERS DR2 spectra, where the classes were found to segregate properties such as Balmer break, colour, star formation rate, and stellar mass. Three main categories of classes were highlighted by the evolutionary tree: (i) red passive galaxies, (ii) blue star forming galaxies and (iii) very blue galaxies with intense emission features. The stacked spectra of these three types match remarkably well what that of the SDSS main classes A, B and D from Fraix-Burnet et al. (2021). Furthermore, additional distinctions among these three categories are made through the branches; notably, different intensities of star-formation are distinguished within the SF-tree, and it appears that a portion of the P-tree and the SF-tree are isolating. Lastly, while the VB-tree is indubitably associated to very high star formation rates, some spectral features could indicate the potential presence of AGNs within the classes, but it remains inconclusive due to the absence of Mg and NeV emission lines.

This novel approach to galaxy classification can easily be extended to higher redshifts with future surveys. Most notably, it will be possible to utilize the JWST spectroscopic observations of very high redshift galaxies, hence contributing to shedding light on the question that are left surrounding the intricate mechanisms governing galaxy formation and evolution.

5 Conclusion and perspectives

Spectroscopic classification of galaxies appears as a powerful but challenging alternative to morphological and parametric classifications. However, this field is still in its early stage of development, and as such, this Ph.D. thesis was anchored in quite a unique context on a problem with many yet unexplored facets.

The high dimensionality of spectra and the ever-growing quantity of data makes the use of an automated method unavoidable; my work has demonstrated Fisher-EM to be a suitable solution. Its unsupervised nature is key: intrinsically, supervised methods, despite their popularity, cannot do more than mimic what they are trained on. This implies the use of an expected classification as a mean for training, and, by lack of anything else, there would not be much of a choice but to use morphological or (multi)-parametric classifications. In the end, this would amount to restricting the new classifier to the limits of the current classifications. Unsupervised techniques, on the other hand, are free from such constraints. But, as a drawback, it also means that no physical prior knowledge is injected in the model; in theory, there is thus no guarantee that the resulting classifications would hold physical meaning. This is why I devoted time during my thesis to investigate this, and the results are clear: the classes obtained do gather galaxies with common physical characteristics. In particular, they often isolate relatively narrow locations of 2D-parametric spaces commonly used for classification (e.g. colour-colour, colour-magnitude, BPT, etc.), indicating our method surpasses the performance of individual parametric methods.

The most significant contribution of this thesis lies in the classification of the galaxy spectra from the VIPERS survey. Classifying galaxies via their spectra has only been recently made possible by modern surveys and statistical tools, and as a result, this work effectively constitutes one of the first few contributions to the matter. Spectra of local galaxies ($z < 0.25$) from the SDSS were previously classified with the same method in Fraix-Burnet et al. (2021), but the VIPERS dataset came with certain specificities that made its analysis significantly more complex. Most importantly, the VIPERS survey probed galaxies of intermediate redshift ranging from 0.4 to 1.2. This, in addition to the fact that the data was affected by various residuals at certain wavelengths, made it necessary to (i) split the sample into 26 subsamples and (ii) mask out part of the spectra with significant residuals. Each of the subsamples were classified individually, and the classes were shown to hold physical meaning despite having 40-55% of the monochromatic fluxes masked. This is explained by (i) the fact that there is a lot a redundancy of information from one wavelength to another, since the monochromatic fluxes are correlated with one another; and (ii) the capacity of

Fisher-EM to retrieve the relevant discriminant information within the spectra with its latent subspace.

A method was devised to synthesise the 26 classifications, and link the classes by similarity at different epochs in a tree-like structure. Through its branches, the evolution of the classes as a function of look-back time can be investigated. It revealed three main categories, which match remarkably well three of the four main categories of the SDSS classification. They separate red passive galaxies, moderately star-forming galaxies and very blue, highly star-forming galaxies. The classification is not limited to these main categories, as further distinctions are made in the different tree branches, providing finer physical separation. While the ambitious end goal would be to obtain an evolutive tree comparable to phylogenetic trees in its meaning, it is important to remain cautious with the current results. In fact, the evolutive interpretation is still largely limited by the simplistic construction of the tree and the incompleteness of the sample. Despite these current limitations, this approach shows promising perspectives.

Due to the exploratory nature of this work, there is still room for improvement on various aspects. As I suggested, the construction method of the tree is perhaps where most subsequent work could be made. I discuss this in more details in the previous chapter (Chapter 4, Sect. 4.6.3), but in short: (i) another distance metric could be used in an effort to balance out the weight of the narrow features versus the continuum; (ii) instead of basing the similarity measurement on the entire spectra, it could be worth exploring the possibility of using the features of the latent subspace, which should encompass most of the discriminant information by construction; (iii) it would be beneficial to remove non-physical and small classes from the tree to avoid unnecessary dead branches; and (iv) it could be interesting to allow more complex structures (e.g. branch merging) to emerge.

The classification algorithm we used has certainly proved its worth. Nonetheless, since the quality of the classification is a direct consequence of the performance of the algorithm, we should remain open to new possibilities, should they prove to be better. Notably, a Bayesian adaptation of Fisher-EM has been developed by Jouvin et al. (2020), which seems to have slightly better performances than the regular version of the algorithm. It could be worth exploring its potential in this specific astrophysical context. Alternative methods could be considered as well, but it should be noted that two characteristics are crucial for spectra classification. First, the method has to be unsupervised for reasons mentioned above; second, it must be designed to perform well with high-dimensional data. This essentially means that a dimension reduction process has to be included to avoid the constraints of the curse of dimensionality.

On a final note, I wish to highlight that the major prospects of this work evidently ensue from the use of additional data. In the foreseeable future, it will be possible to extend the work conducted in this thesis to even higher redshifts with the JWST and other forthcoming

multi-object spectrographs (e.g. Prime Focus Spectrograph at Subaru; Multi-Object Optical and Near-IR Spectrograph at VLT) hence mapping the spectral diversity of galaxies down to the very early Universe. Numerous surveys will be or are currently being conducted (e.g. CEERS, NG-DEEP, COSMOS-Web), and will lead to an unprecedented number of observations at such redshifts. Using appropriate tools to make the most out of all this data is crucial, and in that sense, the work presented in this thesis could be of significant use. In addition to observations at higher redshifts, it would be insightful to include observations at wavelengths other than optical to probe different kinds of physical processes. This essentially amounts to injecting more physical information in the data, in order to increase the relevance of the classes. In theory, this could also be achieved by combining different types of data (e.g. photometric observables, morphological characteristics) to the spectra, although, as far as I know, this has never been done before in astrophysics.

Bibliography

Abu Alfeilat, Haneen Arafat, Ahmad B. A. Hassanat, Omar Lasassmeh, et al. (2019). In: *Big Data* 7, pp. 221–248 (cit. on p. 19).

Arnouts, S., S. Cristiani, L. Moscardini, et al. (1999). In: *Monthly Notices of the Royal Anstronomical Society* 310, pp. 540–556 (cit. on p. 77).

Arnouts, S., E. Le Floc'h, J. Chevallard, et al. (2013). In: *Astronomy and Astrophysics* 558, A67 (cit. on pp. 9, 10).

Baldwin, J. A., M. M. Phillips, and R. Terlevich (1981). In: *Publications of the Astronomical Society of the Pacific* 93, pp. 5–19 (cit. on pp. 10, 93).

Bell, Eric F., Christian Wolf, Klaus Meisenheimer, et al. (2004). In: *The Astrophysical Journal* 608, pp. 752–767 (cit. on p. 10).

Bergh, Sidney van den (1960a). In: *The Astrophysical Journal* 131, p. 558 (cit. on p. 8).

– (1960b). In: *The Astrophysical Journal* 131, p. 558 (cit. on p. 8).

– (1960c). In: *The Astrophysical Journal* 131, p. 215 (cit. on p. 8).

Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft (1999). In: *Database Theory — ICDT'99*. Ed. by Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Catriel Beeri, and Peter Buneman. Vol. 1540. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 217–235 (cit. on p. 20).

Biernacki, C., G. Celeux, and G. Govaert (2000). In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, pp. 719–725 (cit. on p. 30).

Bom, C. R., A. Cortesi, G. Lucatelli, et al. (2021). In: *Monthly Notices of the Royal Astronomical Society* 507, pp. 1937–1955 (cit. on p. 9).

Boquien, M., D. Burgarella, Y. Roehlly, et al. (2019). In: *Astronomy & Astrophysics* 622, A103 (cit. on pp. 36, 37).

Bouveyron, Charles and Camille Brunet (2012). In: *Statist. Comput.* 22, pp. 301–324 (cit. on pp. 30, 32, 84).

Bruzual, G. and S. Charlot (2003). In: *Monthly Notices of the Royal Astronomical Society, Volume 344, Issue 4, pp. 1000-1028* 344, p. 1000 (cit. on p. 39).

Buat, V., S. Boissier, D. Burgarella, et al. (2008). In: *Astronomy and Astrophysics, Volume 483, Issue 1, 2008, pp.107-119* 483, p. 107 (cit. on p. 39).

Burgarella, D., V. Buat, and J. Iglesias-Páramo (2005). In: *Mon. Not. R. Astron. Soc.* 360, pp. 1413–1425 (cit. on p. 37).

Buta, R., S. Mitra, G. de Vaucouleurs, and H. G. Corwin Jr. (1994). In: *The Astronomical Journal* 107, p. 118 (cit. on p. 8).

Calzetti, Daniela, Lee Armus, Ralph C. Bohlin, et al. (2000). In: *Astrophysical Journal, Volume 533, Issue 2, pp. 682-695* 533, p. 682 (cit. on pp. 40, 41).

Casey, Caitlin M. (2012). In: *Monthly Notices of the Royal Astronomical Society* 425, pp. 3094–3103 (cit. on p. 41).

Chabrier, Gilles (2003). In: *Publ. Astron. Soc. Pac.* 115, pp. 763–795 (cit. on pp. 39, 43, 46).

Charlot, Stéphane and S. Michael Fall (2000). In: *Astrophysical Journal, Volume 539, Issue 2, pp. 718-731* 539, p. 718 (cit. on p. 40).

Chattopadhyay, Tanuka, Didier Fraix-Burnet, and Saptarshi Mondal (2019). In: *Publications of the Astronomical Society of the Pacific* 131, p. 108010 (cit. on p. 10).

Cheng, Ting-Yun, Christopher J. Conselice, Alfonso Aragón-Salamanca, et al. (2021). In: *Monthly Notices of the Royal Astronomical Society* 507, pp. 4425–4444 (cit. on p. 9).

Cover, T. and P. Hart (1967). In: *IEEE Trans. Inf. Theory* 13, pp. 21–27 (cit. on p. 23).

Cowie, Lennox L., Antoinette Songaila, Esther M. Hu, and J. G. Cohen (1996). In: *The Astronomical Journal* 112, p. 839 (cit. on p. 103).

Dale, Daniel A., George Helou, Georgios E. Magdis, et al. (2014). In: *The Astrophysical Journal* 784, p. 83 (cit. on p. 41).

De, Tuli, Didier Fraix Burnet, and Asis Kumar Chattopadhyay (2016). en. In: *Communications in Statistics - Theory and Methods* 45, pp. 2638–2653 (cit. on pp. 13, 81).

Dobos, László, István Csabai, Ching-Wa Yip, et al. (2012). en. In: *Monthly Notices of the Royal Astronomical Society* 420, pp. 1217–1238 (cit. on p. 12).

Draine, B. T., G. Aniano, Oliver Krause, et al. (2014). In: *The Astrophysical Journal* 780, p. 172 (cit. on p. 42).

Draine, B. T. and Aigen Li (2007). In: *The Astrophysical Journal* 657, pp. 810–837 (cit. on p. 42).

Dubois, J., D. Fraix-Burnet, J. Moultaka, P. Sharma, and D. Burgarella (2022). In: *Astronomy and Astrophysics* 663, A21 (cit. on pp. 13, 35).

Dubois, J., M. Siudek, D. Fraix-Burnet, and J. Moultaka (2023). In: in prep (cit. on p. 74).

Elmegreen, D. M. and B. G. Elmegreen (1982). In: *Monthly Notices of the Royal Astronomical Society* 201, pp. 1021–1034 (cit. on p. 8).

Elmegreen, Debra Meloy and Bruce G. Elmegreen (1987). In: *The Astrophysical Journal* 314, p. 3 (cit. on p. 8).

Feigelson, Eric D. and G. Jogesh Babu (2012). Cambridge, England, UK: Cambridge University Press (cit. on p. 16).

Ferrari, F., R. R. de Carvalho, and M. Trevisan (2015). In: *The Astrophysical Journal* 814, p. 55 (cit. on p. 21).

Fisher, R. A. (1936). In: *Annals of Eugenics* 7, pp. 179–188 (cit. on pp. 18, 21).

Fraix-Burnet, D., C. Bouveyron, and J. Moultaka (2021). en. In: *Astronomy & Astrophysics* 649, A53 (cit. on pp. 12, 13, 44, 66–69, 74, 81, 87, 88, 98, 104, 109, 111).

Fraix-Burnet, Didier (2023). In: *Monthly Notices of the Royal Astronomical Society* (cit. on p. 9).

Fraix-Burnet, Didier, Marc Thuillard, and Asis Kumar Chattopadhyay (2015). In: *Frontiers in Astronomy and Space Sciences* 2 (cit. on p. 12).

Fritz, J., A. Franceschini, and E. Hatziminaoglou (2006). In: *Monthly Notices of the Royal Astronomical Society, Volume 366, Issue 3, pp. 767-786* 366, p. 767 (cit. on p. 42).

Garilli, B., L. Guzzo, M. Scodeggio, et al. (2014). en. In: *Astronomy & Astrophysics* 562, A23 (cit. on p. 75).

Guzzo, L., M. Scodeggio, B. Garilli, et al. (2014). en. In: *Astronomy & Astrophysics* 566, A108 (cit. on p. 75).

Haines, C. P., A. Iovino, J. Krywult, et al. (2017). In: *Astronomy and Astrophysics* 605, A4 (cit. on p. 102).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). New York, NY, USA: Springer (cit. on pp. 16, 25, 82).

Helou, G., B. T. Soifer, and M. Rowan-Robinson (1985). In: *Astrophys. J.* 298, pp. L7–L11 (cit. on p. 42).

Hotelling, H. (1933). en. In: *Journal of Educational Psychology* 24, pp. 417–441 (cit. on p. 20).

Hubble, E. P. (1926). In: *The Astrophysical Journal* 64, pp. 321–369 (cit. on pp. 2, 6).

– (1936) (cit. on pp. 1–3, 5–7).

Huertas-Company, M., R. Gravet, G. Cabrera-Vives, et al. (2015). In: *The Astrophysical Journal Supplement Series* 221, p. 8 (cit. on p. 9).

Huertas-Company, M. and F. Lanusse (2023). In: *Publications of the Astronomical Society of Australia* 40, e001 (cit. on p. 24).

Ilbert, O., S. Arnouts, H. J. McCracken, et al. (2006). In: *Astronomy and Astrophysics, Volume 457, Issue 3, October III 2006, pp.841-856* 457, p. 841 (cit. on p. 77).

Inoue, Akio K. (2010). In: *Monthly Notices of the Royal Astronomical Society, Volume 401, Issue 2, pp. 1325-1333* 401, p. 1325 (cit. on p. 40).

– (2011). In: *Mon. Not. R. Astron. Soc.* 415, pp. 2920–2931 (cit. on p. 40).

Johns, Alan, Bonita Seaton, Jonathan Gal-Edd, et al. (2008). In: *Observatory Operations: Strategies, Processes, and Systems II* 7016, p. 70161D (cit. on p. 15).

Jouvin, Nicolas, Charles Bouveyron, and Pierre Latouche (2020). en. In: *arXiv:2012.04620 [stat]* (cit. on p. 112).

– (2021). In: *Statistics and Computing* 31 (cit. on pp. 66, 68).

Juneau, Stéphanie, Mark Dickinson, David M. Alexander, and Samir Salim (2011). In: *The Astrophysical Journal* 736, p. 104 (cit. on pp. 10, 11, 93).

Kasa, Siva Rajesh and Vaibhav Rajan (2020). In: *arXiv e-prints*, arXiv:2007.12786, arXiv:2007.12786 (cit. on p. 66).

Kauffmann, Guinevere, Timothy M. Heckman, Simon D. M. White, et al. (2003). In: *Monthly Notices of the Royal Astronomical Society* 341, pp. 33–53 (cit. on pp. 10, 99, 100).

Kennicutt Jr., Robert C. (1992). en. In: *The Astrophysical Journal Supplement Series* 79, p. 255 (cit. on pp. 8, 11).

Khramtsov, V., I. B. Vavilova, D. V. Dobrycheva, et al. (2022). In: *Kosmichna Nauka i Tekhnologiya* 28, pp. 27–55 (cit. on p. 9).

Kramer, M.A. (1992). en. In: *Computers & Chemical Engineering* 16, pp. 313–328 (cit. on p. 20).

Kramer, Mark A. (1991). en. In: *AIChE Journal* 37, pp. 233–243 (cit. on p. 20).

Kroupa, Pavel (2001). In: *Monthly Notices of the Royal Astronomical Society, Volume 322, Issue 2, pp. 231-246* 322, p. 231 (cit. on p. 40).

Lee, Joon Hyeop, Myung Gyoon Lee, Changbom Park, and Yun-Young Choi (2008). In: *Monthly Notices of the Royal Astronomical Society* 389, pp. 1791–1804 (cit. on p. 12).

Leitherer, Claus, I.-Hui Li, Daniela Calzetti, and Timothy M. Heckman (2002). In: *Astrophys. J. Suppl. Ser.* 140, pp. 303–329 (cit. on pp. 40, 41).

Leka, K. D. and G. Barnes (2007). In: *The Astrophysical Journal* 656, pp. 1173–1186 (cit. on p. 21).

MacQueen, J. (1967). In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Vol. 5.1. Ewing, NJ, USA: University of California Press, pp. 281–298 (cit. on p. 24).

Maraston, Claudia (2005). In: *Monthly Notices of the Royal Astronomical Society, Volume 362, Issue 3, pp. 799-825* 362, p. 799 (cit. on pp. 39, 40).

Marchetti, A., B. Garilli, B. R. Granett, et al. (2017). en. In: *Astronomy & Astrophysics* 600, A54 (cit. on pp. 75, 78, 79).

Meiksin, Avery (2006). In: *Monthly Notices of the Royal Astronomical Society, Volume 365, Issue 3, pp. 807-812* 365, p. 807 (cit. on p. 43).

Morgan, W. W. (1958). In: *Publications of the Astronomical Society of the Pacific* 70, p. 364 (cit. on p. 8).

– (1959). In: *Publications of the Astronomical Society of the Pacific* 71, p. 394 (cit. on p. 8).

Moutard, T., S. Arnouts, O. Ilbert, et al. (2016a). In: *Astron. Astrophys.* 590, A102 (cit. on p. 77).

Moutard, T., S. Arnouts, O. Ilbert, et al. (2016b). In: *Astron. Astrophys.* 590, A103 (cit. on p. 77).

Nevin, R., L. Blecha, J. Comerford, et al. (2023). In: *Monthly Notices of the Royal Astronomical Society* 522, pp. 1–28 (cit. on p. 21).

Noll, S., D. Burgarella, E. Giovannoli, et al. (2009). In: *Astron. Astrophys.* 507, pp. 1793–1813 (cit. on p. 37).

Pearson, Karl (1901). en. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, pp. 559–572 (cit. on p. 20).

Reynolds, J. H. (1927). In: *The Observatory* 50, pp. 185–189 (cit. on p. 8).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). In: *ACL Anthology*, pp. 97–101 (cit. on p. 24).

Roberts, Morton S. and M. P. Haynes (1994). In: *Annual Review of Astronomy and Astrophysics* 32, pp. 115–152 (cit. on p. 8).

Salpeter, Edwin E. (1955). In: *Astrophys. J.* 121, p. 161 (cit. on pp. 39, 40).

Sandage, Allan (1961) (cit. on p. 7).

Sandage, Allan and John Bedke (1994). Vol. 638 (cit. on p. 7).

Sandage, Allan and G. A. Tammann (1981) (cit. on p. 7).

Schutter, A. and L. Shamir (2015). en. In: *Astronomy and Computing* 12, pp. 60–66 (cit. on p. 9).

Scodeggio, M., L. Guzzo, B. Garilli, et al. (2018). In: *Astronomy & Astrophysics* 609, A84 (cit. on p. 75).

Shamir, Lior (2009). In: *Monthly Notices of the Royal Astronomical Society* 399, pp. 1367–1372 (cit. on p. 9).

Sheth, Kartik, Michael W. Regan, Nicholas Z. Scoville, and Linda E. Strubbe (2003). In: *The Astrophysical Journal* 592, pp. L13–L16 (cit. on p. 8).

Siudek, M., K. Małek, A. Pollo, et al. (2018a). en. In: *Astronomy & Astrophysics* 617, A70 (cit. on pp. 10, 77).

Siudek, M., K. Małek, A. Pollo, et al. (2018b). en. In: *arXiv:1805.09905 [astro-ph]* (cit. on p. 10).

Strateva, Iskra, Željko Ivezić, Gillian R. Knapp, et al. (2001). In: *The Astronomical Journal* 122, pp. 1861–1874 (cit. on p. 8).

Sánchez Almeida, J., J. A. L. Aguerri, C. Muñoz-Tuñón, and A. de Vicente (2010). In: *The Astrophysical Journal* 714, pp. 487–504 (cit. on p. 13).

Sánchez Almeida, J. and C. Allende Prieto (2013). In: *The Astrophysical Journal* 763, p. 50 (cit. on p. 13).

Vaucouleurs, Gerard de (1959). In: *Handbuch der Physik* 53, p. 275 (cit. on p. 7).

Vavilova, I. B., D. V. Dobrycheva, M. Yu. Vasylenko, et al. (2021). In: *Astronomy and Astrophysics* 648, A122 (cit. on p. 9).

Vavilova, I. B., V. Khramtsov, D. V. Dobrycheva, et al. (2022). In: *Kosmichna Nauka i Tekhnologiya* 28, pp. 03–22 (cit. on p. 9).

Walmsley, Mike, Chris Lintott, Tobias Géron, et al. (2022). In: *Monthly Notices of the Royal Astronomical Society* 509, pp. 3966–3988 (cit. on p. 9).

Wang, Li-Li, A.-Li Luo, Shi-Yin Shen, et al. (2018). en. In: *Monthly Notices of the Royal Astronomical Society* 474, pp. 1873–1885 (cit. on p. 12).

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.

List of publications

A

Publications

1. Dubois, J., D. Fraix-Burnet, J. Moultaka, P. Sharma, and D. Burgarella. 'Unsupervised Classification of CIGALE Galaxy Spectra'. Astronomy and Astrophysics 663 (1 July 2022): A21. https://doi.org/10.1051/0004-6361/202141729.
2. Dubois, J. M. Siudek, D. Fraix-Burnet, J. Moultaka. 'Unsupervised classification of spectra of $0.4 < z < 1.2$ galaxies from the VIMOS Public Extragalactic Redshift Survey (VIPERS)', in prep.

Referred proceedings

1. Dubois J., Fraix-Burnet D., Moultaka J., « An unsupervised approach to galaxy spectra classification », EAS-2022: Proceedings of the annual meeting of the European Astronomical Society. In review for a publication as a referred contribution on Memorie della Società Astronomica Italiana.
2. Dubois J., Fraix-Burnet D., Moultaka J., « Clustering of galaxy spectra: an unsupervised approach with Fisher-EM », ML4Astro-2022: Proceedings of the International conference on machine-learning for astrophysics. Accepted for publication as a referred contribution on Springer Nature.

Contributed talks

1. EAS Annual Meeting (Valencia, Spain | July 2022) – 'An unsupervised approach to galaxy spectra classification'
2. PNCG Annual Meeting (Strasbourg, France | June 2022) – 'An unsupervised approach to galaxy spectra classification'
3. ML4Astro (Catania, Italy | May 2022) – 'Clustering of galaxy spectra: an unsupervised approach with Fisher-EM'
4. SFdS Annual Meeting (Remote | June 2021) – 'GMM-based unsupervised classification in latent subspace: an application to galaxy spectra'

5. SF2A Annual Meeting (Remote | June 2021) – 'Unsupervised classification of CIGALE galaxy spectra'

B

Classification of a CIGALE-simulated sample: results for all S/N

B.0.1 Parameter distribution

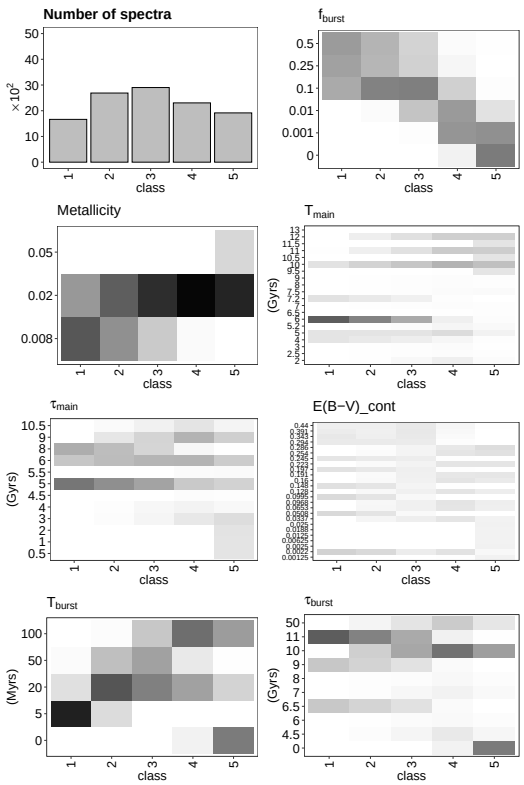


Fig. B.1.: S/N=1 (see Fig. 3.7).

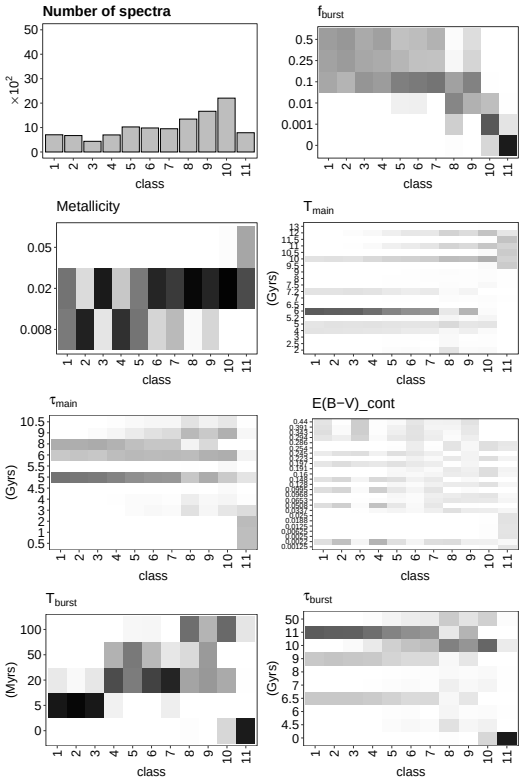


Fig. B.2.: S/N=3 (see Fig. 3.7).

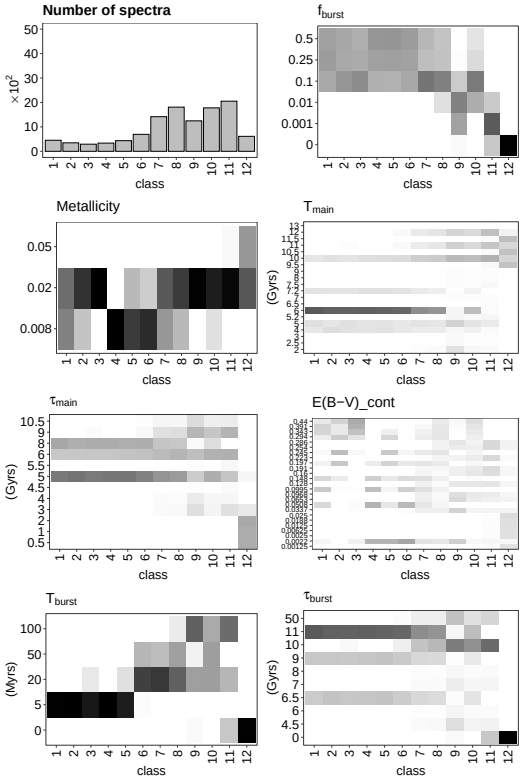


Fig. B.3.: S/N=5 (see Fig. 3.7).

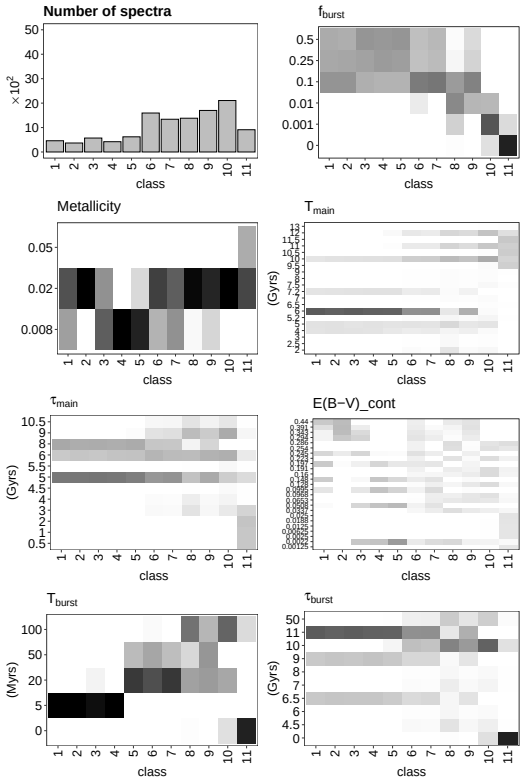


Fig. B.4.: S/N=10 (see Fig. 3.7).

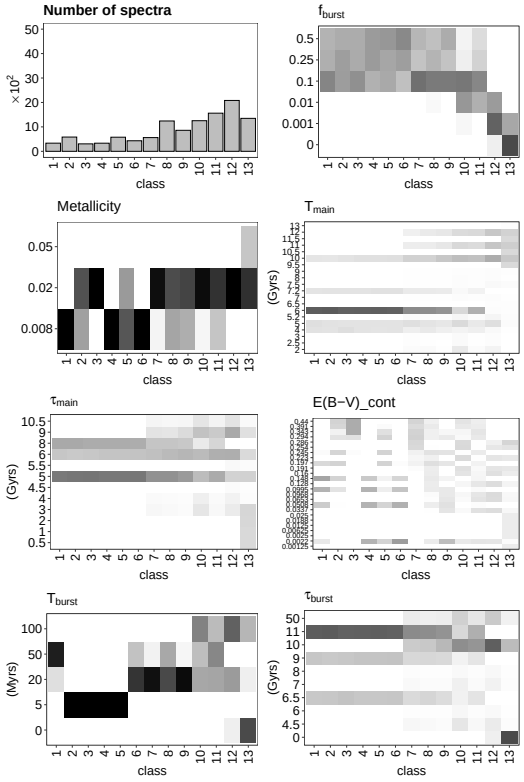


Fig. B.5.: S/N=100 (see Fig. 3.7).

B.0.2 LDA analysis

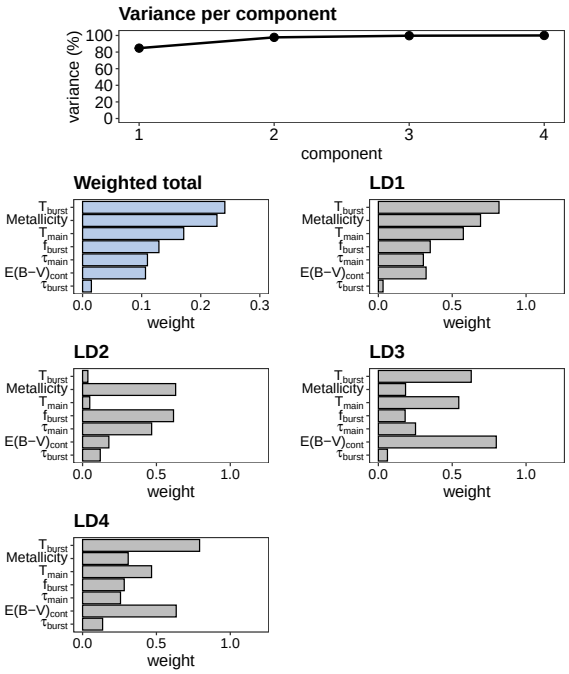


Fig. B.6.: S/N=1 (see Fig. 3.9).

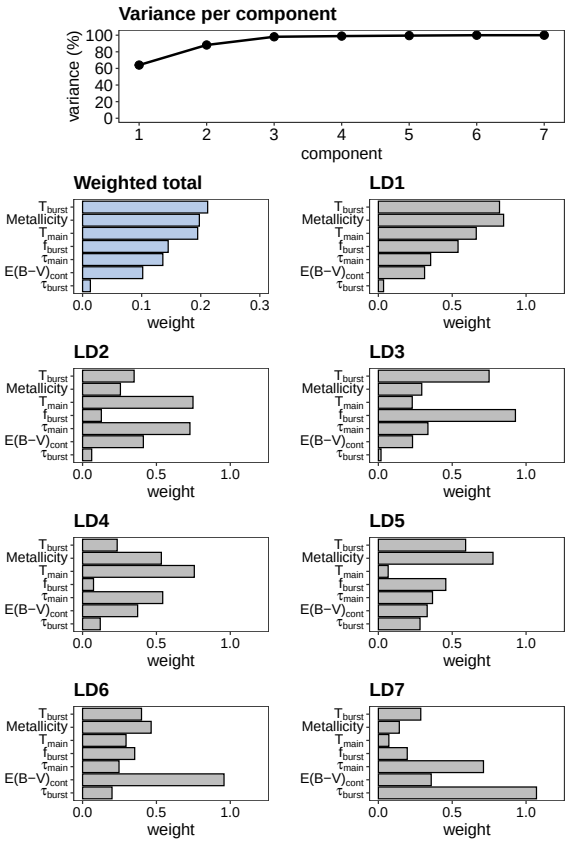


Fig. B.7.: S/N=3 (see Fig. 3.9).

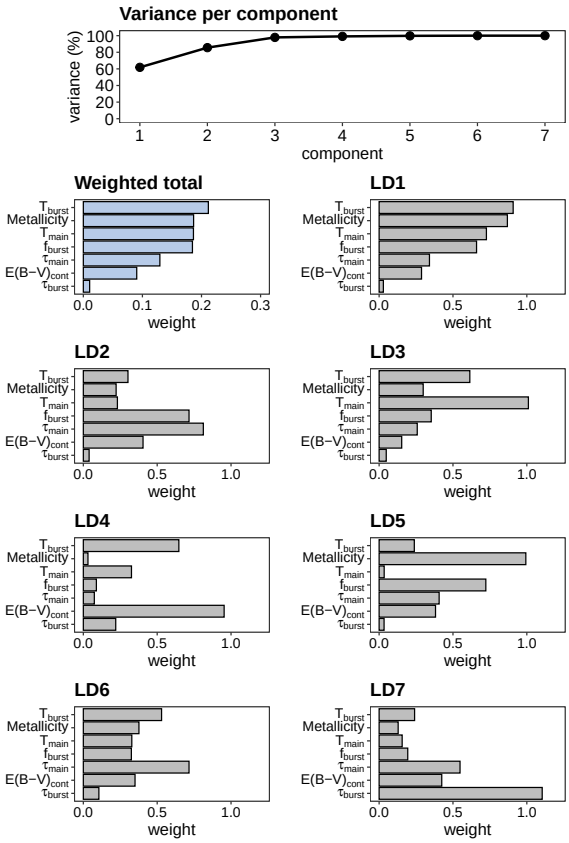


Fig. B.8.: S/N=5 (see Fig. 3.9).

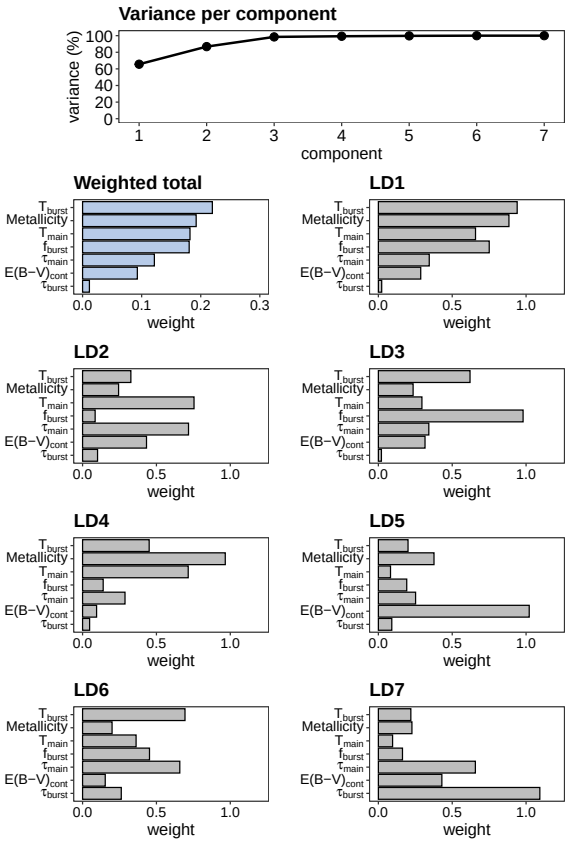


Fig. B.9.: S/N=10 (see Fig. 3.9).

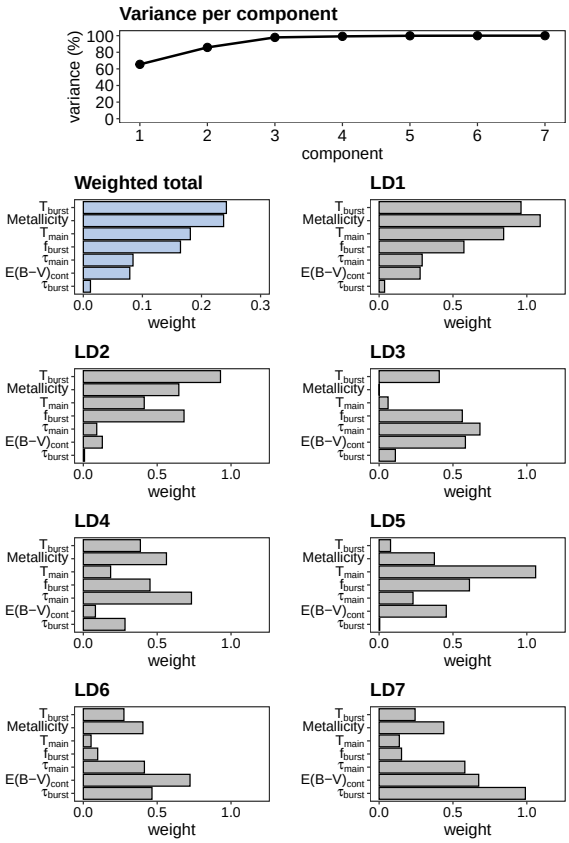


Fig. B.10.: S/N=100 (see Fig. 3.9).

B.0.3 Mean spectra

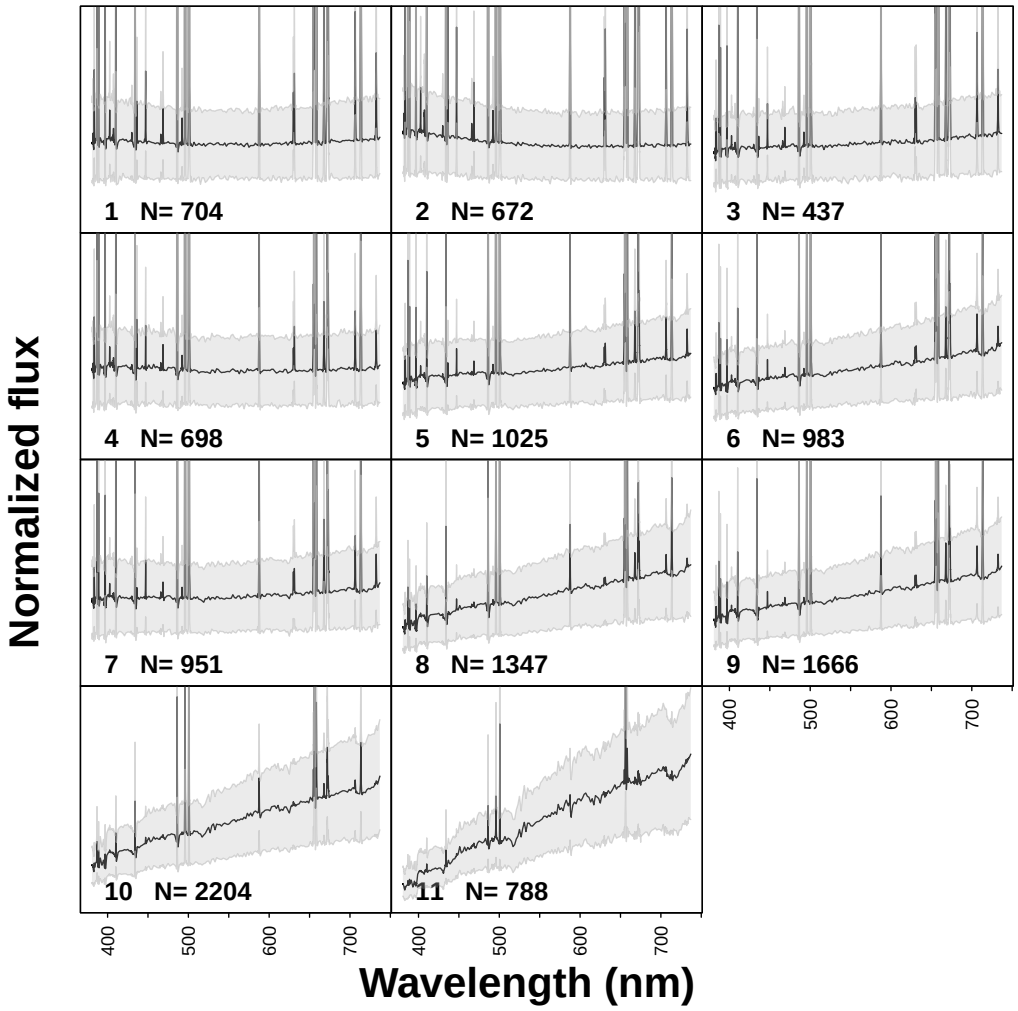


Fig. B.11.: S/N=3 (see Fig. 3.5).

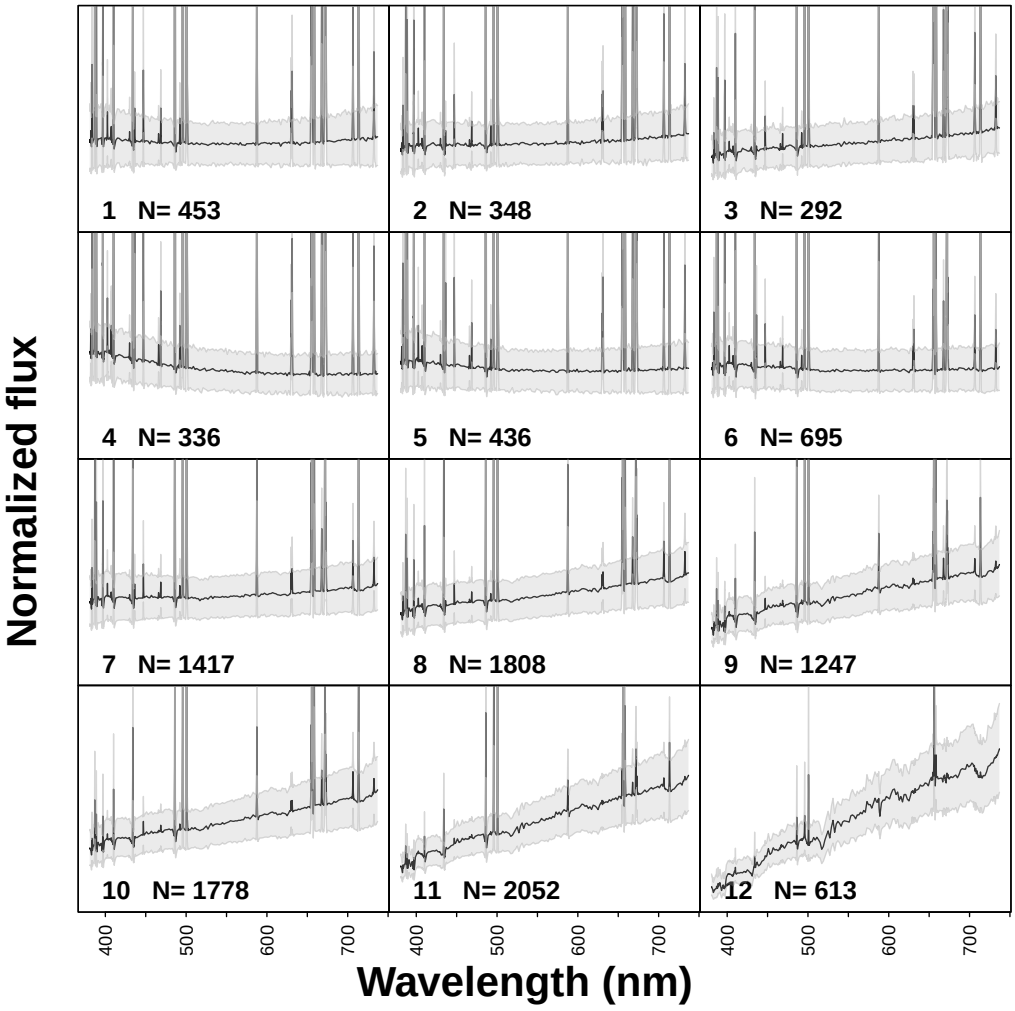


Fig. B.12.: S/N=5 (see Fig. 3.5).

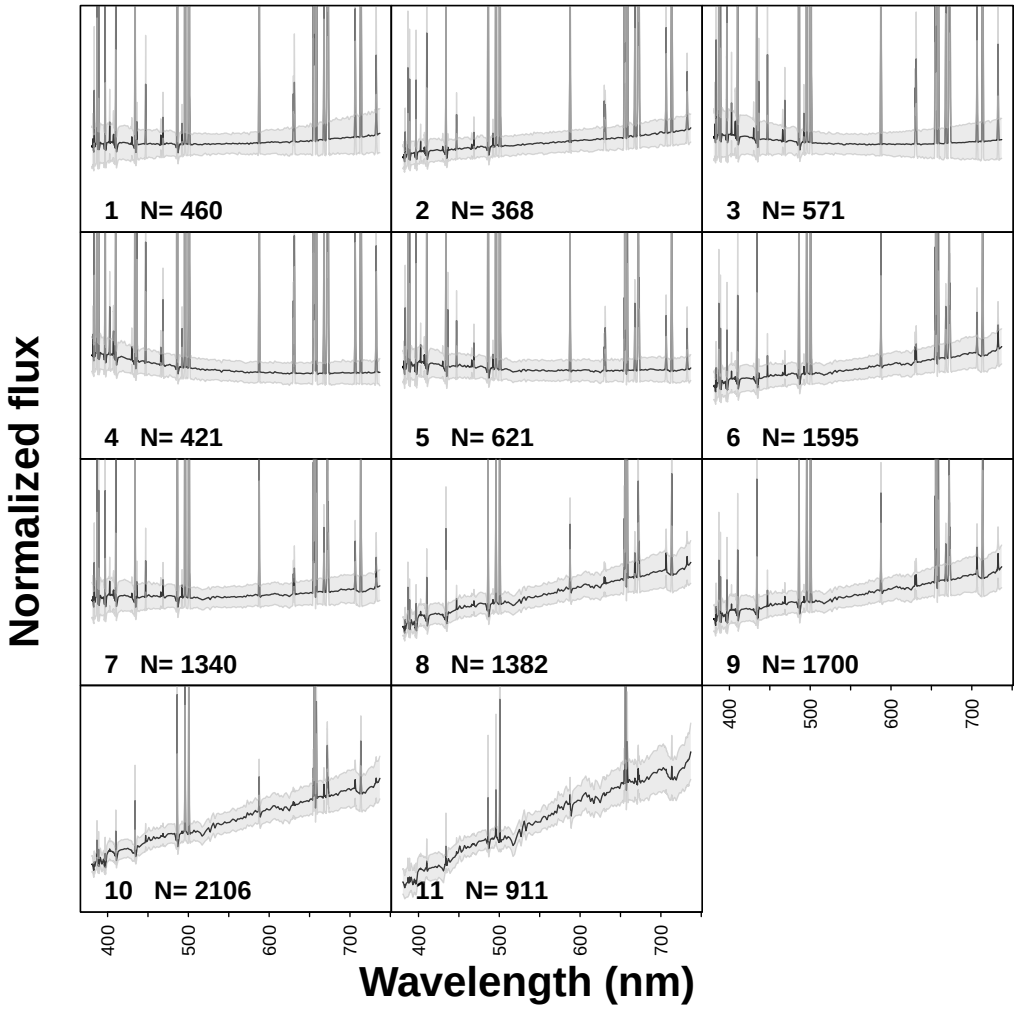


Fig. B.13.: S/N=10 (see Fig. 3.5).

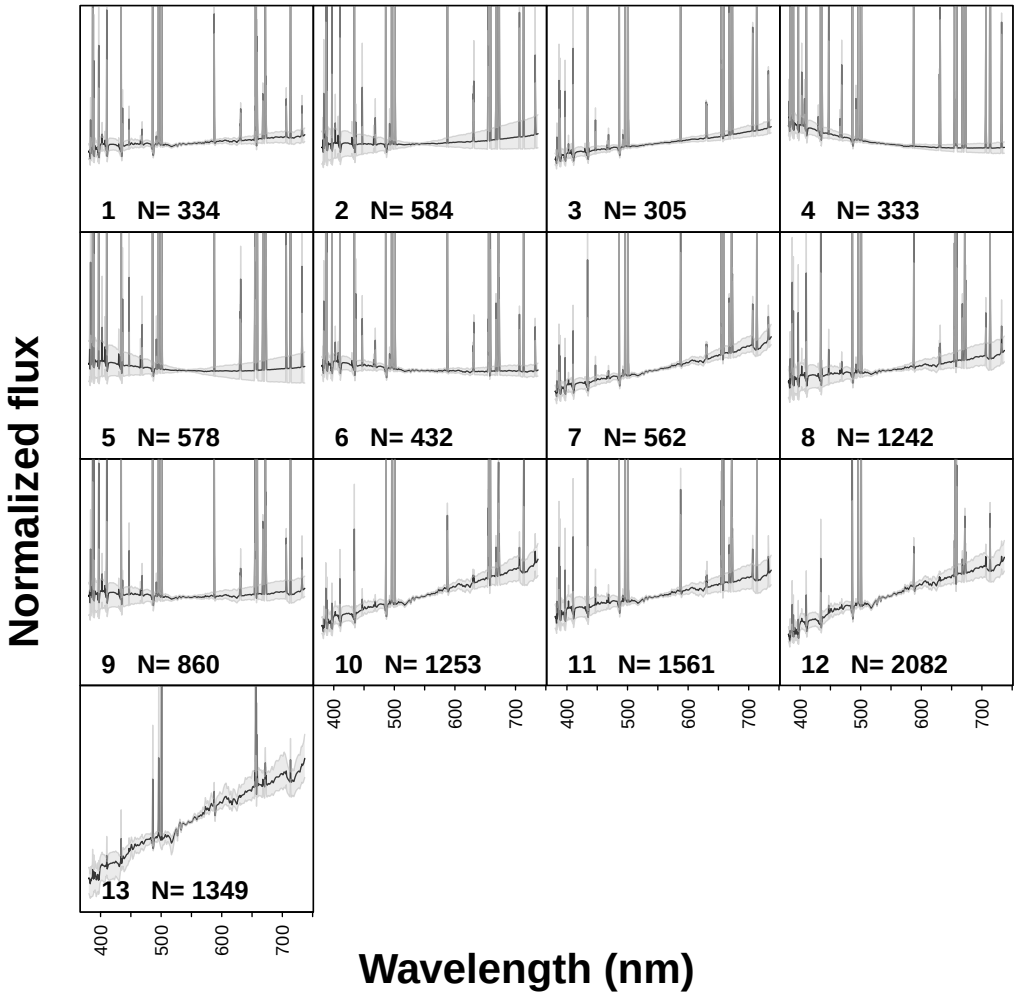


Fig. B.14.: S/N=100 (see Fig. 3.5).

C

Classification of a CIGALE-simulated sample: toy model

To try to visualise the conditions for the gap at K=13 in the ICL curve for the noiseless case (Sect. 3.4.1 and Fig. 3.3), we constructed a toy model by trial and error. The simplest sample that can be built to reproduce this behaviour is a sample of five variables and 1000 observations to ensure a perfect reproducibility.

We consider the following matrix made with the following five variables:

$$
\begin{array}{ll}
\text{Var1}[1:500] = 1 & \text{Var1}[501:1000] = 2 \\
\text{Var2}[1:300] = \mathcal{N}(10, 0.01) & \text{Var2}[301:1000] = \mathcal{N}(15, 0.01) \\
\text{Var3}[1:800] = 1 & \text{Var3}[801:1000] = 2 \\
\text{Var4}[1:500] = \mathcal{N}(100, 0.01) & \text{Var4}[501:1000] = \mathcal{N}(150, 0.05) \\
\text{Var5}[1:200] = 1 & \text{Var5}[201:1000] = \mathcal{N}(4, 0.1)
\end{array}
$$

where VarX$[i:j]$ designates the indices from i to j of variable VarX, and $\mathcal{N}(\mu, \sigma^2)$ means that the values are drawn from a normal distribution of mean μ and standard deviation σ.

This sample (Fig C.2) yields an ICL curve Fig C.1 with a gap at K=3 (*Fisher-EM* never converges) and a much higher value at K=4 than at K=2. This behaviour is identical to the one at K=13 in Fig. 3.3.

Adding some dispersion in Var2 by increasing σ^2 from 0.01 to 0.05,

$$
\text{Var2}[1:300] = \mathcal{N}(10, 0.05) \quad \text{Var2}[301:1000] = \mathcal{N}(15, 0.05)
$$

as represented by the red points in Fig C.2, the *Fisher-EM* analysis always yields a solution (Fig C.3).

This behaviour is thus very similar to the one obtained on the CIGALE sample, and is thus explained by the very peculiar distribution of the observations in the multivariate data space.

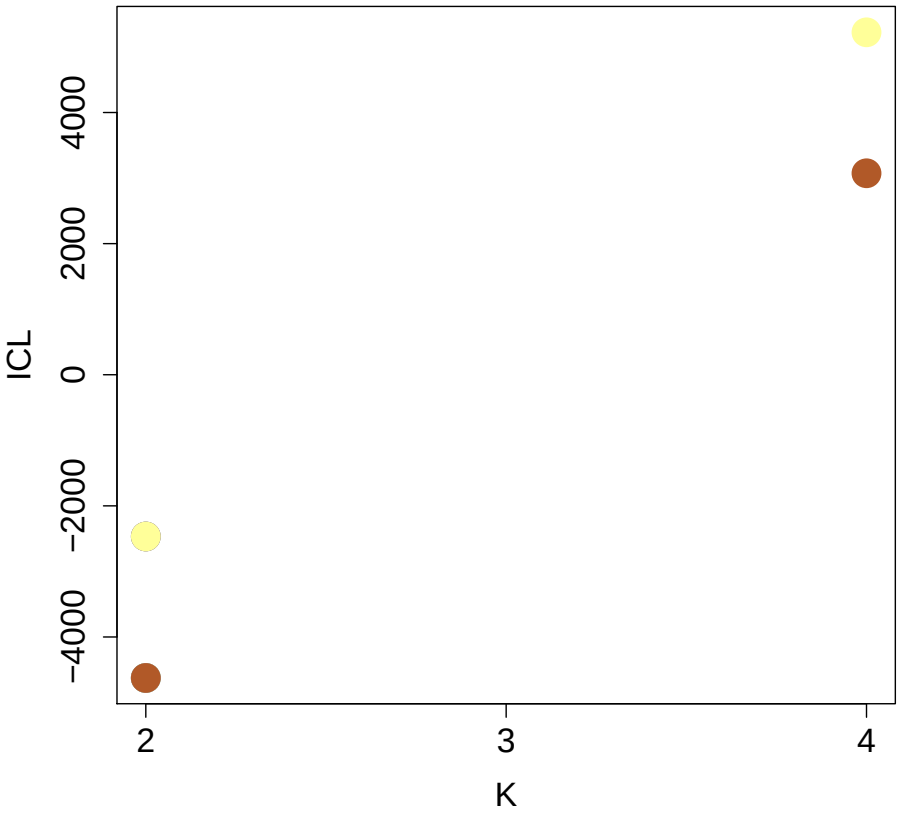


Fig. C.1.: All the ICL values as a function of the number of cluster K obtained for the toy model. Each point corresponds to a successful run of *Fisher-EM* for one of the 12 statistical models. This figure should be compared with Fig. 3.3.

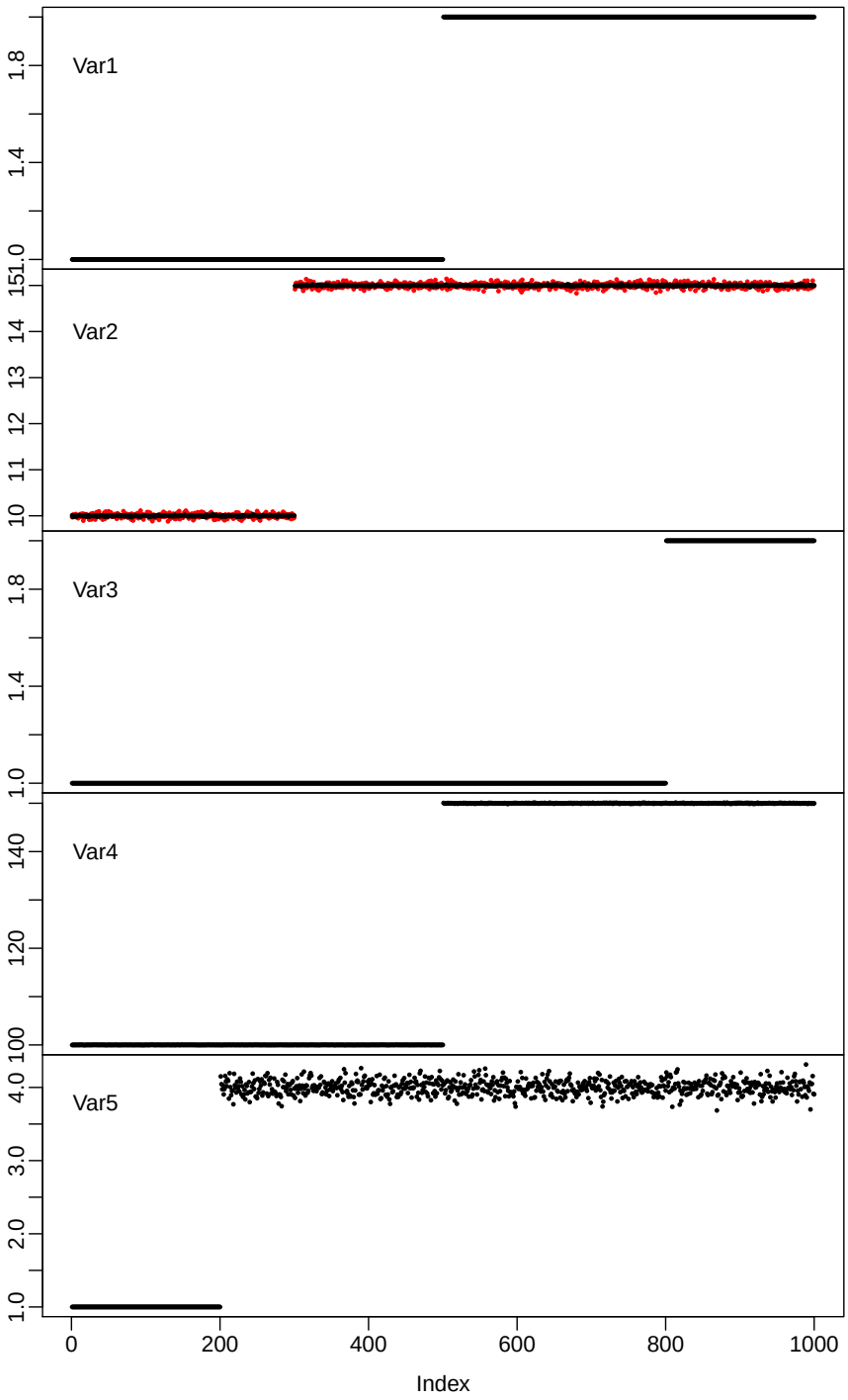


Fig. C.2.: Values of the five variables for the toy model with 1000 observations that yield the ICL curve in Fig C.1. The points in red in the second panel (Var2) show a slightly increased dispersion that yields the ICL curve in Fig C.3.

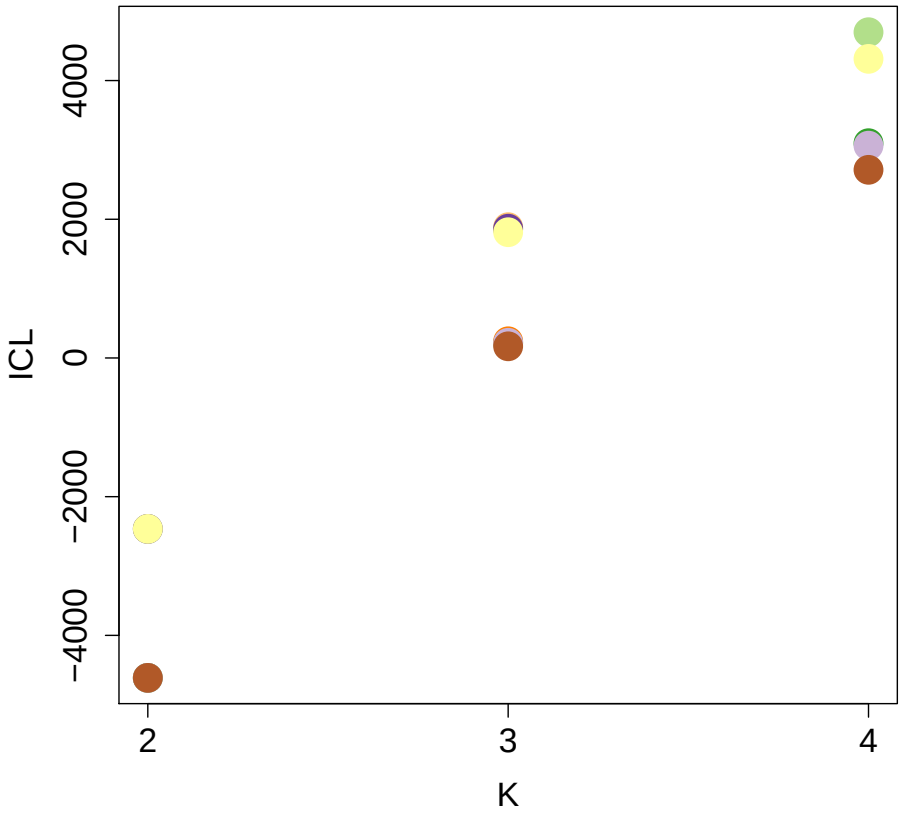


Fig. C.3.: Same as Fig. C.1 with the slightly more dispersed variable Var2.

D

VIPERS classification: stacked spectra

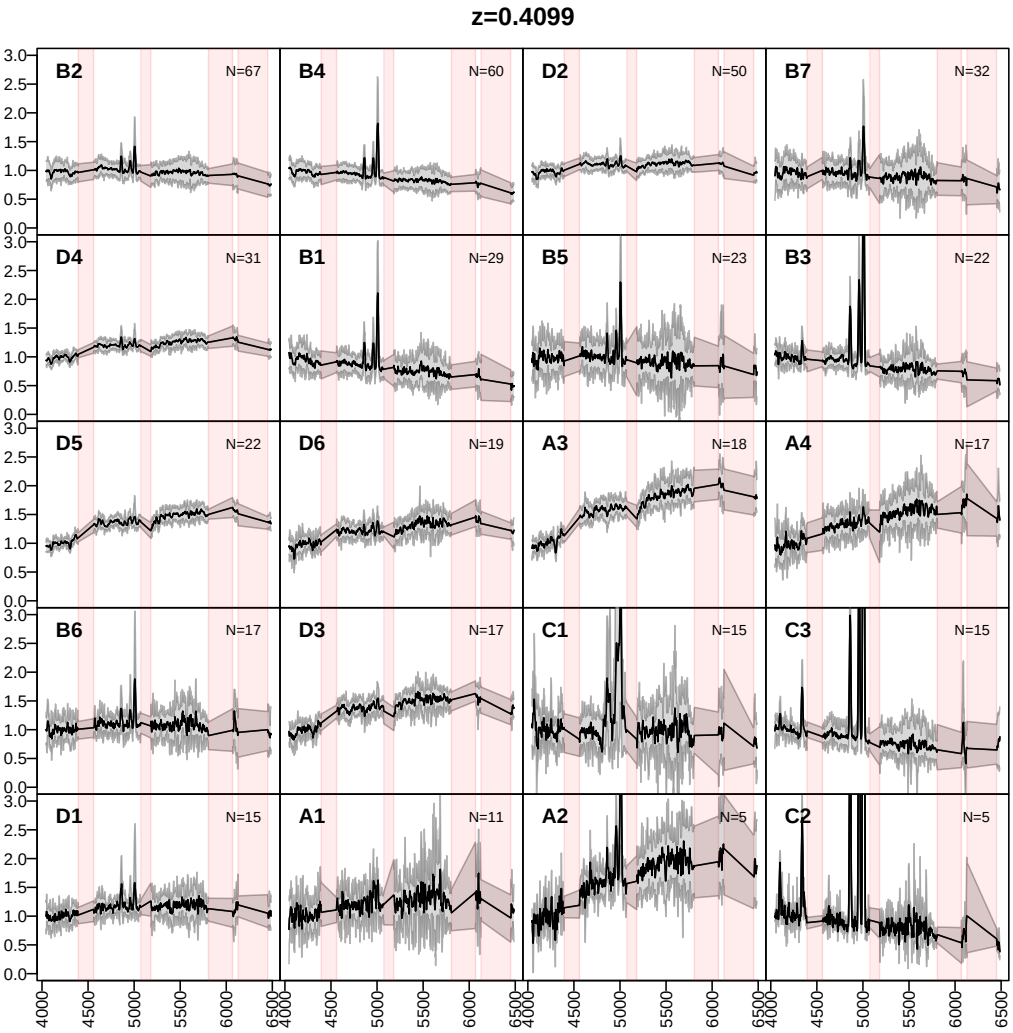


Fig. D.1.: Stacked spectra of the classes of bin 1 (see Fig. 4.10 for further information)

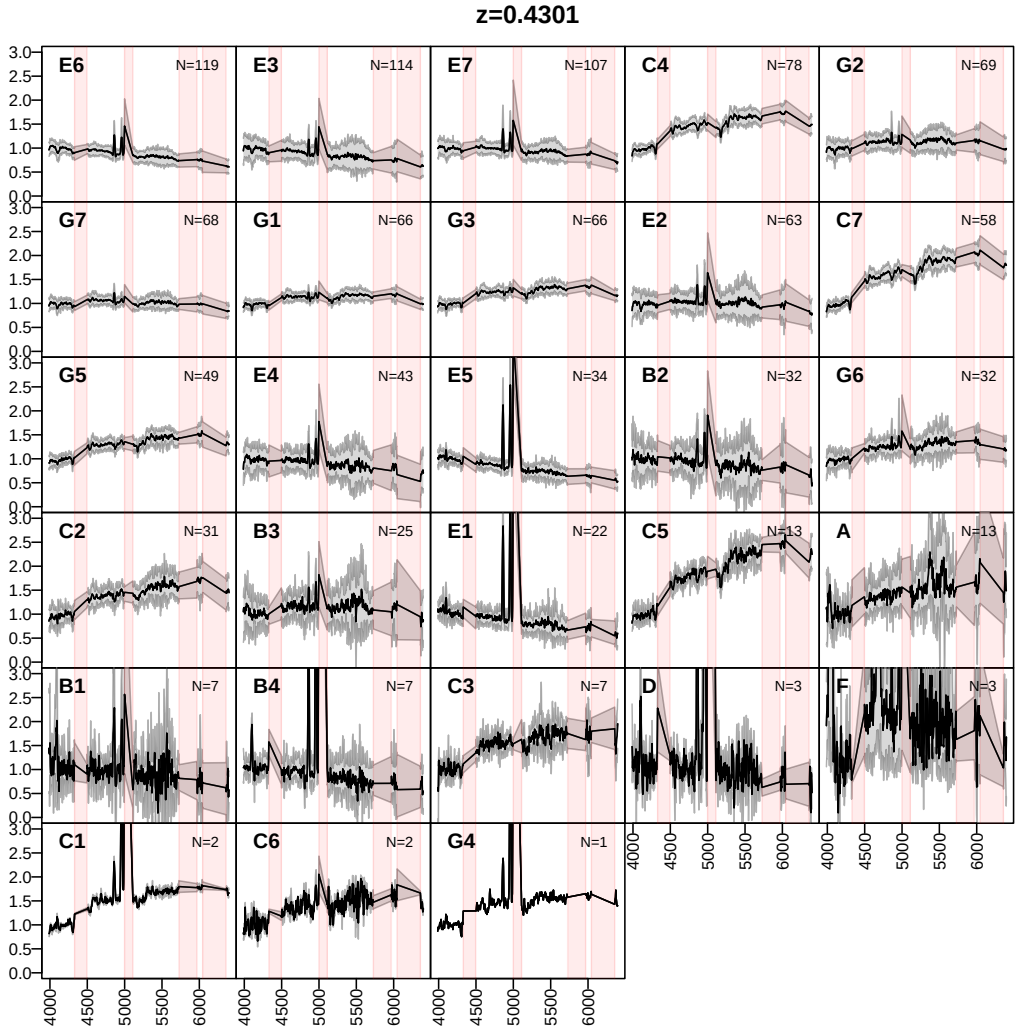


Fig. D.2.: Stacked spectra of the classes of bin 2 (see Fig. 4.10 for further information)

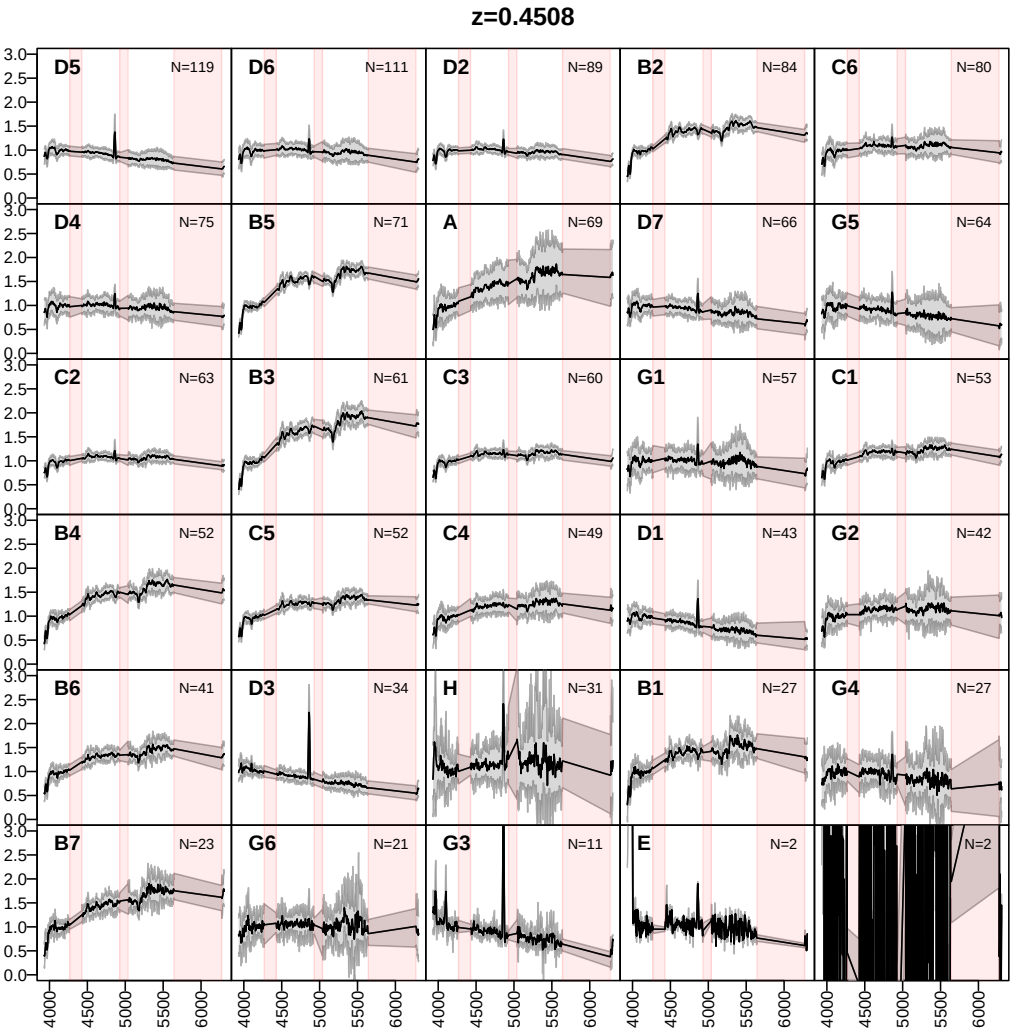


Fig. D.3.: Stacked spectra of the classes of bin 3 (see Fig. 4.10 for further information)

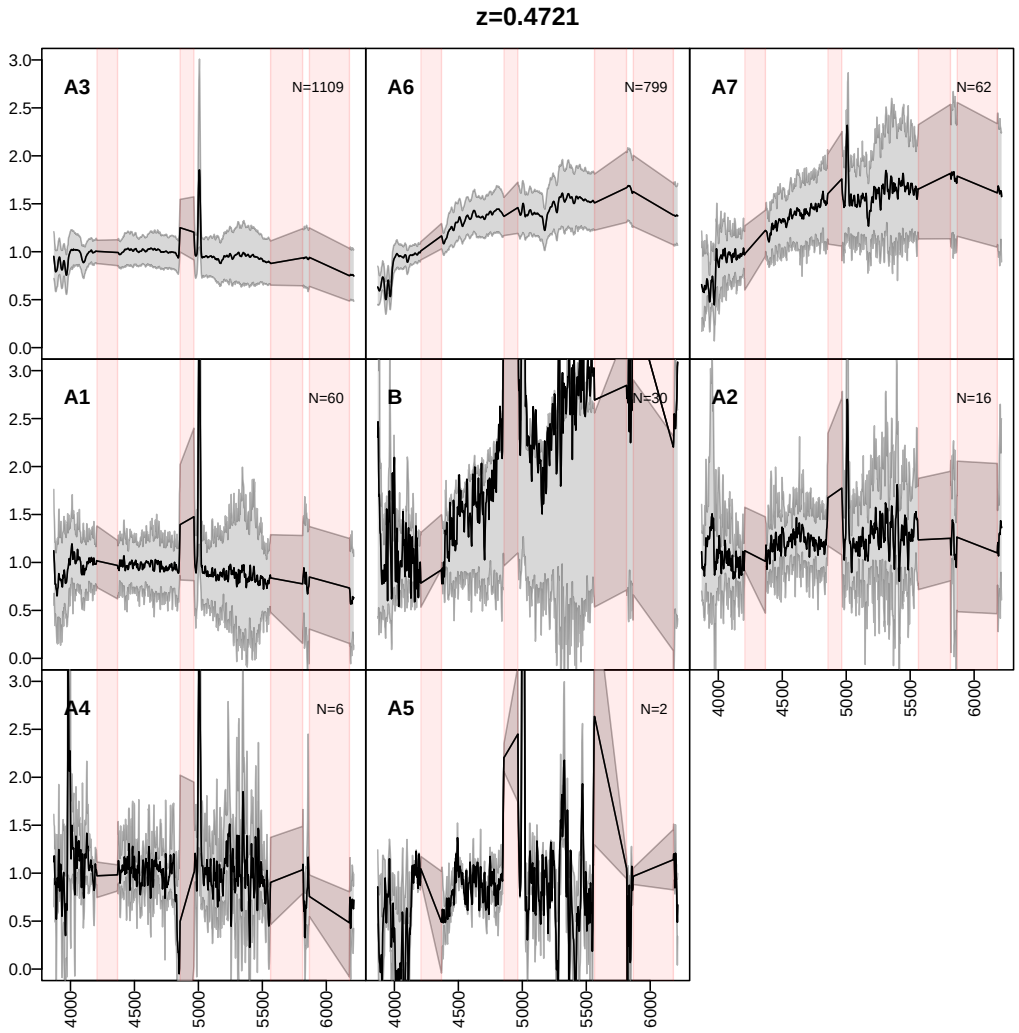


Fig. D.4.: Stacked spectra of the classes of bin 4 (see Fig. 4.10 for further information)

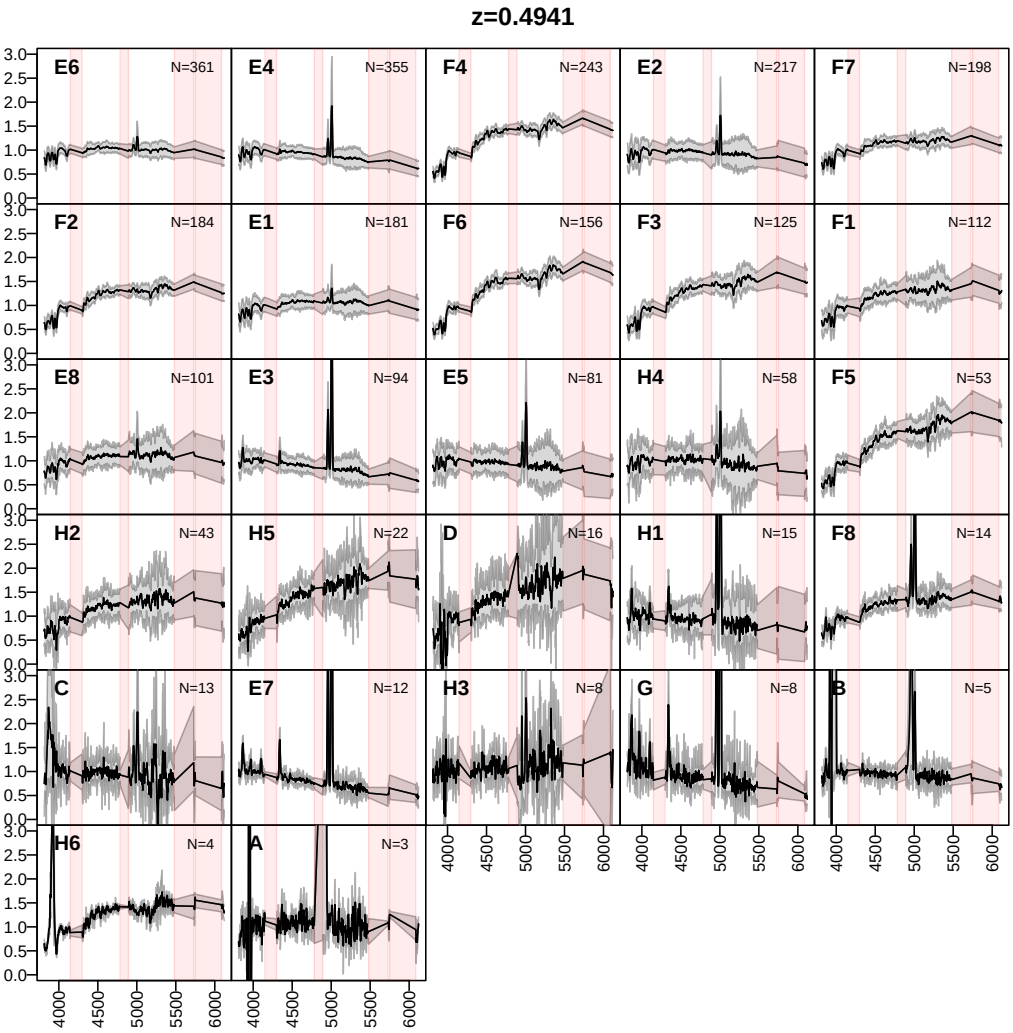


Fig. D.5.: Stacked spectra of the classes of bin 5 (see Fig. 4.10 for further information)

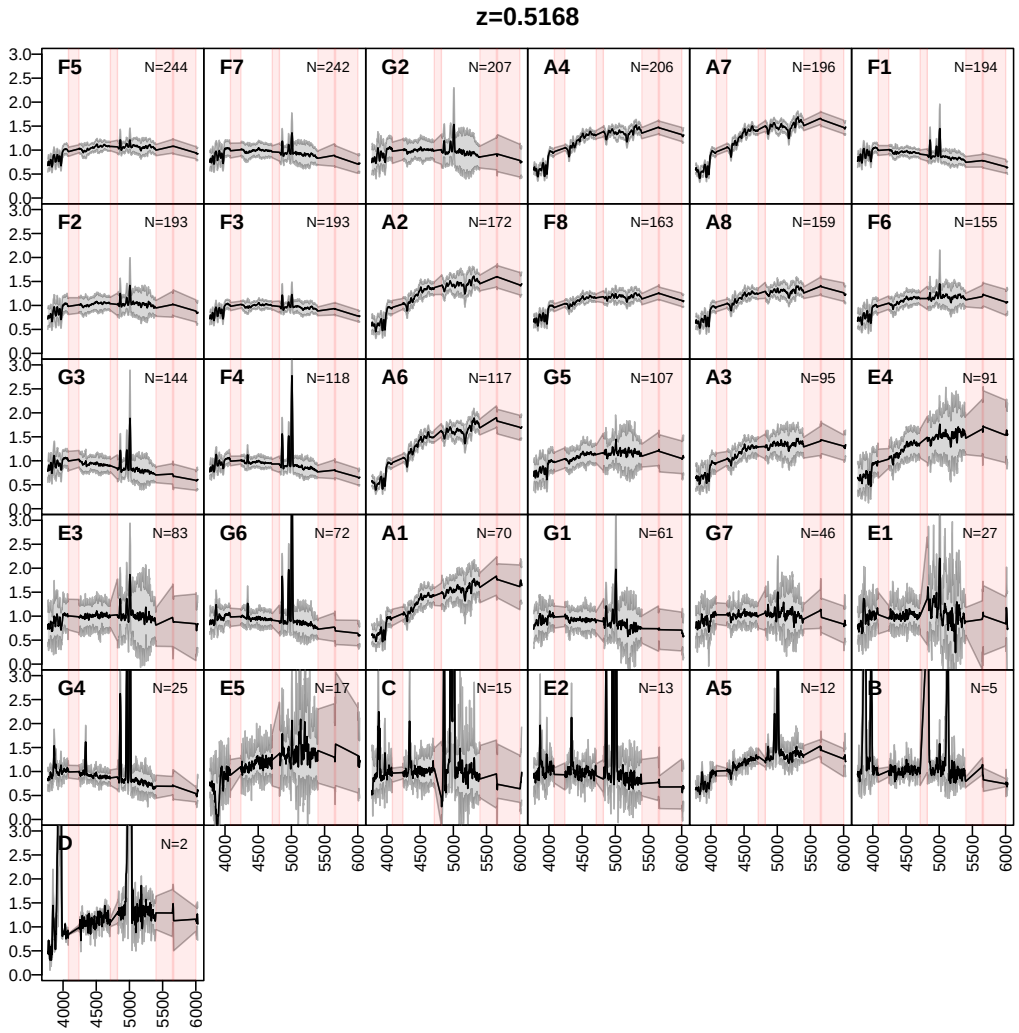


Fig. D.6.: Stacked spectra of the classes of bin 6 (see Fig. 4.10 for further information)

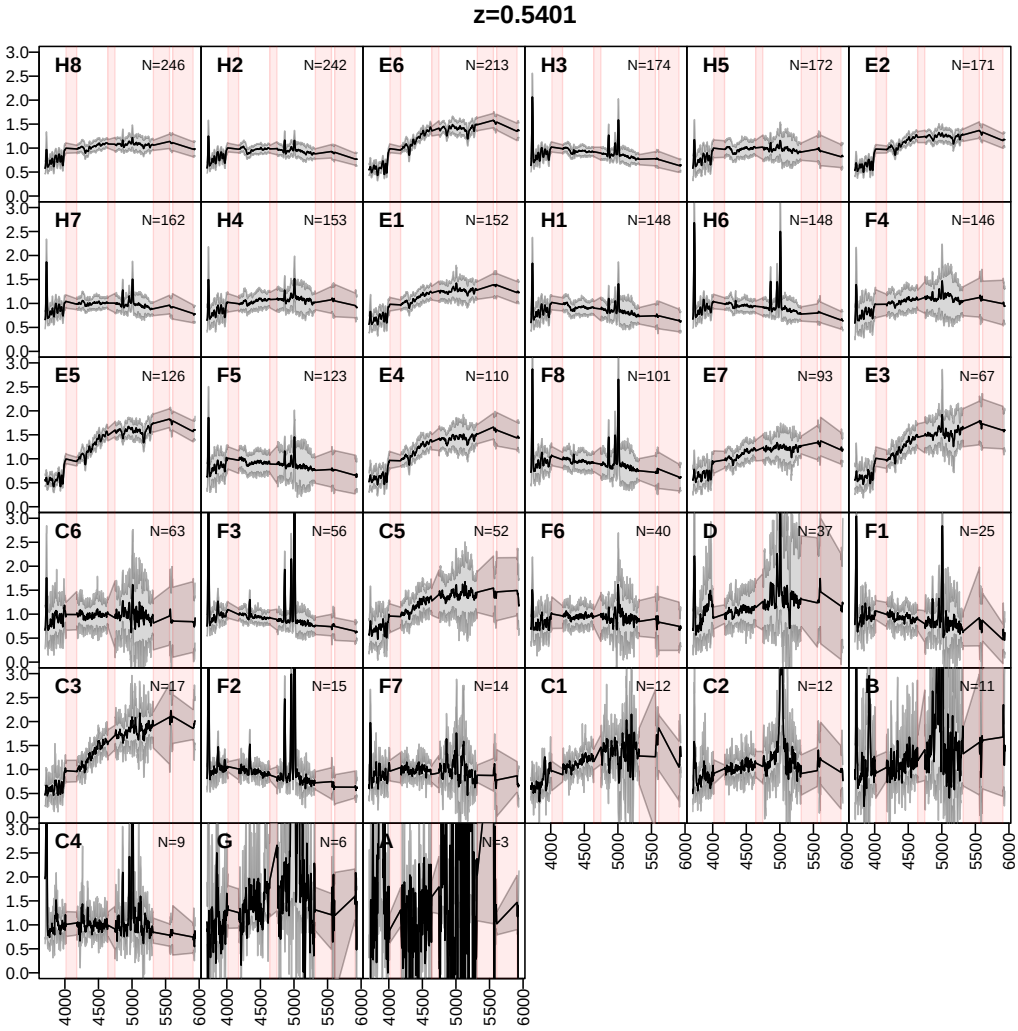


Fig. D.7.: Stacked spectra of the classes of bin 7 (see Fig. 4.10 for further information)

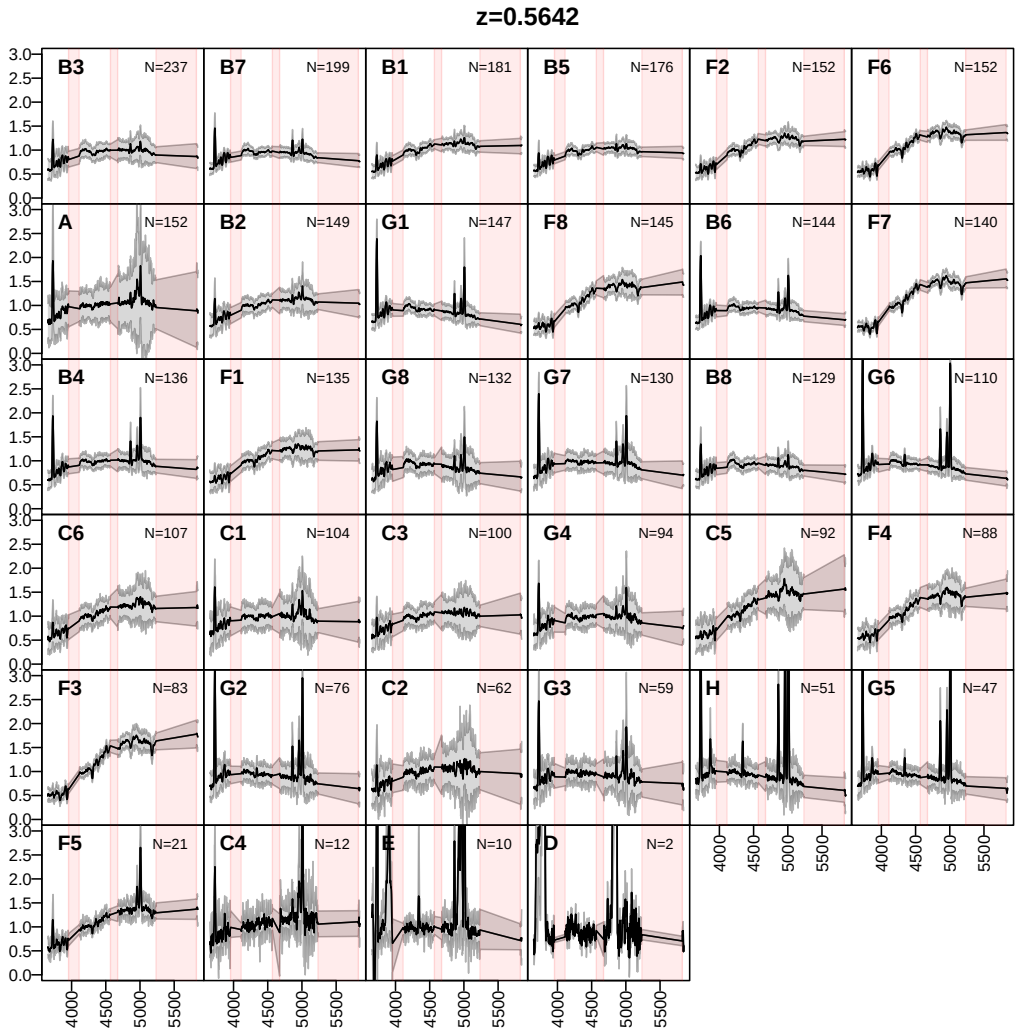


Fig. D.8.: Stacked spectra of the classes of bin 8 (see Fig. 4.10 for further information)

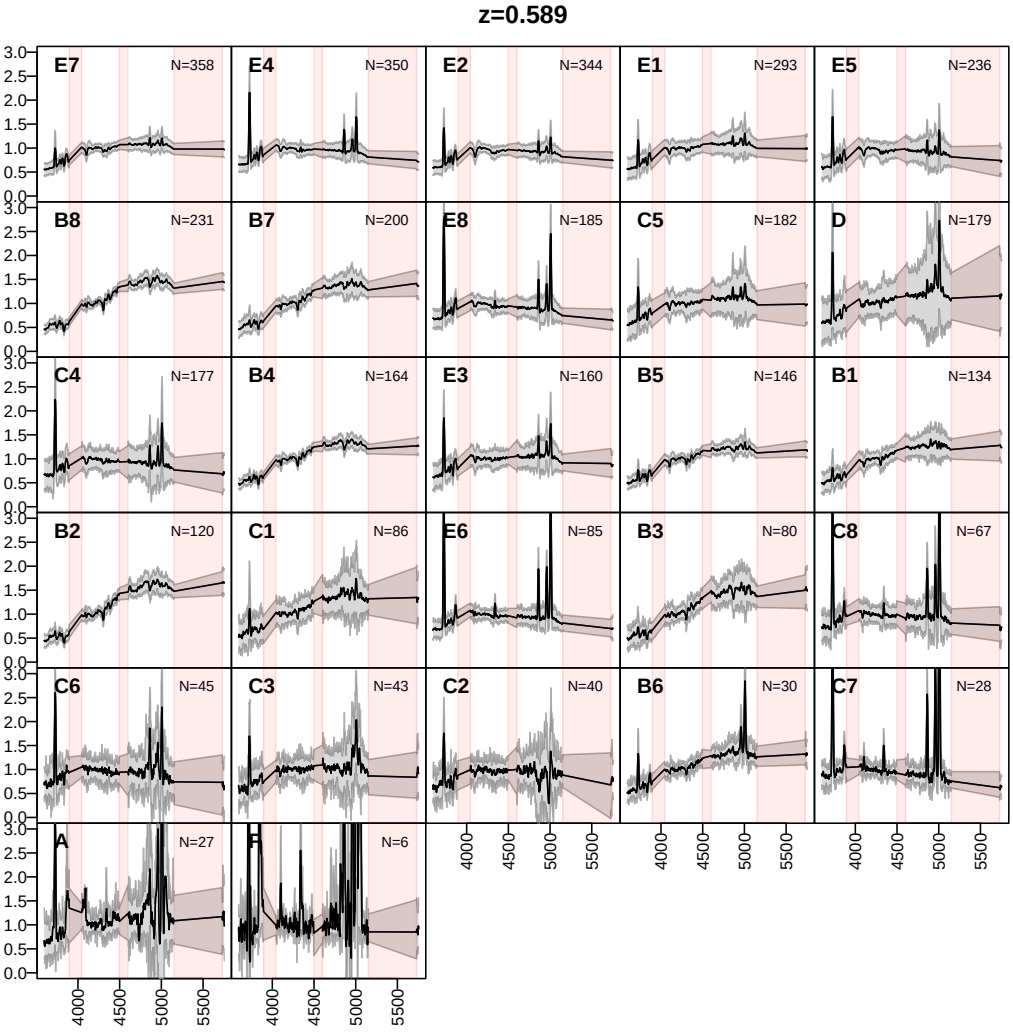


Fig. D.9.: Stacked spectra of the classes of bin 9 (see Fig. 4.10 for further information)

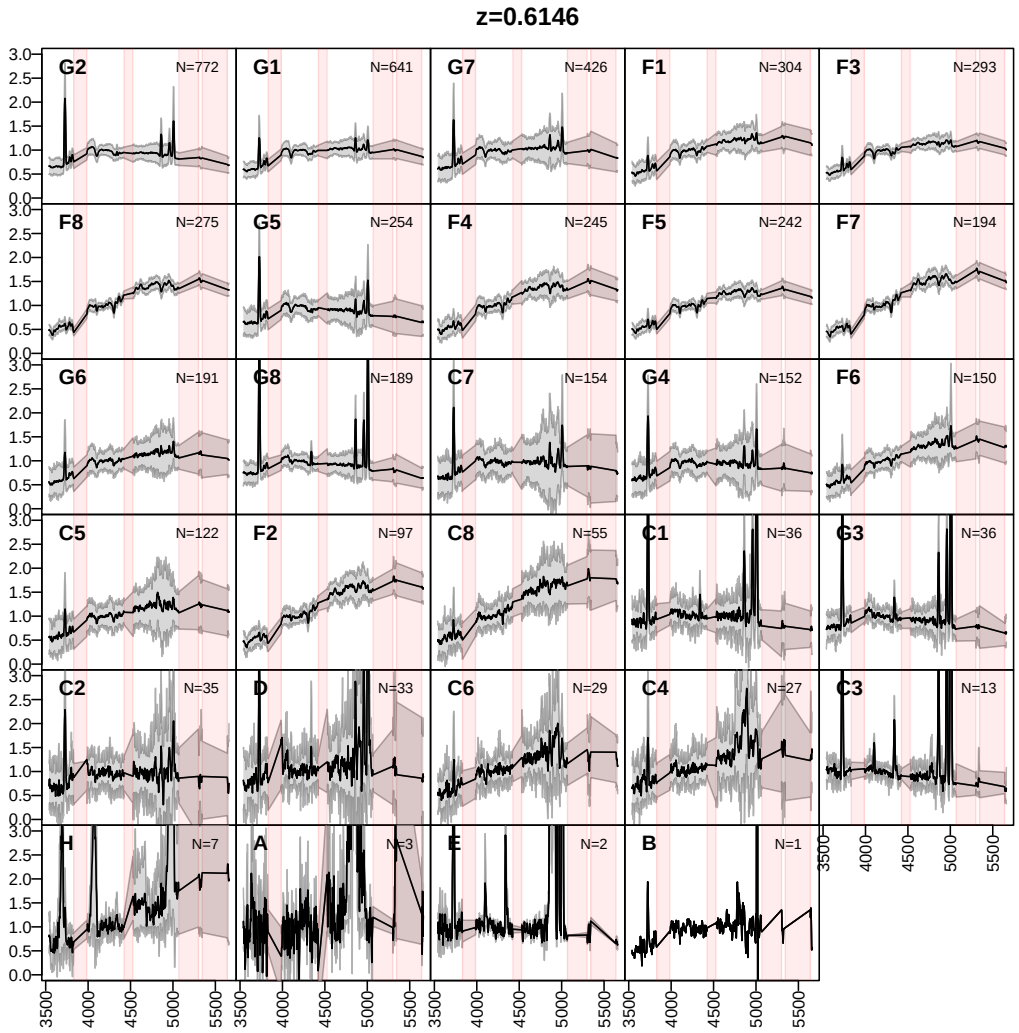


Fig. D.10.: Stacked spectra of the classes of bin 10 (see Fig. 4.10 for further information)

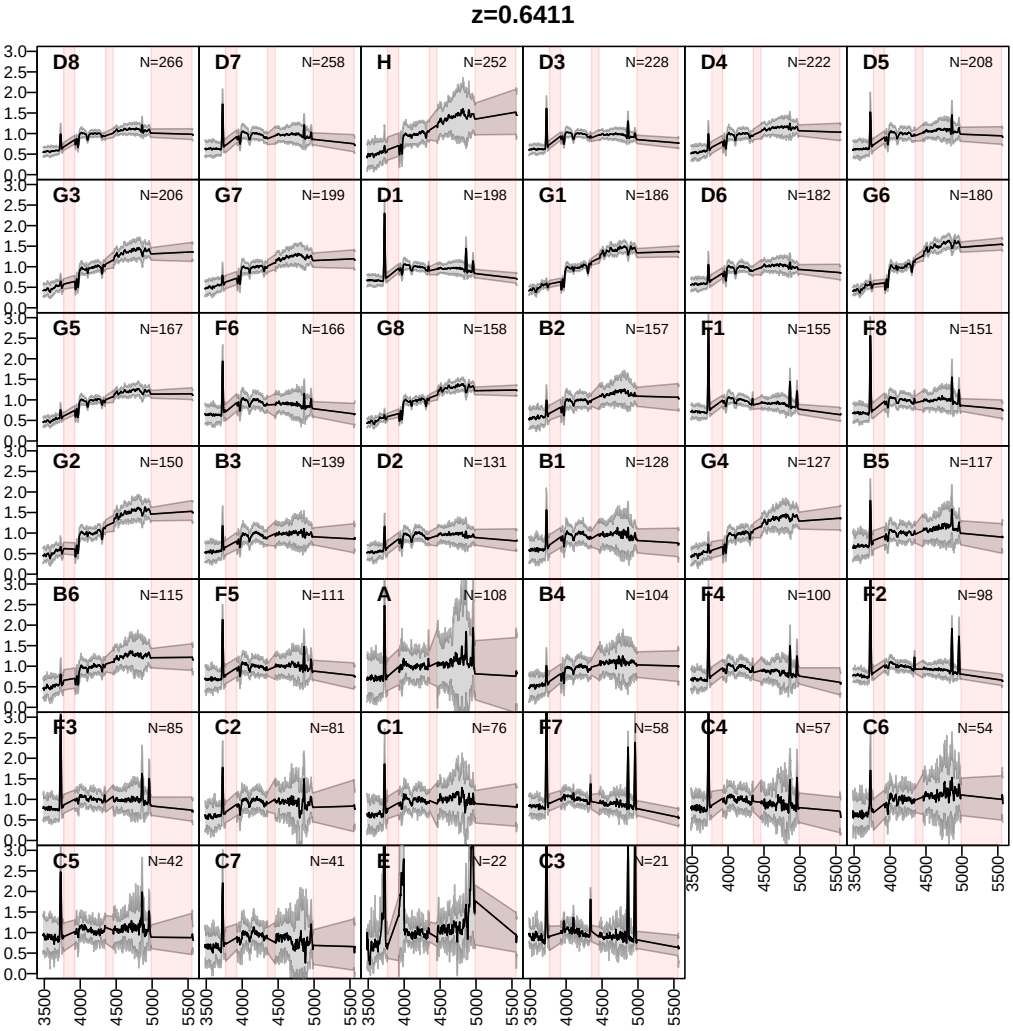


Fig. D.11.: Stacked spectra of the classes of bin 11 (see Fig. 4.10 for further information)

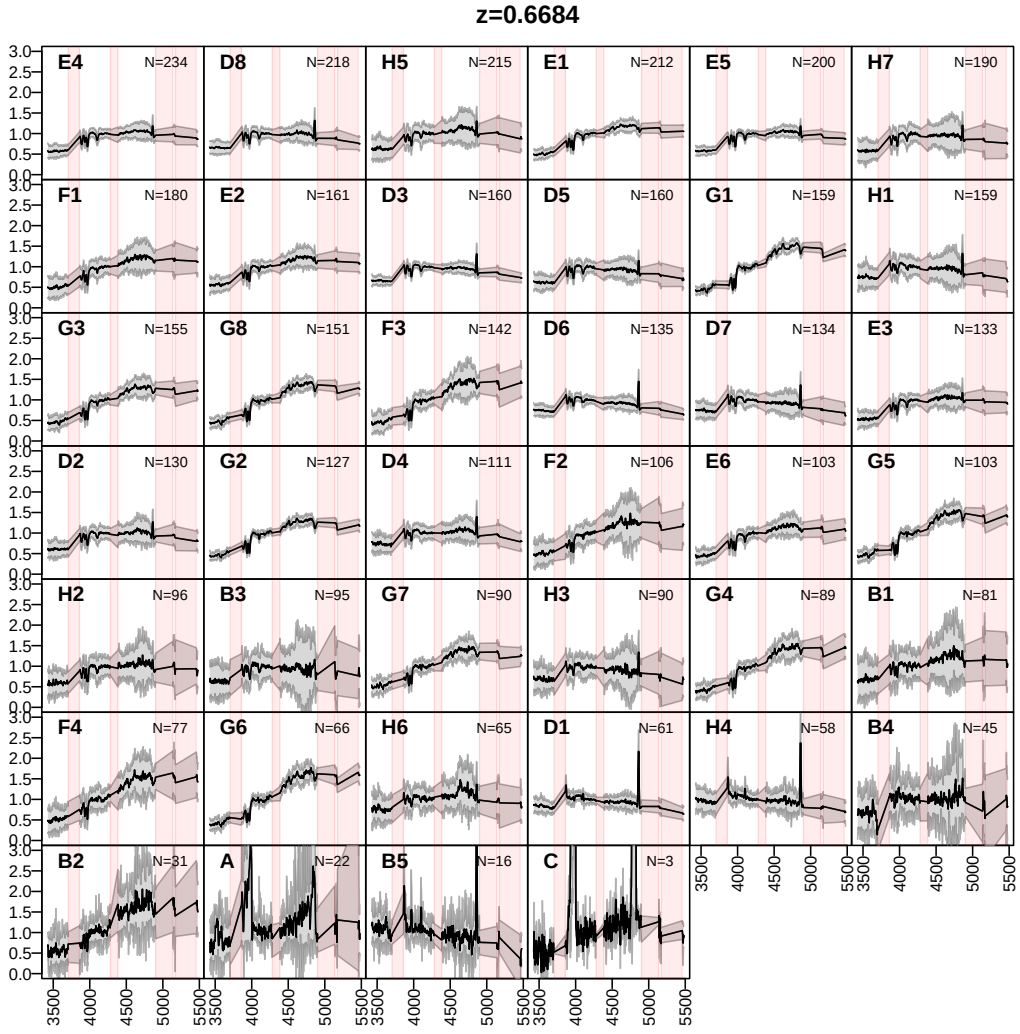


Fig. D.12.: Stacked spectra of the classes of bin 12 (see Fig. 4.10 for further information)

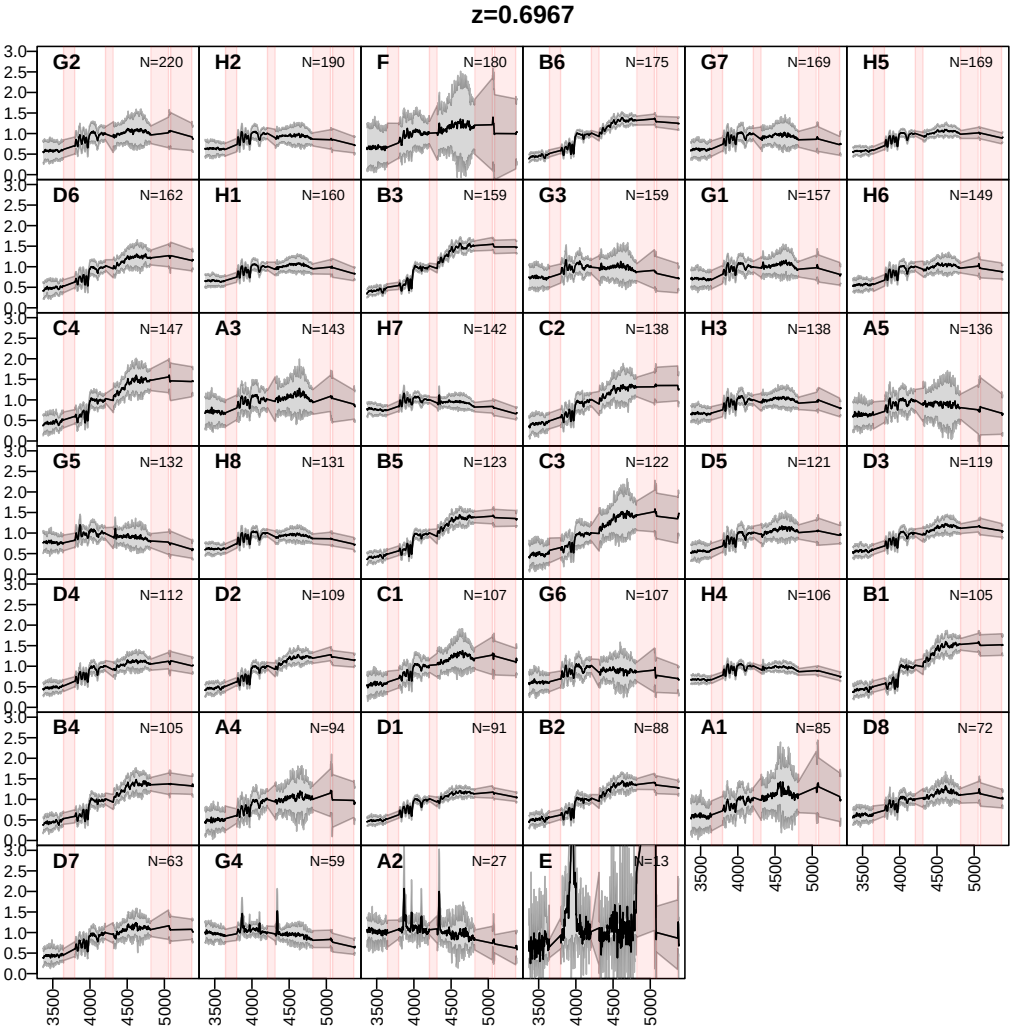


Fig. D.13.: Stacked spectra of the classes of bin 13 (see Fig. 4.10 for further information)

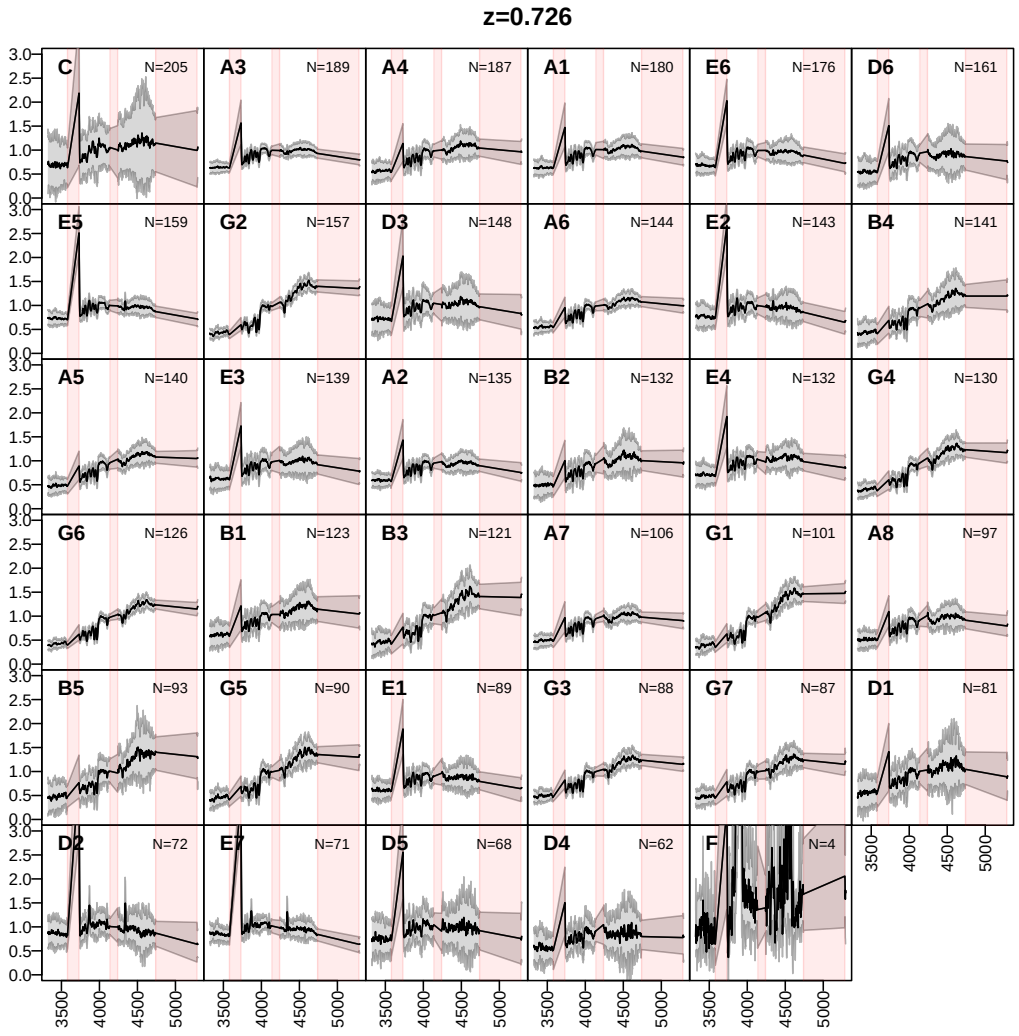


Fig. D.14.: Stacked spectra of the classes of bin 14 (see Fig. 4.10 for further information)

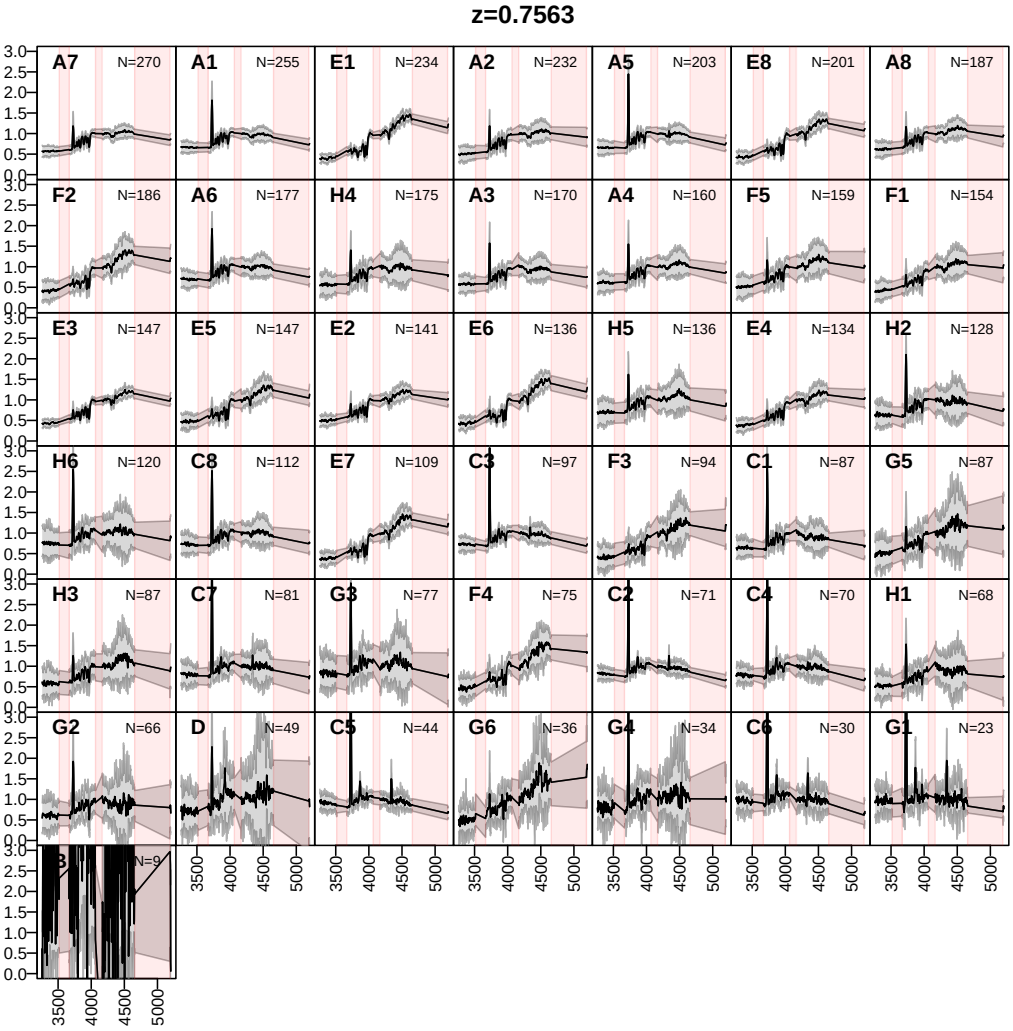


Fig. D.15.: Stacked spectra of the classes of bin 15 (see Fig. 4.10 for further information)

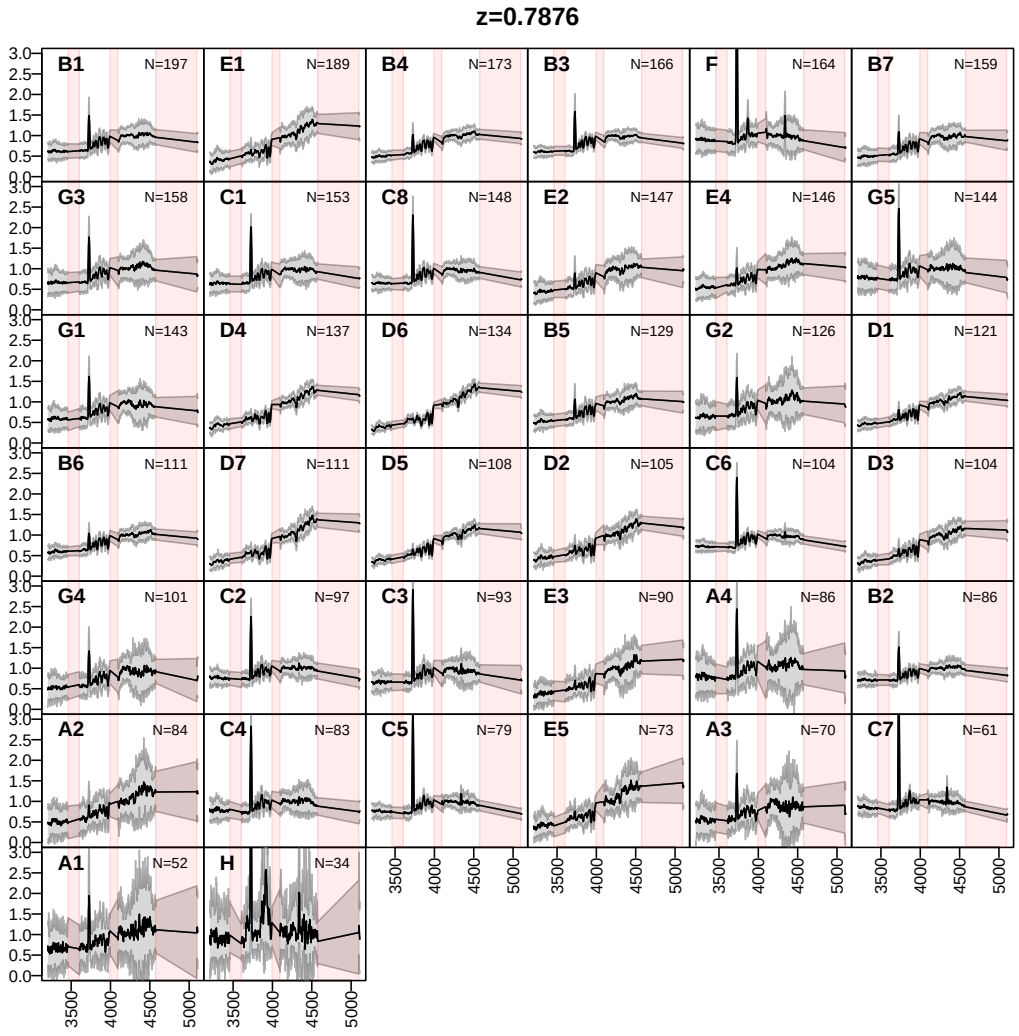


Fig. D.16.: Stacked spectra of the classes of bin 16 (see Fig. 4.10 for further information)

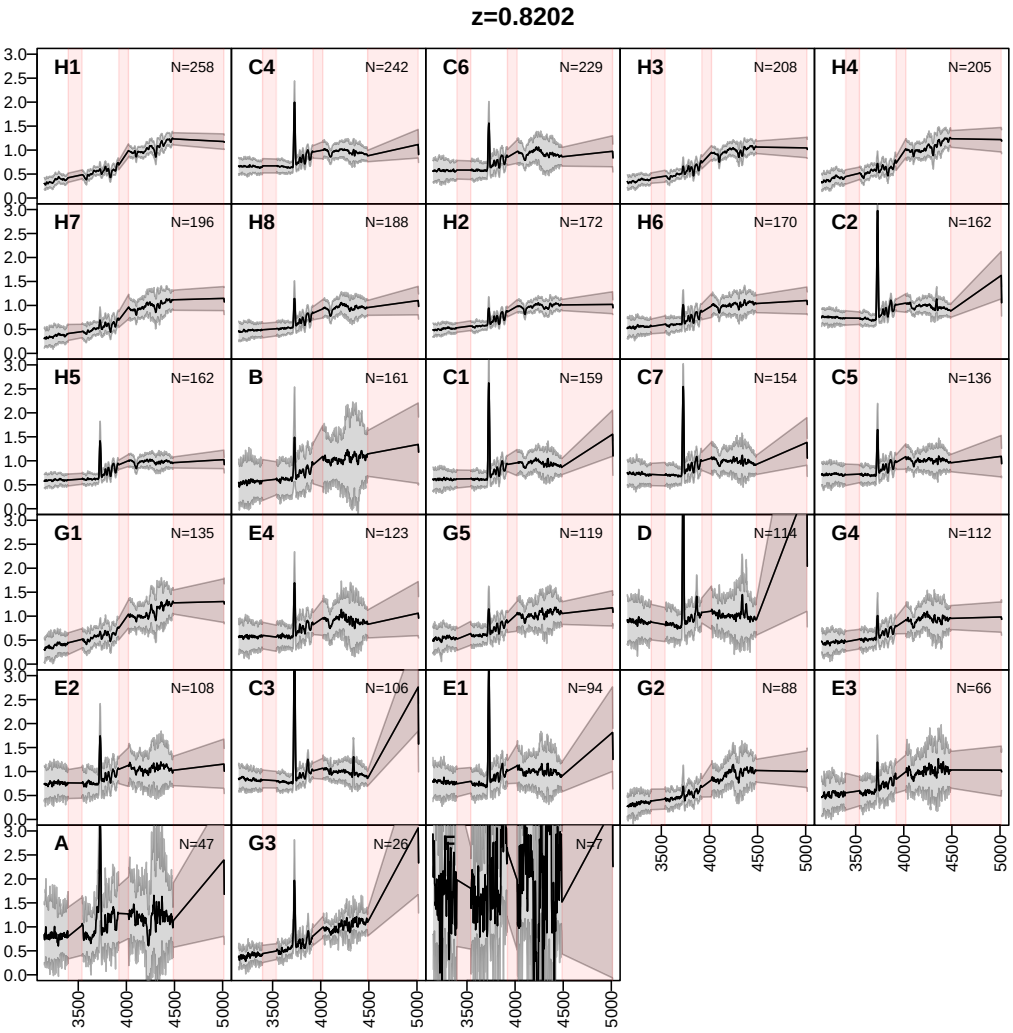


Fig. D.17.: Stacked spectra of the classes of bin 17 (see Fig. 4.10 for further information)

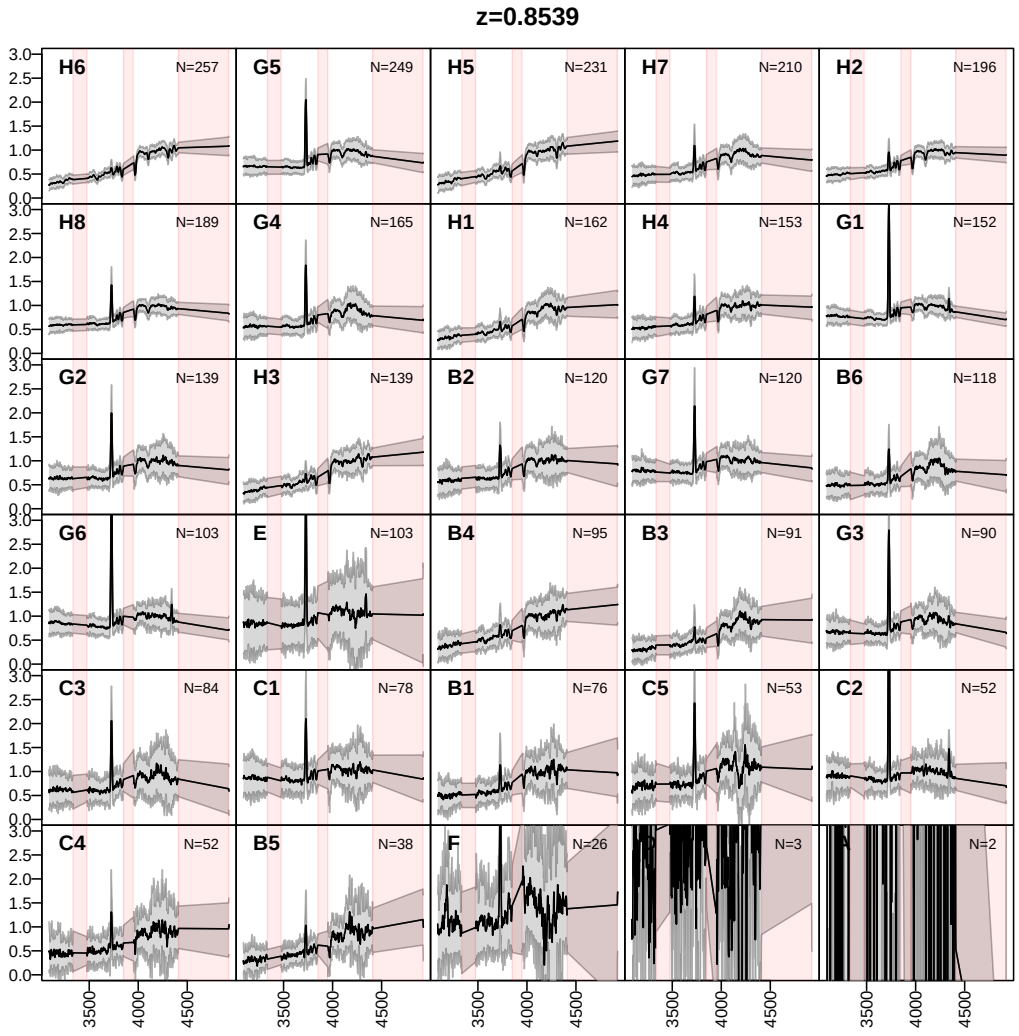


Fig. D.18.: Stacked spectra of the classes of bin 18 (see Fig. 4.10 for further information)

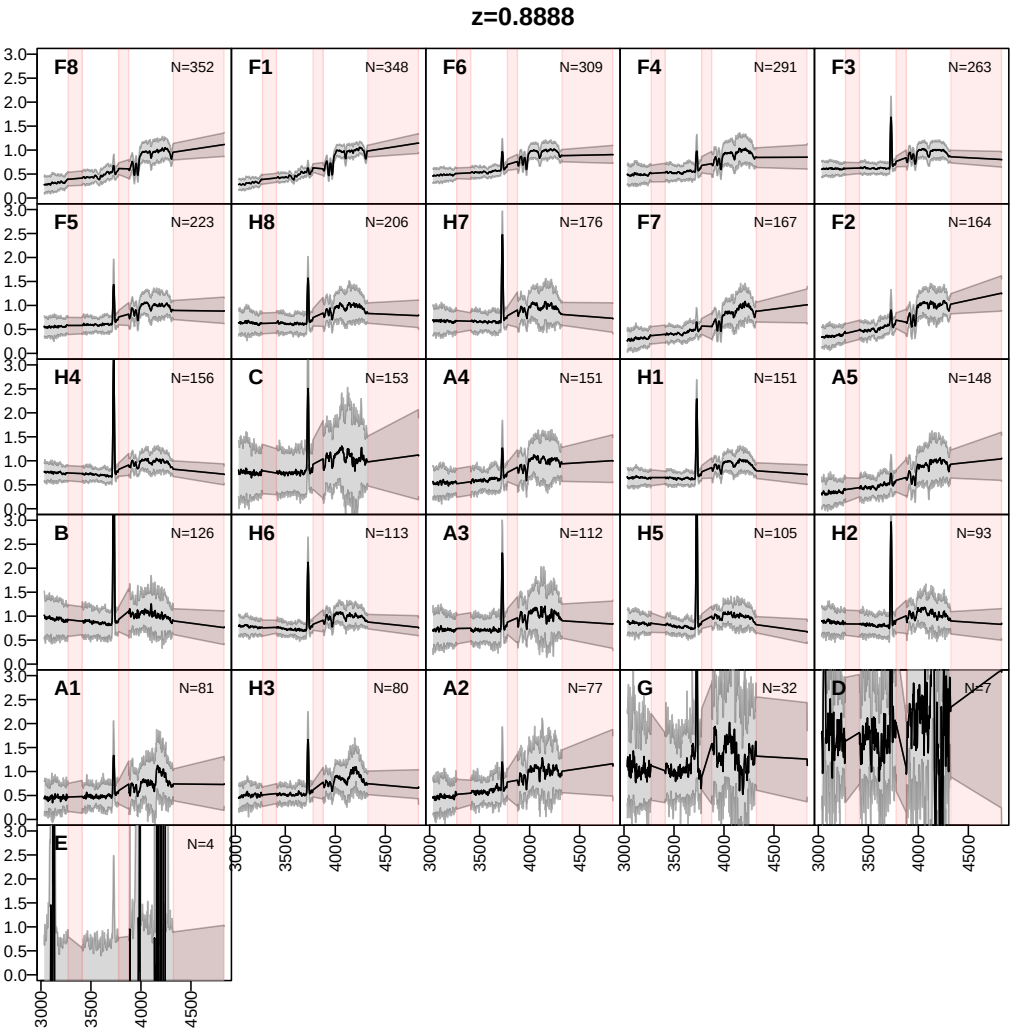


Fig. D.19.: Stacked spectra of the classes of bin 19 (see Fig. 4.10 for further information)

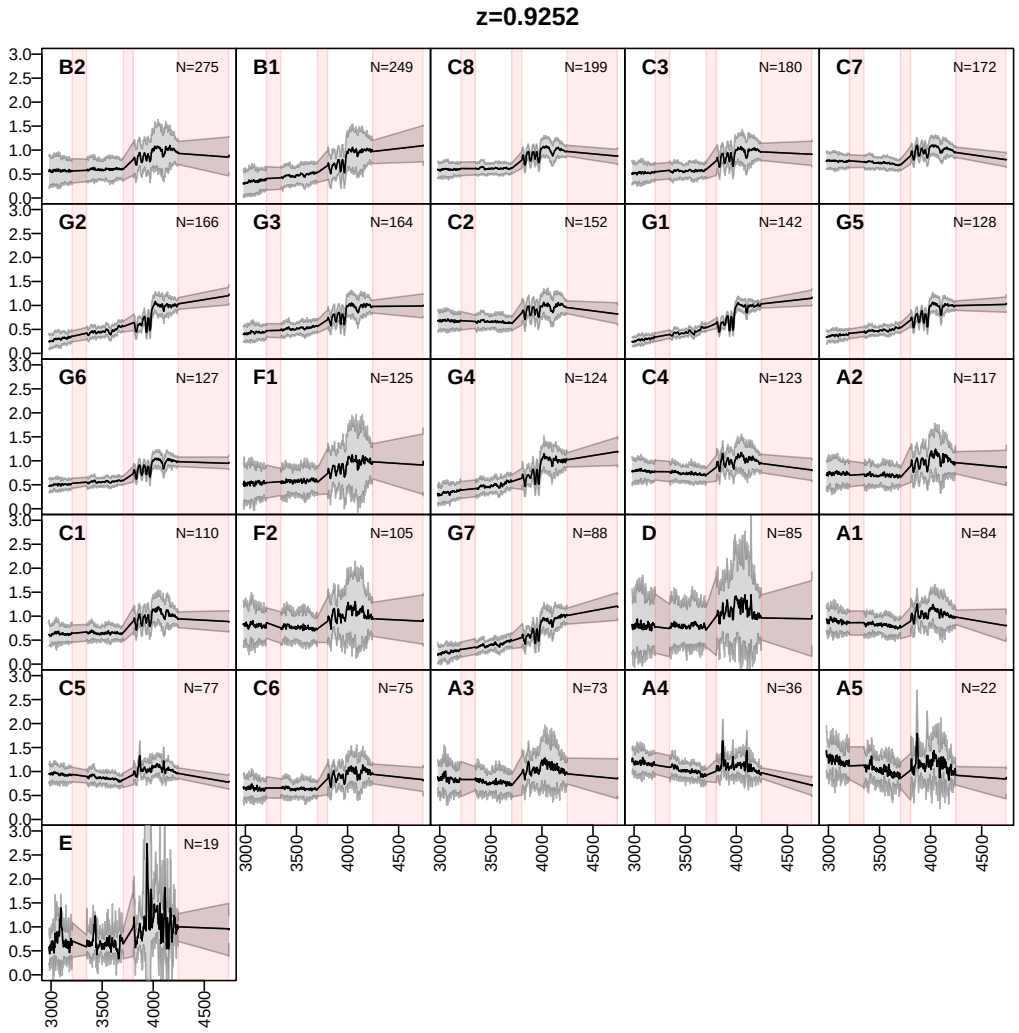


Fig. D.20.: Stacked spectra of the classes of bin 20 (see Fig. 4.10 for further information)

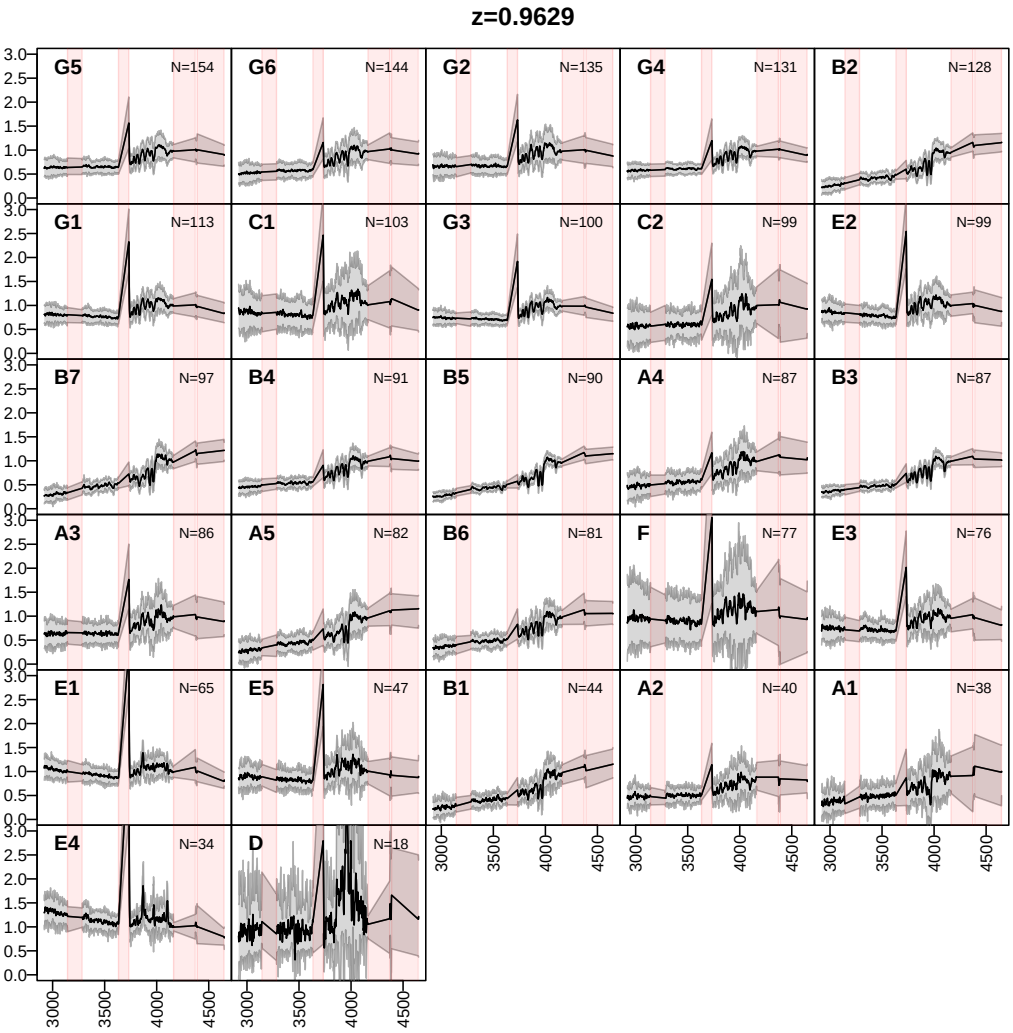


Fig. D.21.: Stacked spectra of the classes of bin 21 (see Fig. 4.10 for further information)

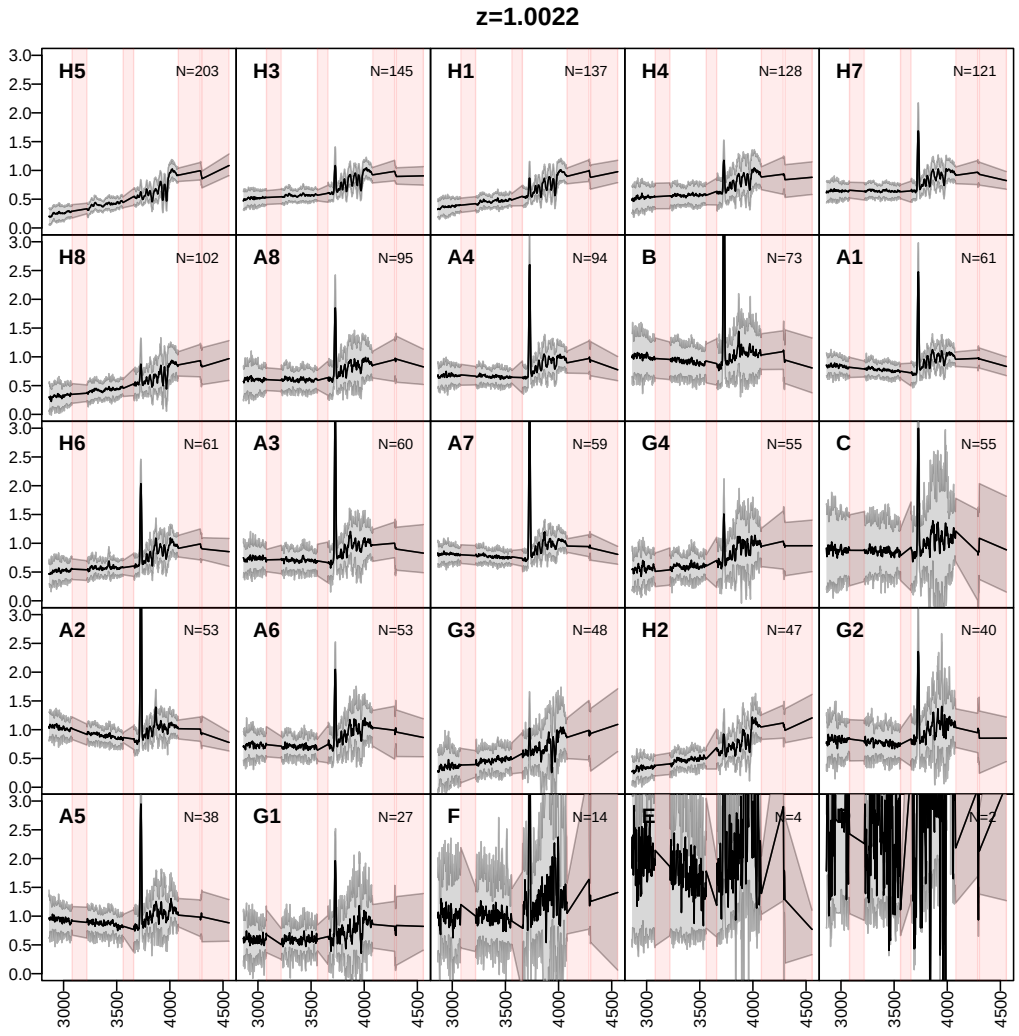


Fig. D.22.: Stacked spectra of the classes of bin 22 (see Fig. 4.10 for further information)

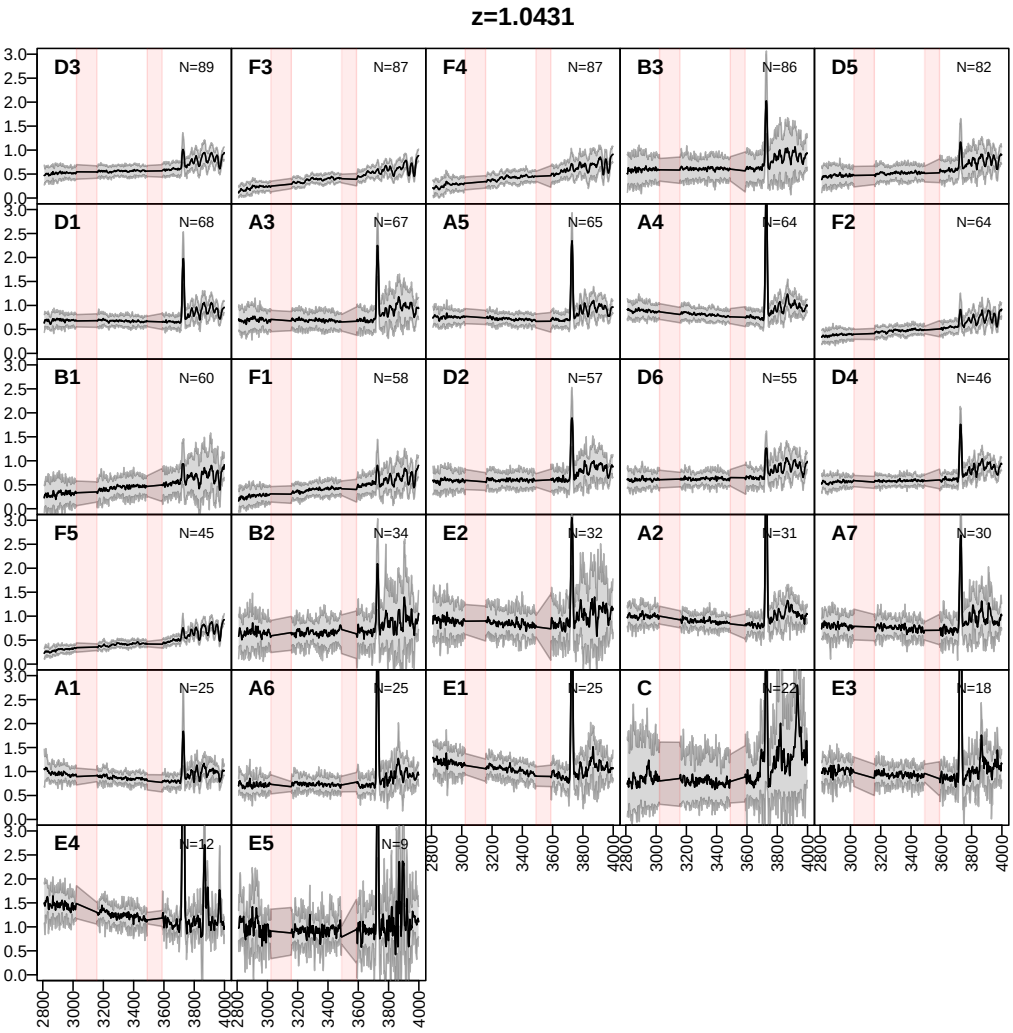


Fig. D.23.: Stacked spectra of the classes of bin 23 (see Fig. 4.10 for further information)

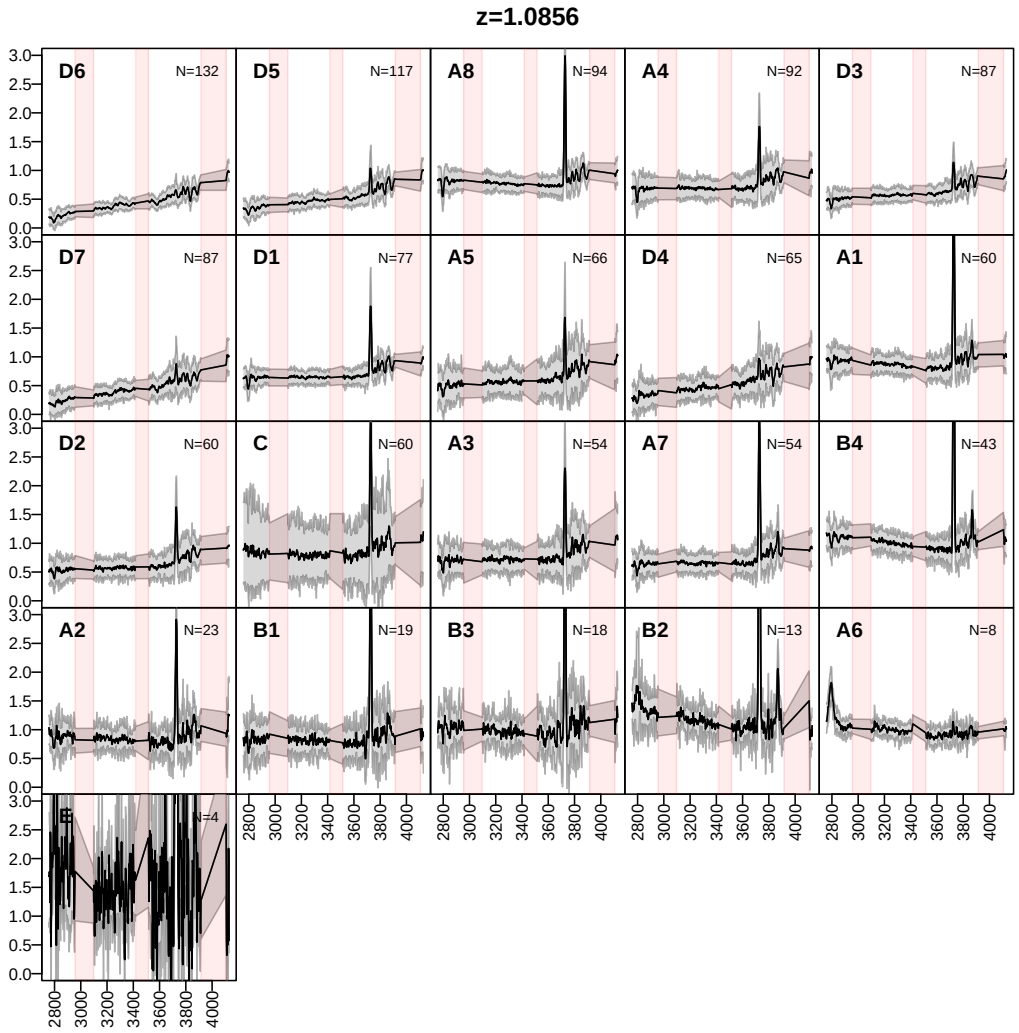


Fig. D.24.: Stacked spectra of the classes of bin 24 (see Fig. 4.10 for further information)

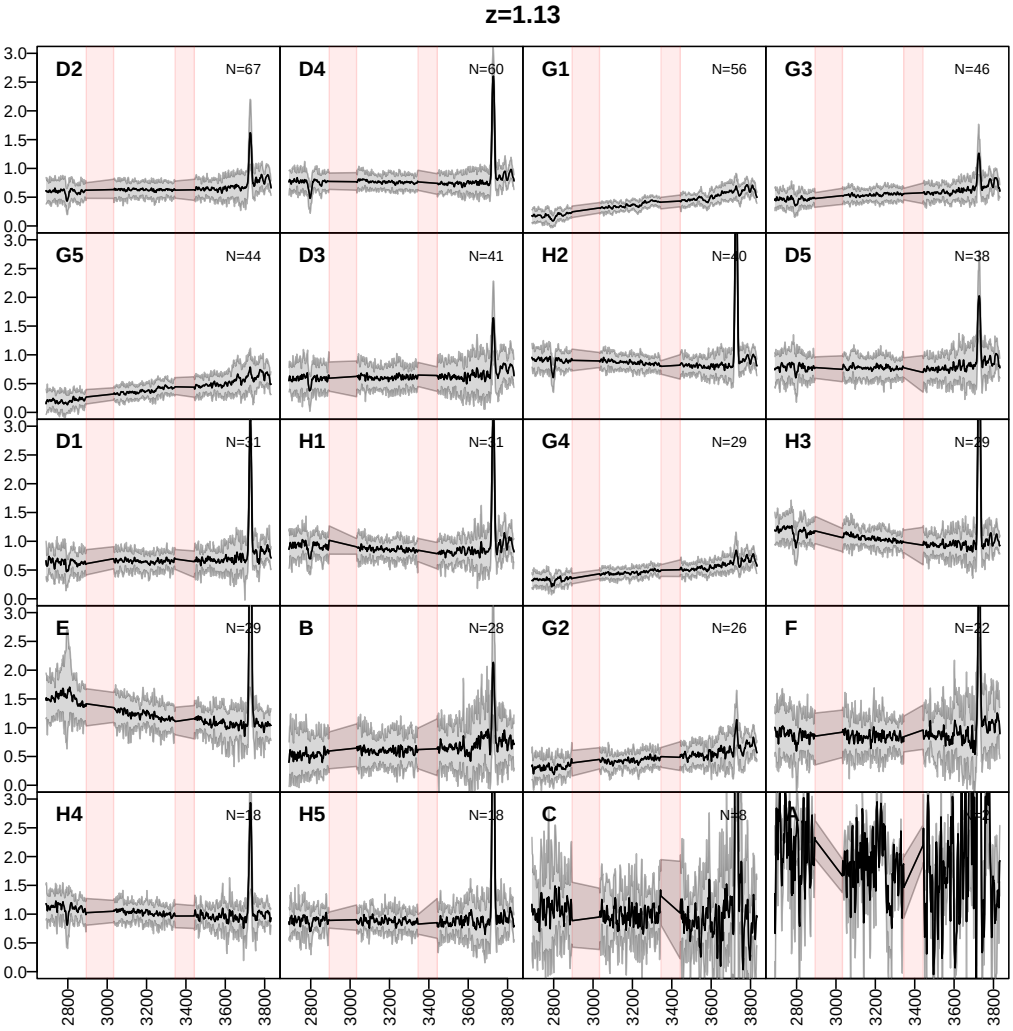


Fig. D.25.: Stacked spectra of the classes of bin 25 (see Fig. 4.10 for further information)

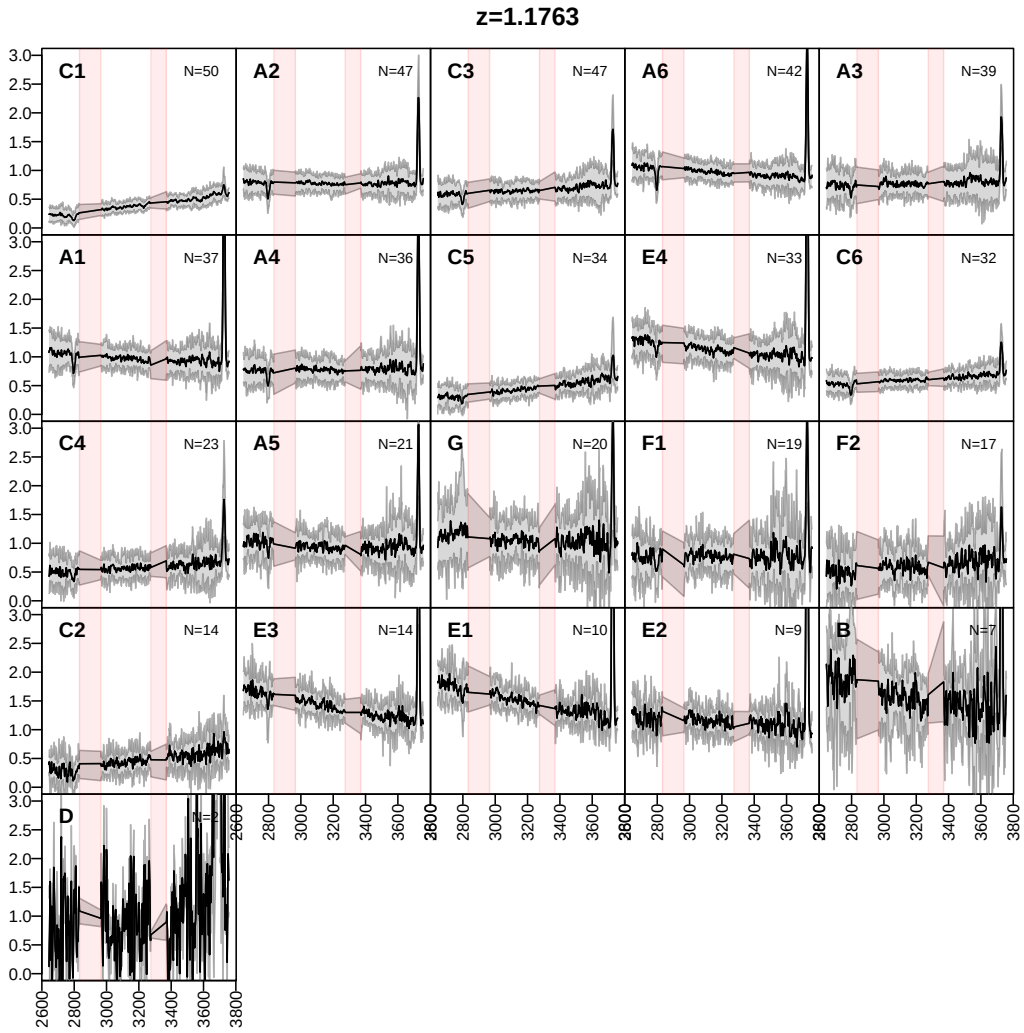


Fig. D.26.: Stacked spectra of the classes of bin 26 (see Fig. 4.10 for further information)

E

VIPERS classification: MEx diagrams

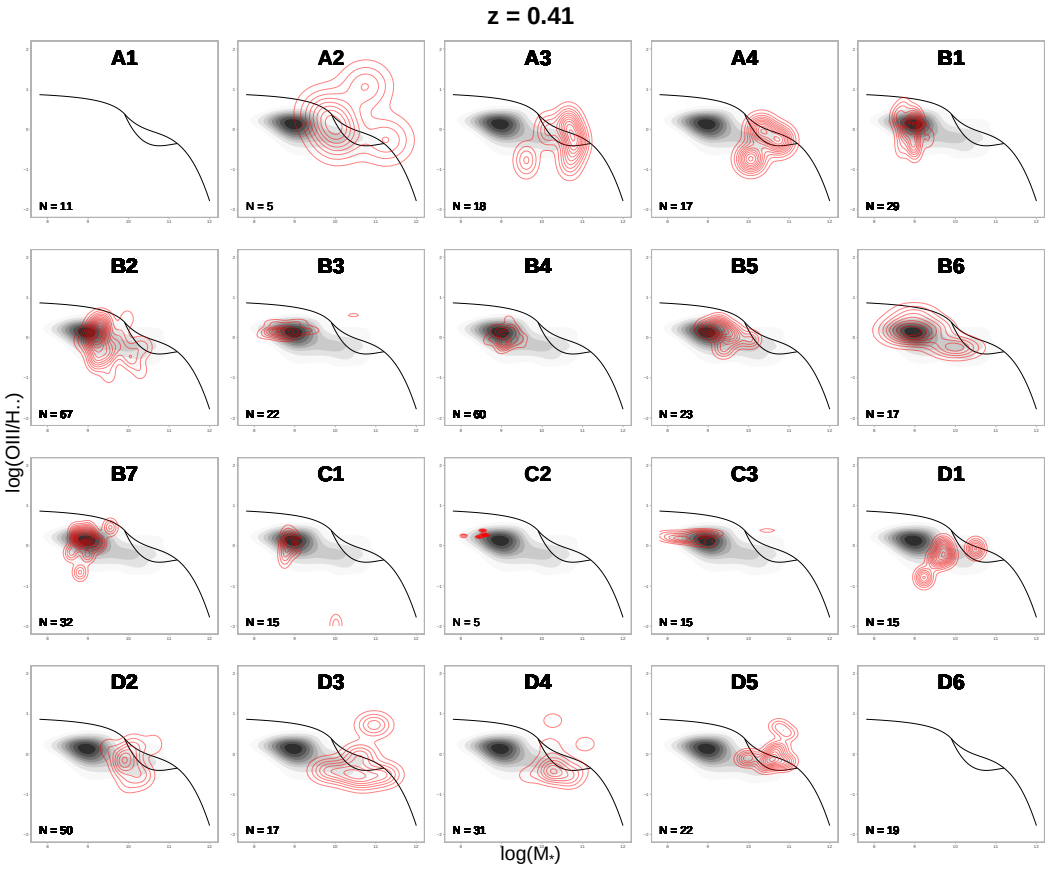


Fig. E.1.: MEx diagram of the classes of bin 1 (see Fig. 4.11 for further information)

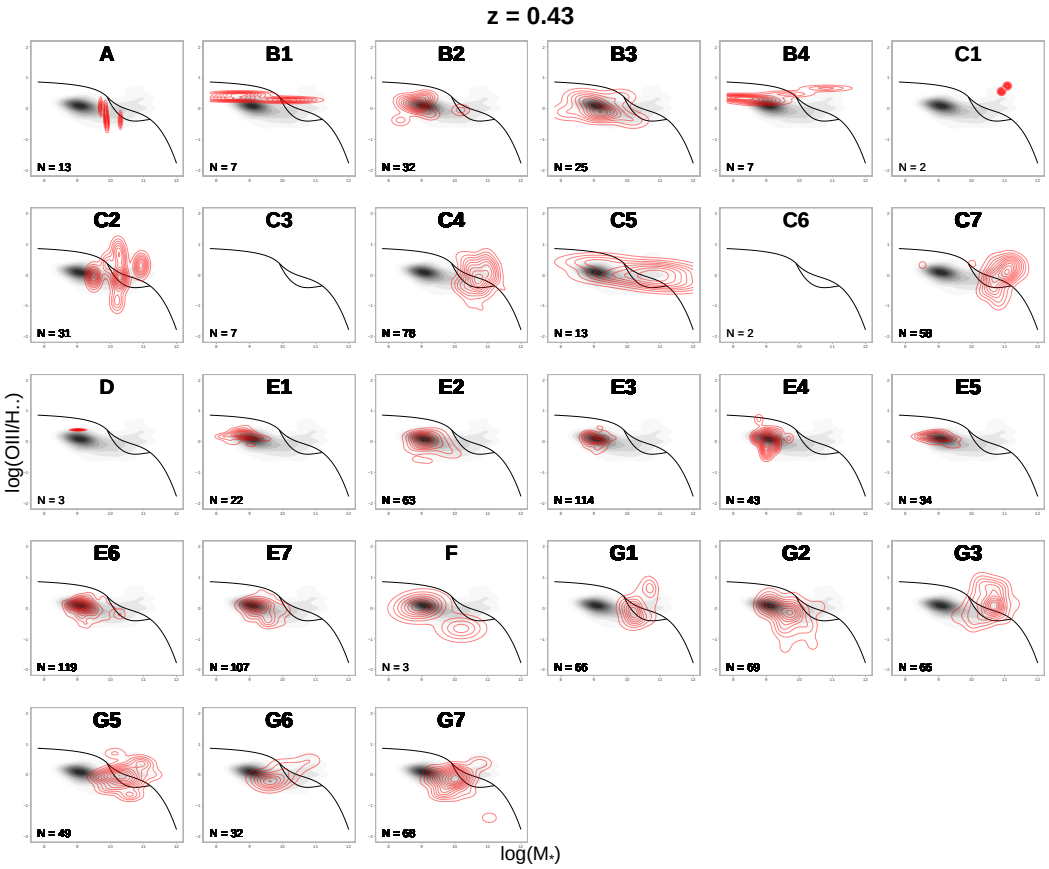


Fig. E.2.: MEx diagram of the classes of bin 2 (see Fig. 4.11 for further information)

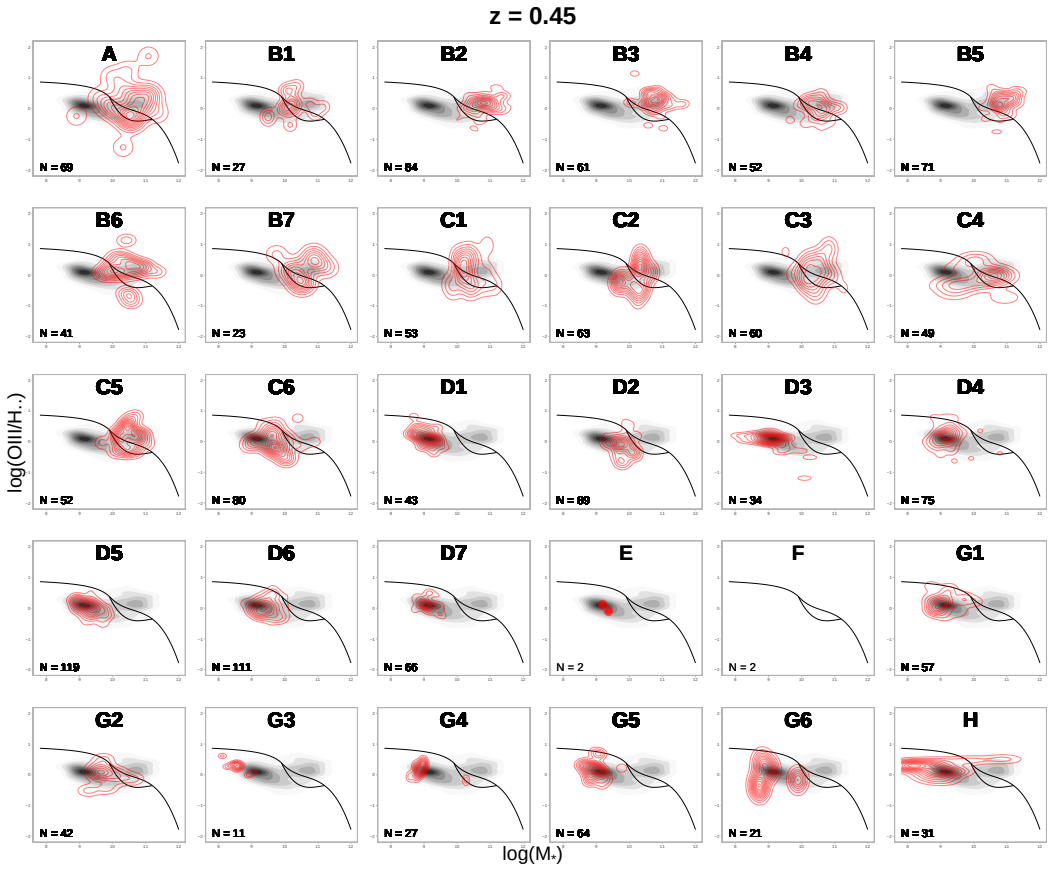


Fig. E.3.: MEx diagram of the classes of bin 3 (see Fig. 4.11 for further information)

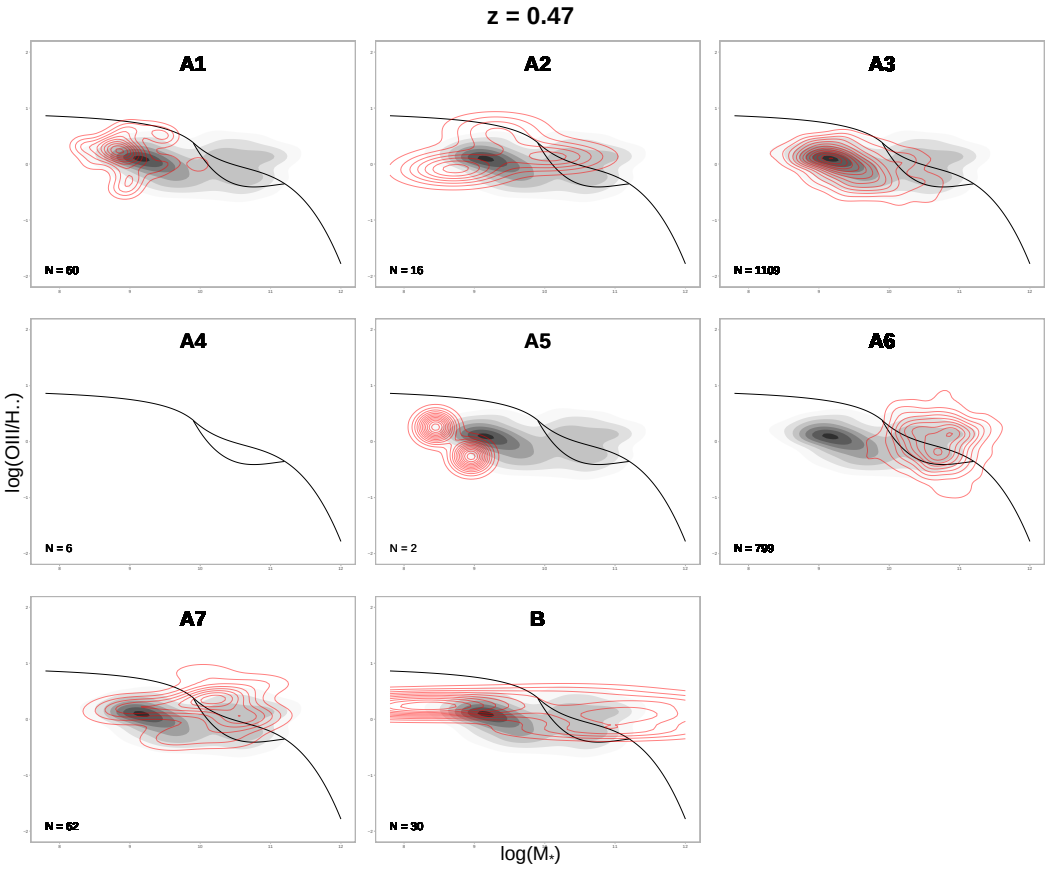


Fig. E.4.: MEx diagram of the classes of bin 4 (see Fig. 4.11 for further information)

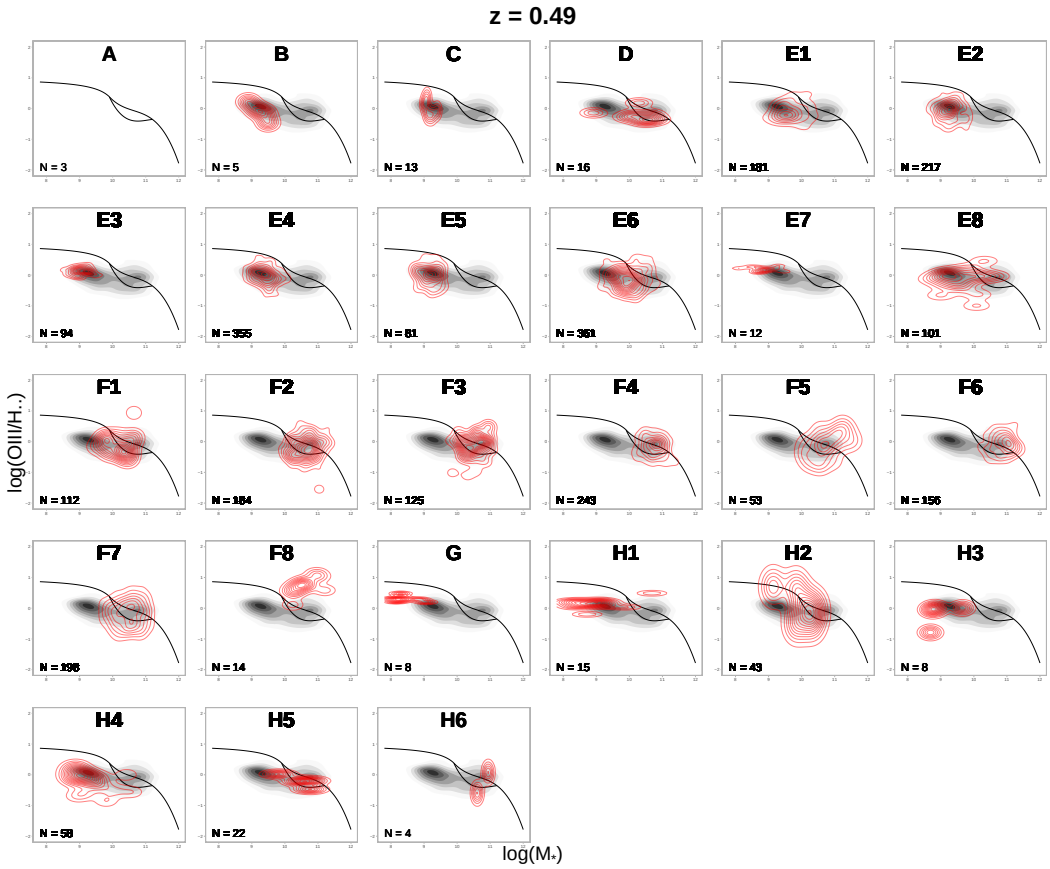


Fig. E.5.: MEx diagram of the classes of bin 5 (see Fig. 4.11 for further information)

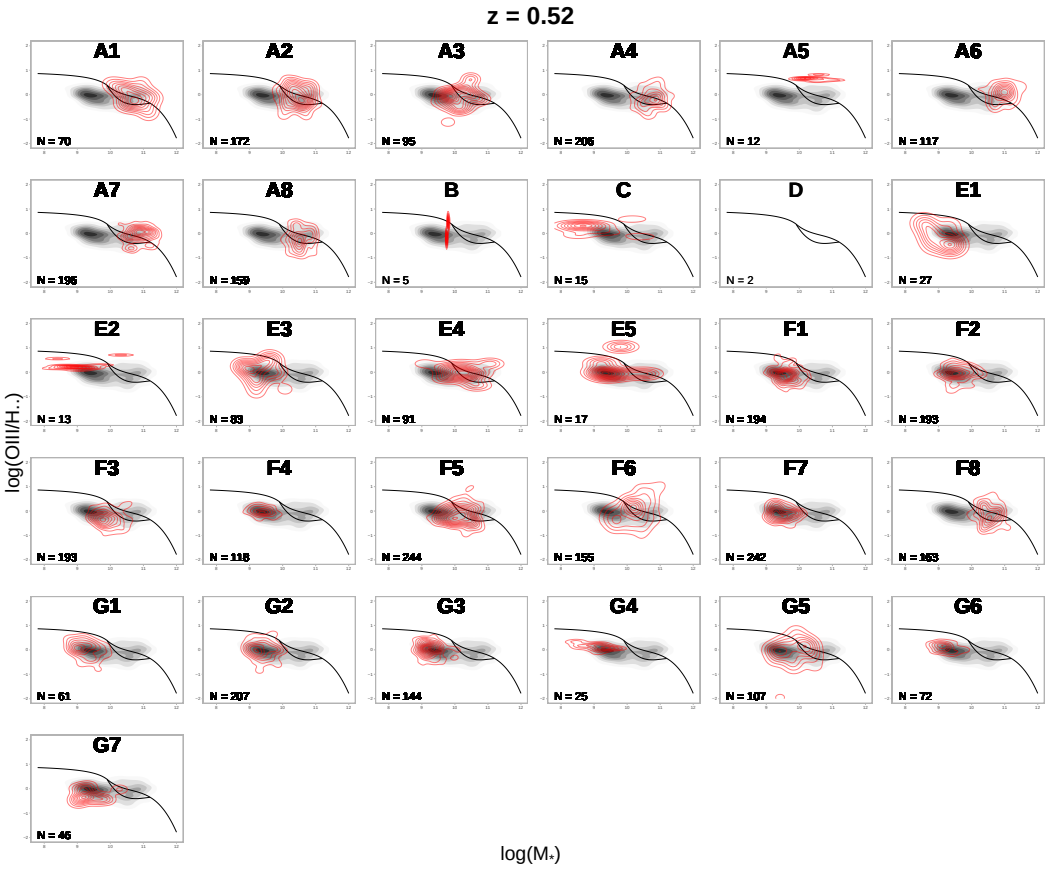


Fig. E.6.: MEx diagram of the classes of bin 6 (see Fig. 4.11 for further information)

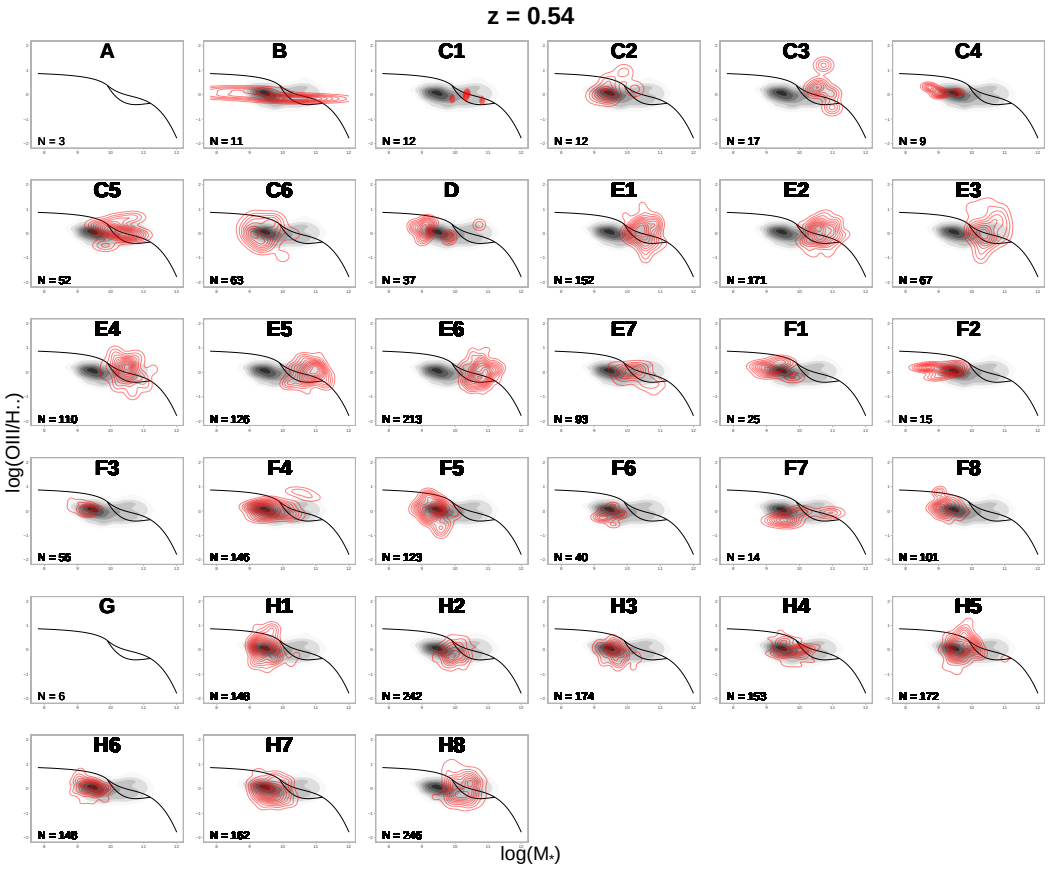


Fig. E.7.: MEx diagram of the classes of bin 7 (see Fig. 4.11 for further information)

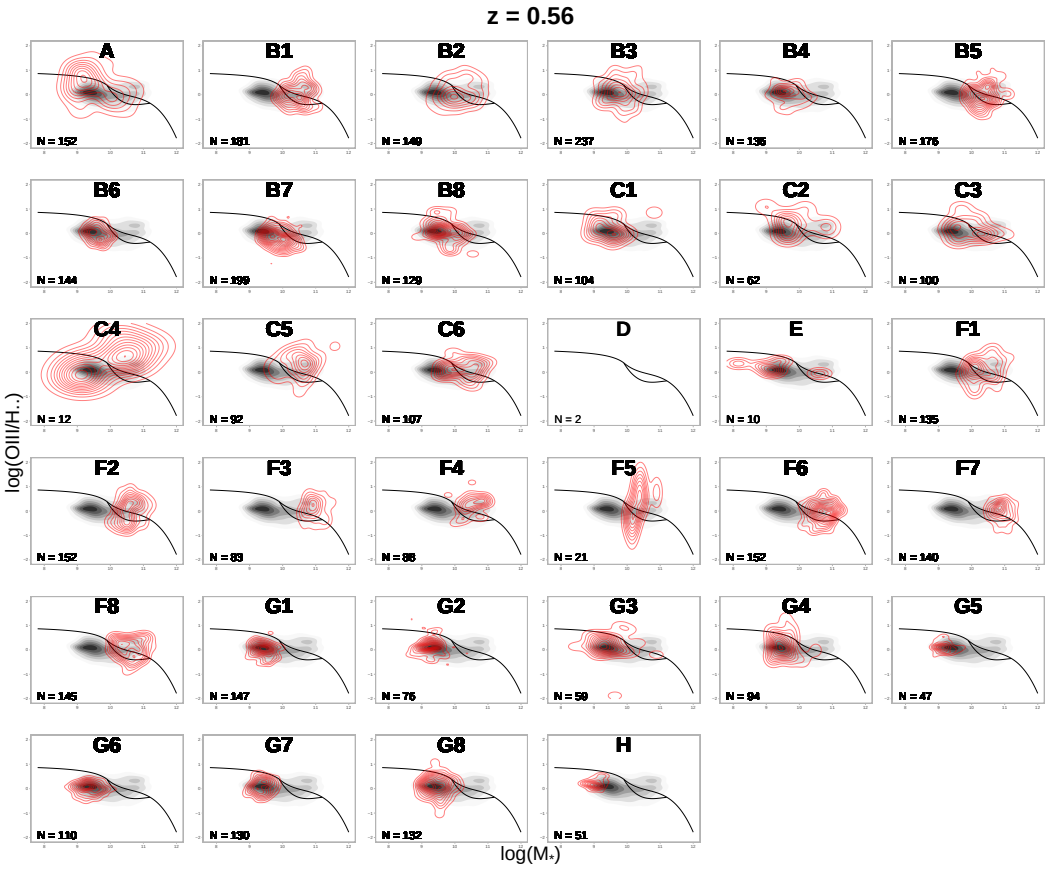


Fig. E.8.: MEx diagram of the classes of bin 8 (see Fig. 4.11 for further information)

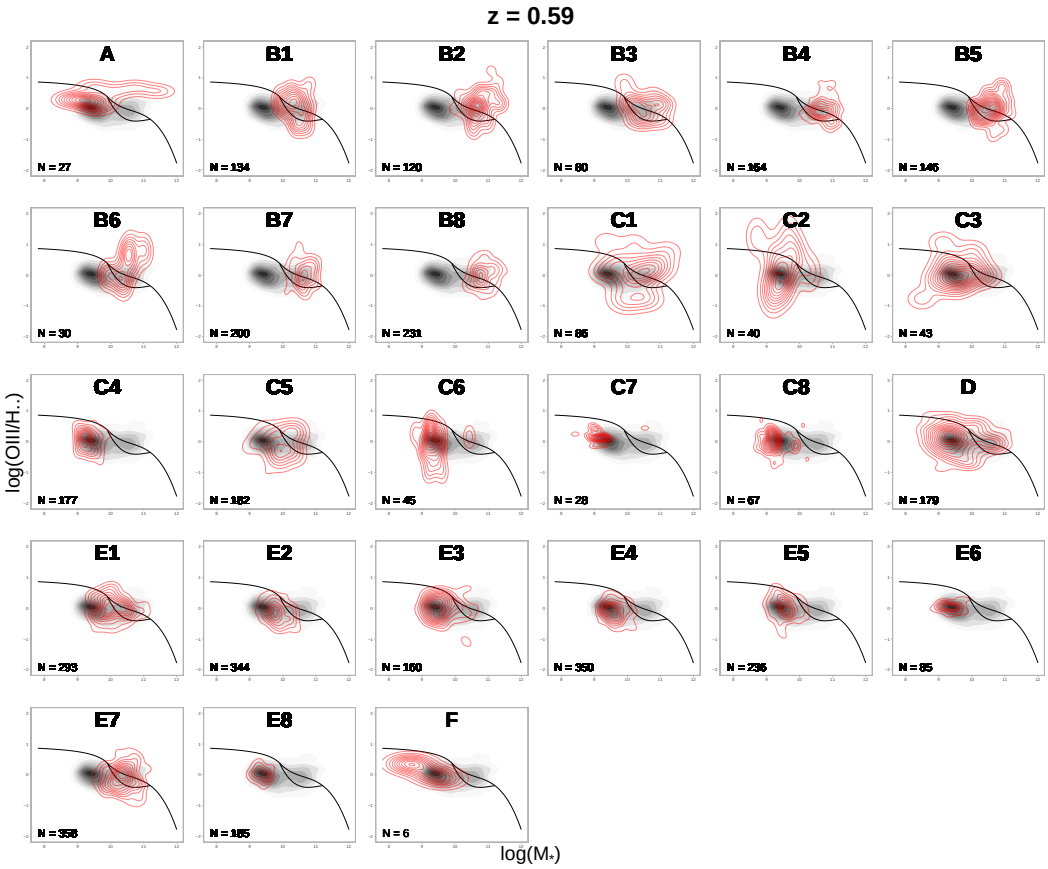


Fig. E.9.: MEx diagram of the classes of bin 9 (see Fig. 4.11 for further information)

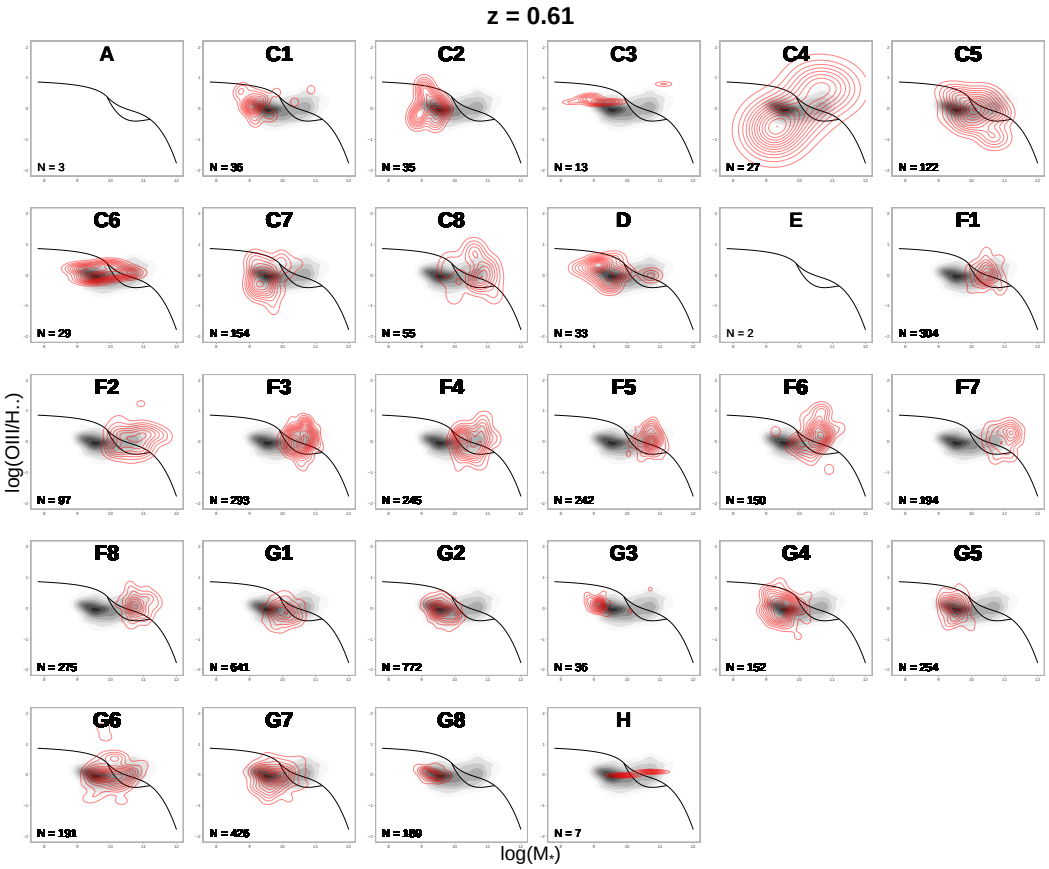


Fig. E.10.: MEx diagram of the classes of bin 10 (see Fig. 4.11 for further information)

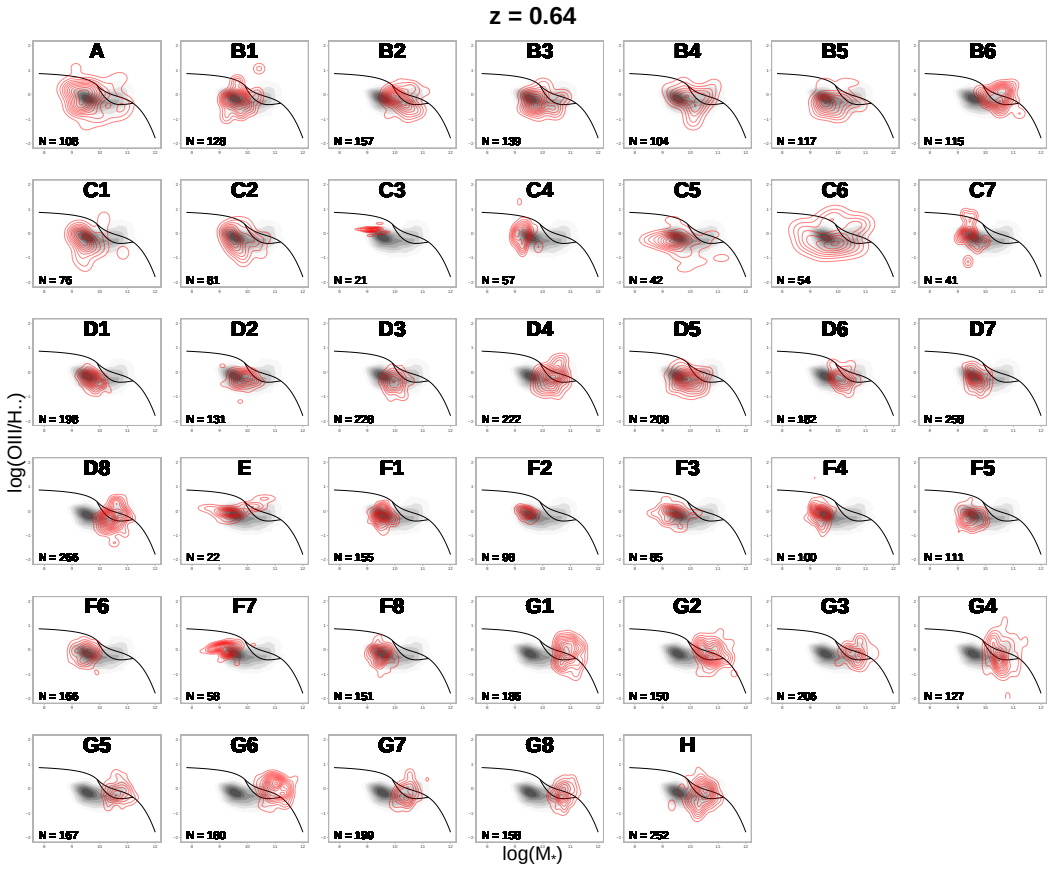


Fig. E.11.: MEx diagram of the classes of bin 11 (see Fig. 4.11 for further information)

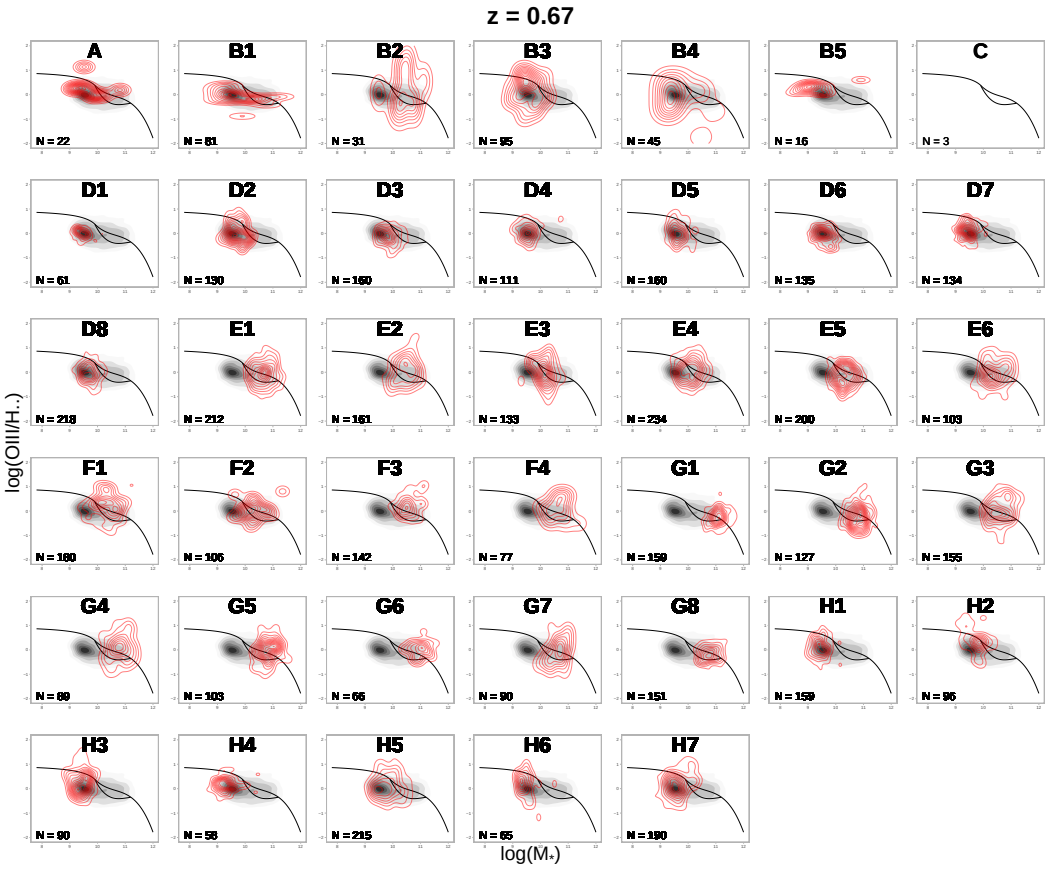


Fig. E.12.: MEx diagram of the classes of bin 12 (see Fig. 4.11 for further information)

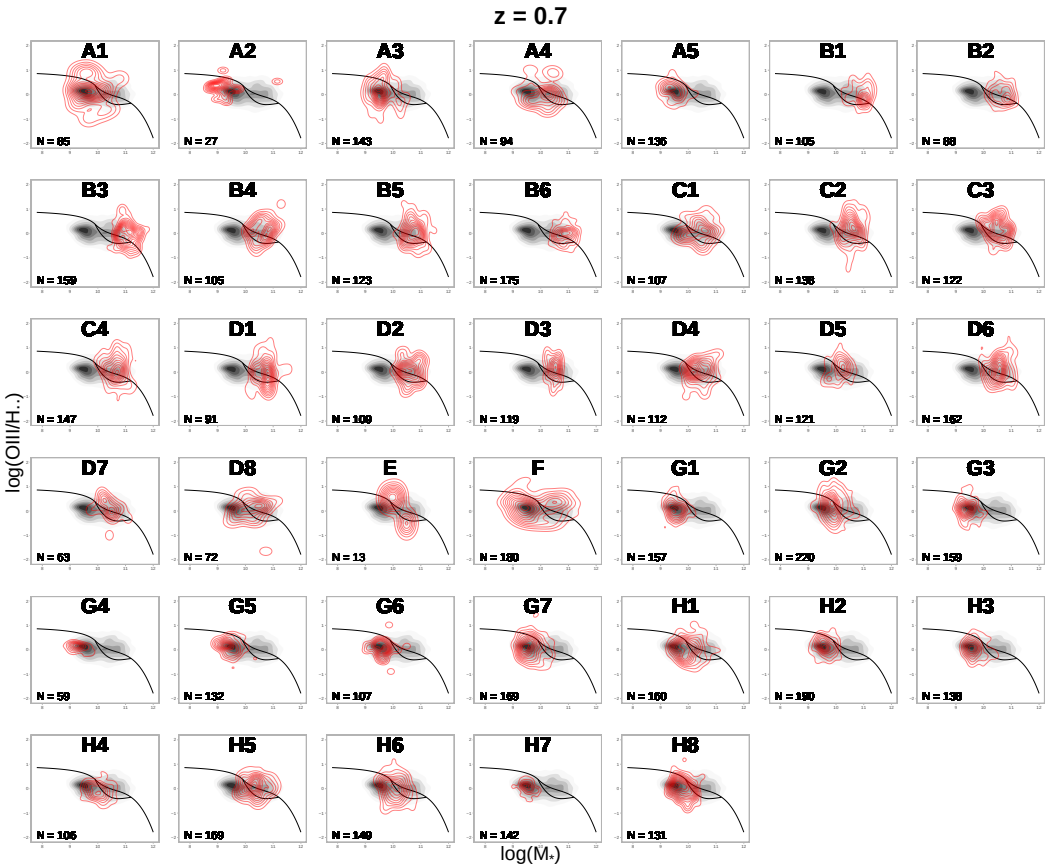


Fig. E.13.: MEx diagram of the classes of bin 13 (see Fig. 4.11 for further information)

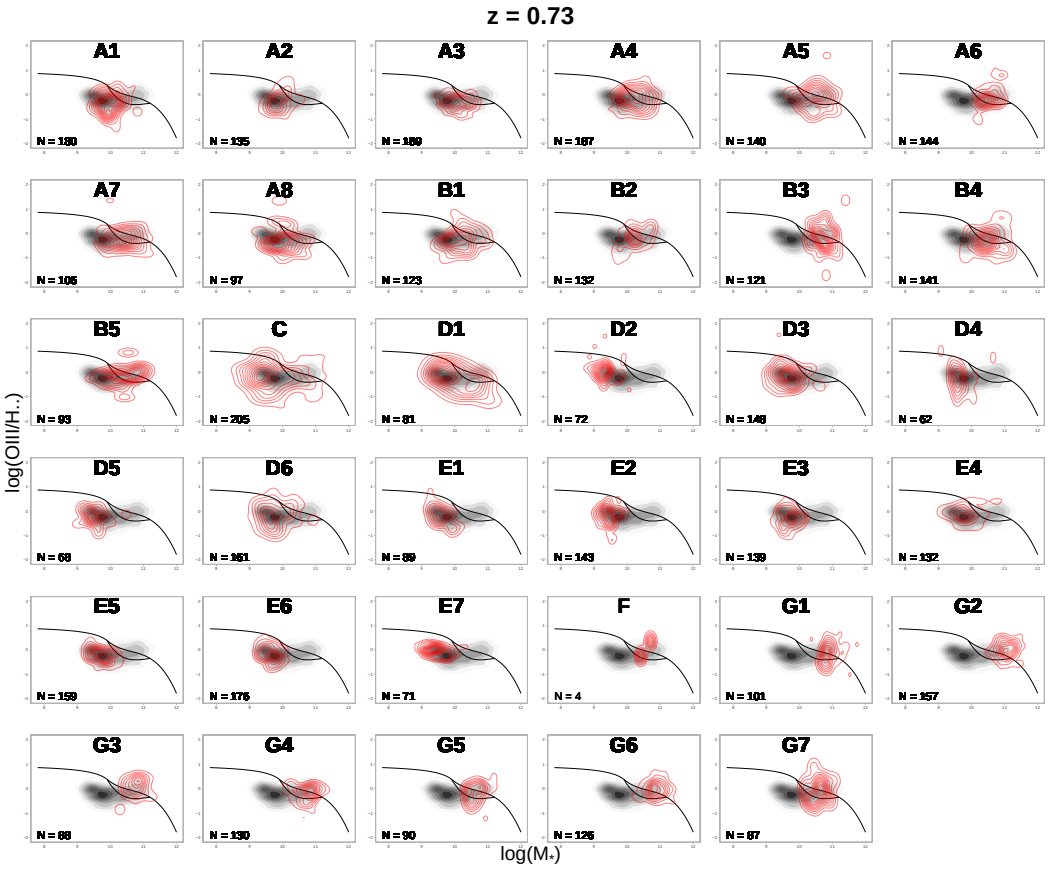


Fig. E.14.: MEx diagram of the classes of bin 14 (see Fig. 4.11 for further information)

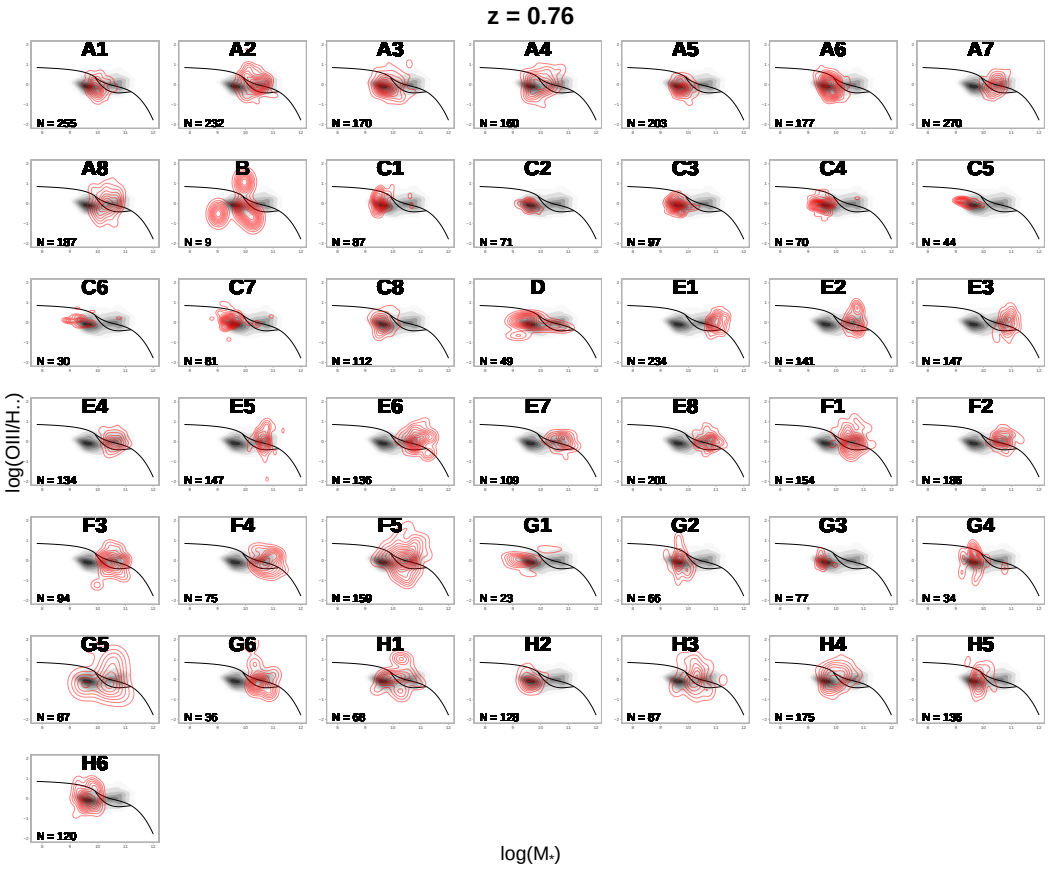


Fig. E.15.: MEx diagram of the classes of bin 15 (see Fig. 4.11 for further information)

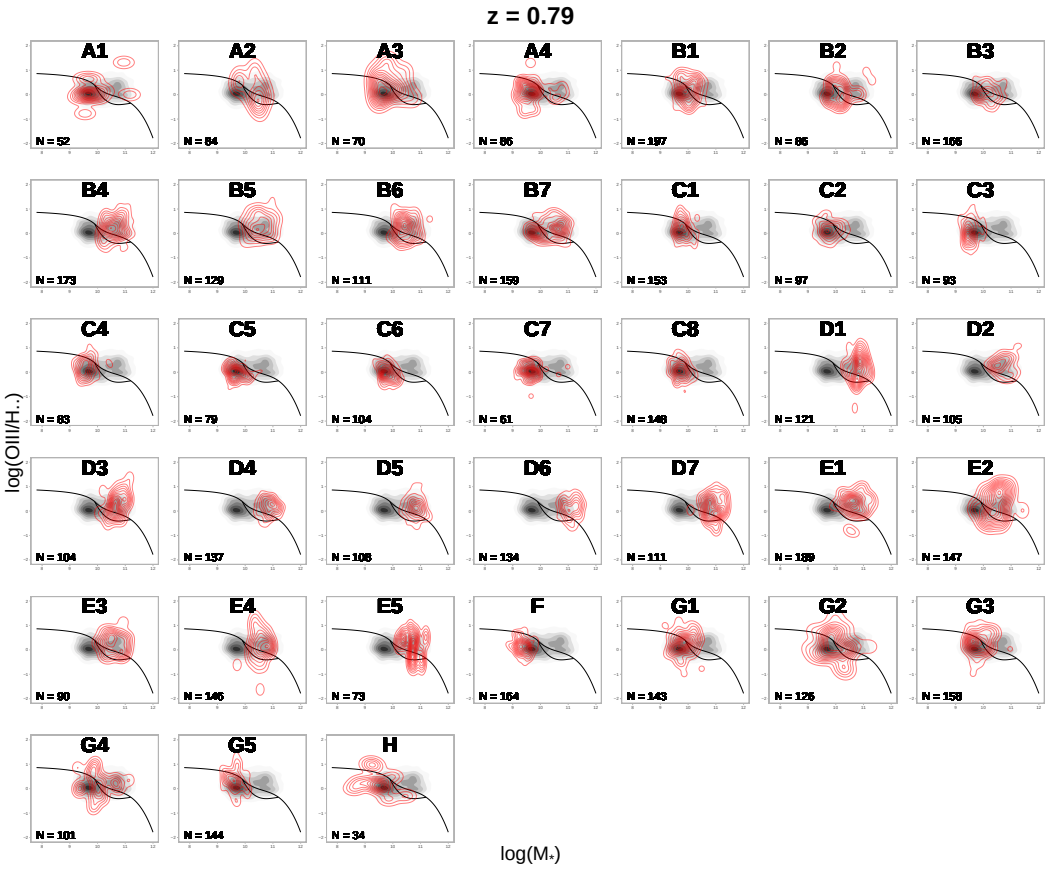


Fig. E.16.: MEx diagram of the classes of bin 16 (see Fig. 4.11 for further information)

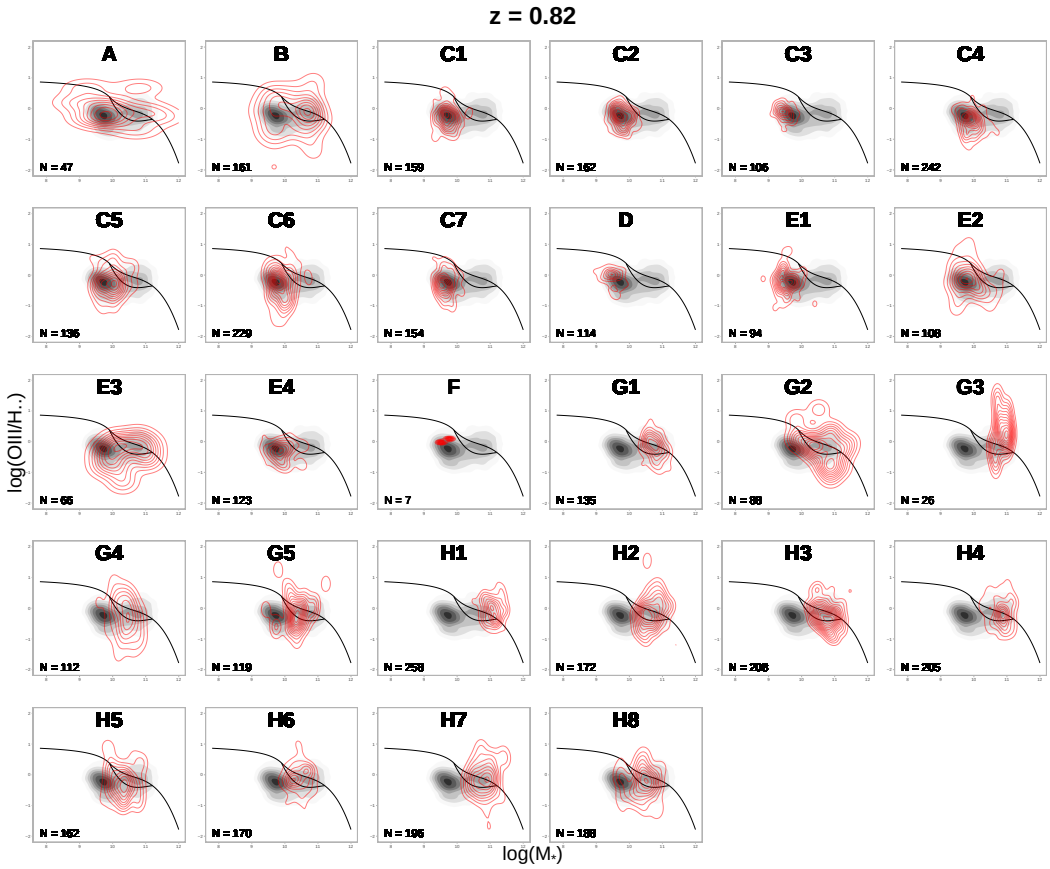


Fig. E.17.: MEx diagram of the classes of bin 17 (see Fig. 4.11 for further information)

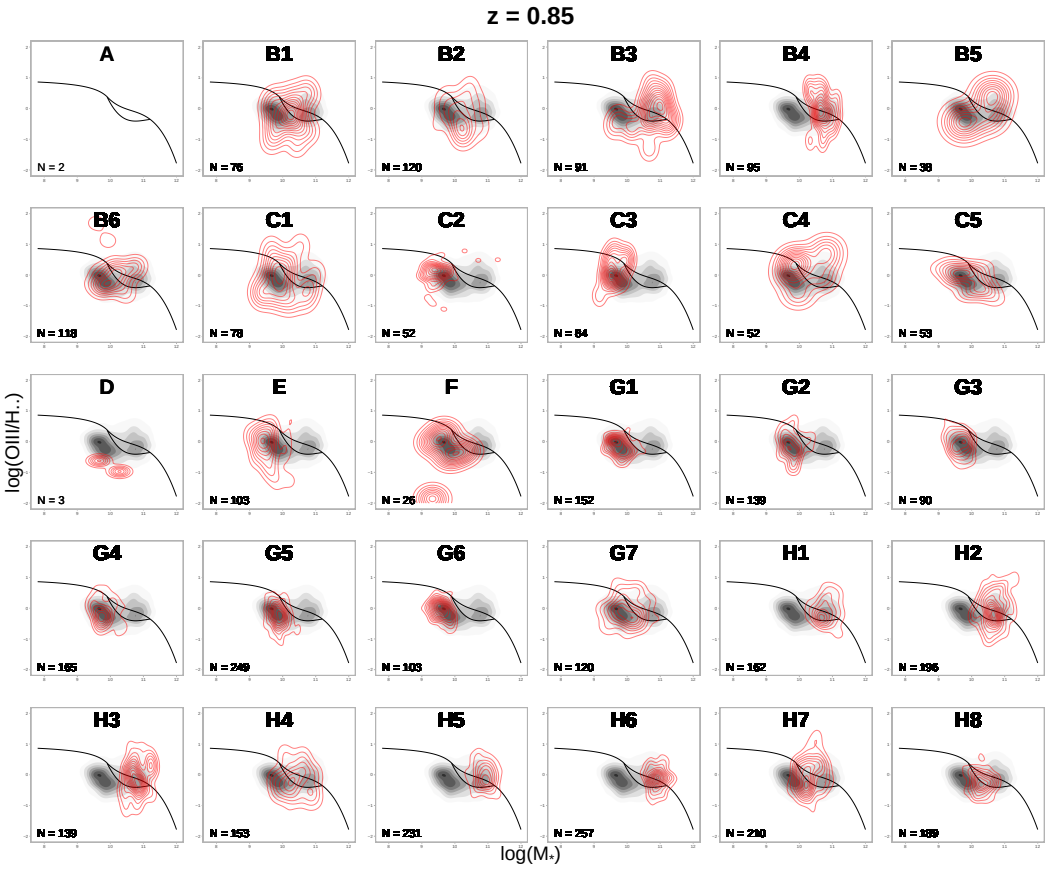


Fig. E.18.: MEx diagram of the classes of bin 18 (see Fig. 4.11 for further information)

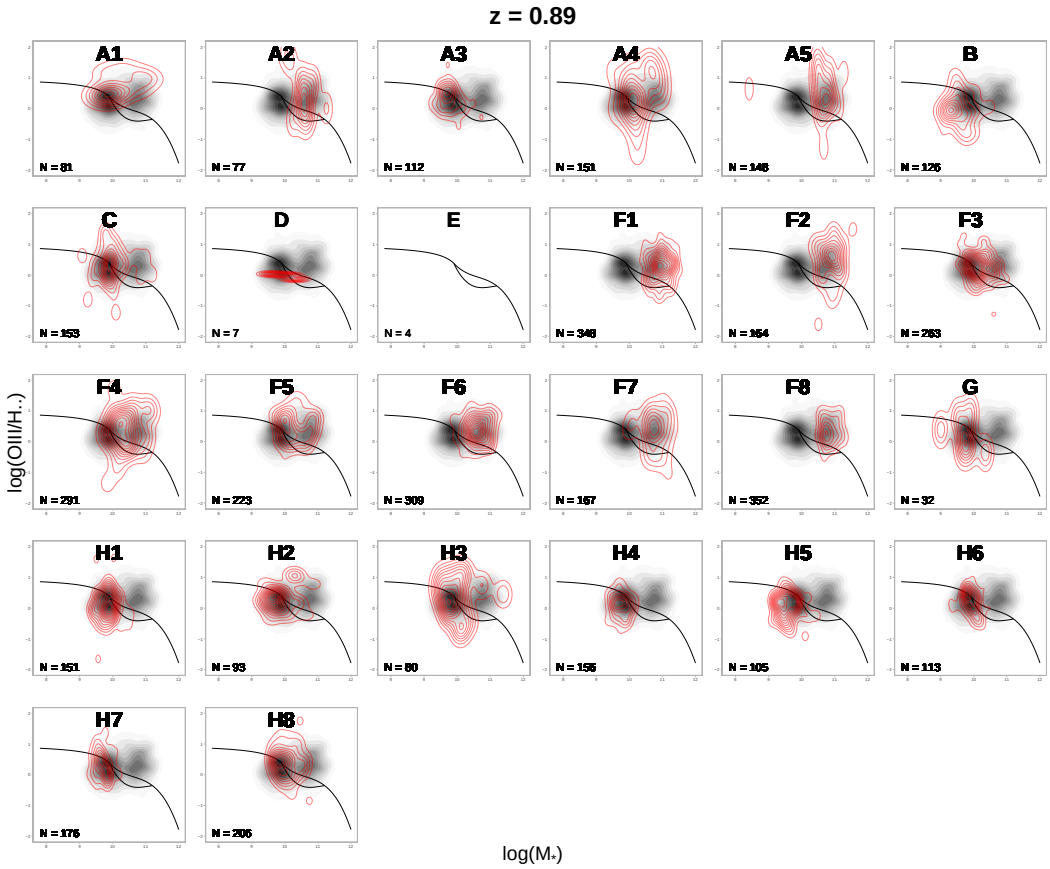


Fig. E.19.: MEx diagram of the classes of bin 19 (see Fig. 4.11 for further information)