Aix*Marseille université
Socialement engagée

NNT/NL :
2020AIXM0001/001ED000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 10 octobre 2022 par

Grégoire AUFORT

Statistique Computationnelle Bayésienne pour l'étude des Distributions Spectrales d'énergie des galaxies

Discipline
Mathématiques

École doctorale
ED 184 MATHEMATIQUES ET INFORMATIQUE

Laboratoire/Partenaires de recherche
Institut de Mathématiques de Marseille
Laboratoire d'Astrophysique de Marseille

Composition du jury

Sylvain LE CORFF Institut Polytechnique de Paris	Rapporteur
Marc HUERTAS-COMPANY Observatoire de Paris	Rapporteur
Véronique BUAT Aix-Marseille Université	Examinatrice
Florence FORBES INRIA	Examinatrice
Clotilde LAIGLE Institut d'Astrophysique de Paris	Examinateur
Nicolas CHOPIN ENSAE	Président du jury
Pierre PUDLO Aix-Marseille Université	Directeur de thèse
Denis BURGARELLA Laboratoire d'Astrophysique de Marseille	Directeur de thèse

Je soussigné, Grégoire Aufort, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Pierre Pudlo et Denis Burgarella, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 15 Mai 2022

Résumé

Le développement de nouveaux outils de mesure et d'observation en astrophysique permet la collecte de données de plus en plus nombreuses, précises et variées. Ces données peuvent être des images complètes ou des mesures du flux lumineux à certaines longueurs d'onde (de la spectroscopie haute résolution sur de petites parties du spectre électromagnétique, ou de la photométrie, moins résolue mais couvrant une plus grande partie du spectre). L'exploitation de cette manne d'information nécessite toutefois le développement de nouveaux outils statistiques afin d'être efficace et précis. On s'intéresse en particulier à de nouveaux outils de statistique bayésienne pour l'étude des distributions spectrales d'énergie des galaxies.

Apres une introduction à l'analyse des distributions spectrales d'énergie, la première partie de cette thèse propose un algorithme de calcul bayésien approché (Approximate Bayesian Computation, ABC) pour le choix de modèle d'histoire de formation stellaire à partir de données photométriques. Cet algorithme est basé sur la simulation d'un ensemble échantillonné selon la distribution a priori de chaque modèle, puis sur l'apprentissage d'un classifieur dont la sortie est utilisée directement comme estimation de la probabilité a posteriori de chaque modèle. La méthode est appliquée à des données issues du relevé COSMOS pour l'identification de galaxies dont le taux de formation stellaire a subi une violente altération dans un passé proche, que ce soit une augmentation (dite starburst) ou une diminution (quenching). De telles altérations participeraient à expliquer les variations observées dans le rapport entre la masse stellaire d'une galaxie et son taux de formation stellaire observés.

La seconde partie de la thèse propose un nouvel algorithme d'échantillonnage préférentiel adaptatif multiple : TAMIS (Tempered Anti-Truncated Multiple Importance Sampling). En introduisant une suite de distributions cibles auxiliaires auto-calibrées, TAMIS résout le problème d'initialisation et de réglage des hyper-paramètres qui limite l'utilisation automatique de l'échantillonnage préférentiel adaptatif. Cet algorithme est robuste au fléau de la dimension ainsi qu'à une mauvaise initialisation et ne requiert que relativement peu d'évaluations de la densité de cible, sans utiliser son gradient.

La troisième partie présente le code CIGALE utilisé pour la modélisation des distributions spectrale des galaxies à partir de modèles physiques. Le calcul de la SED attendue à paramètres connus se fait par étapes successives (détermination de l'histoire de formation stellaire puis calcul des émissions lumineuses de la population d'étoiles correspondante, ajout des émissions du gaz nébulaire, absorption et ré-émission par la poussière, décalage vers le rouge dû à la distance). Pour accélérer le temps de calcul, nous proposons le remplacement des calculs explicites de certaines étapes par une approximation par réseau de neurones. Cette approximation permet de diminuer

de façon drastique le temps de calcul et d'interpoler des valeurs sur des grilles précalculées.
Enfin la dernière partie présente un modèle statistique complet pour l'estimation des paramètres et le choix de modèle bayésiens prenant en compte à la fois les données photométriques et spectroscopiques, puis l'implémentation et l'application de TAMIS à ce problème spécifique.

Mots clés : Distribution spectrale d'énergie, galaxie, statistique bayésienne, échantillonnage préférentiel, apprentissage statistique

Abstract

The development of new measurements and observation tools in astrophysics allows the collection of increasingly numerous, precise and varied data. This data can be full images or light flux measurements at certain wavelengths (high resolution spectroscopy on narrow parts of the light spectrum, or lower resolution but more spread over the spectrum photometry). The exploitation of this wealth of information, however, requires the development of new statistical tools in order to be effective and precise. We are particularly interested in new tools of Bayesian statistics for the study of the Spectral Energy Distributions of galaxies.
After an introduction to the analysis of spectral energy distributions, the first part of this thesis proposes an Approximate Bayesian Computation algorithm (ABC) for the choice of Star Formation History models from data photometric. This algorithm is based on simulating a sample set according to the prior distribution of each model, then training a classifier whose output is used directly as an estimate of the posterior probability of each model. The method is applied to data from the COSMOS survey for the identification of galaxies whose star formation rate has undergone a violent alteration in the near past, either an increase (called starburst) or a decrease (quenching). Such alterations would help explain the variations observed in the relationship between the stellar mass of a galaxy and its observed star formation rate.
The second part of the thesis proposes a new Multiple Adaptive Importance Sampling algorithm: TAMIS (Tempered Anti-Truncated Multiple Importance Sampling). By introducing a sequence of self-calibrated auxiliary target distributions, TAMIS solves the hyper-parameter initialization and tuning problem that limits the automatic use of Adaptive Importance Sampling. This algorithm is robust to the curse of dimensionality as well as poor initialization, and requires relatively few evaluations of the target density, without using its gradient.
The third part presents the CIGALE code used for modeling the spectral distributions of galaxies from physical models. The calculation of the expected SED with known parameters is done in successive stages (determination of the Star Formation History, then computation of the light emissions of the corresponding stellar population, addition of the emissions of the nebular gas, absorption and re-emission by the dust, redshift due to distance). To speed up the computation, we propose the replacement of the explicit computations of certain steps by a neural network approximation. This approximation makes it possible to drastically reduce the computation time and interpolate values on precomputed grids.
Finally the last part presents a complete statistical model for the Bayesian parameter inference and model choice taking into account both photometric and spectroscopic data, then the implementation and application of TAMIS to this specific problem.

Keywords: Bayesian statistics,Spectral Energy Distribution,Importance Sampling, Machine Learning, Galaxies

Remerciements

Je tiens en premier lieu a remercier mes co-directeurs de thèse, Pierre Pudlo et Denis Burgarella. Leur soutien, expertise, patience et ouverture ont été fondamentaux à la réussite de ce travail proprement inter-disciplinaire.

Je remercie également Véronique Buat et Laure Ciesla pour leur encadrement et leur soutien durant le stage qui m'a introduit à la recherche en astrophysique. C'est un domaine d'application passionnant que je n'aurais certainement jamais étudié autrement.

Merci à Yannick et Médéric pour leur aide quant à l'utilisation de CIGALE et leur experience du SED fitting.

Merci à Patrice, Samuel, Jean-Charles, François-Xavier, Olivier, Stéphane pour leur soutien et leur sympathie; à Bastien et Redda pour des conversations aussi accaparantes que généralement peu productives ; à Jorge et Jana et aux autres doctorants du LAM et de l'I2M, ainsi qu'aux différentes équipes pédagogiques avec lesquelles j'ai enseigné durant ces quatre ans.
Merci également à ceux que j'aurais oublié de citer.

Merci aux rapporteurs (Sylvain Le Corff et Marc Huertas-Company), qui ont eu à lire ma thèse pendant l'été, et aux membres du jury (notamment Nicolas Chopin qui a accepté de présider).

Je remercie enfin évidemment mes parents, frères et amis (notamment Samy) pour leur soutien, spécialement durant la pandémie !

“Im Anfang war die Tat."
“Au commencement était l’action."

Goethe - Faust

Contents

List of Figures

List of Tables

1. Introduction

Sommaire

1.1. Motivation and context

Advances in the development of observational tools in astronomy allow for the collection of an ever increasing amount of data about galaxies. For example the Cosmic Evolution Survey (COSMOS) collected multi-wavelength data for over two million galaxies, spanning 75% of the age of the Universe. More recent tools such as the MOONS spectrograph or the James Webb Space Telescope (JWST) will amplify this phenomenon.
Once analyzed, this enormous amount of data will inform us about the nature, physical properties, history and evolution of those gigantic systems of billions of stars, gas and dust. As the quality and the quantity of the collected data increase, we are able to create and fit models of growing complexity to better understand the physical phenomena at play. This complexity, combined with the number of galaxies to study, is a computational challenge and requires also developing new statistical tools. The rise of Bayesian methods in Physics and Astronomy since the 1990's (c.f Loredo, 2013) allows for a principled quantification of the uncertainty surrounding the estimates of the physical quantities of interest, but often at the cost of the computationally expensive use of Monte Carlo algorithms, mostly as Markov Chain Monte Carlo - such as the popular Nested Sampling algorithm (Skilling, 2006), Metropolis-Hastings (Metropolis and Ulam, 1949) or emcee (Foreman-Mackey, Hogg, Lang, et al., 2013)- or grid sampling (e.g Roehlly, Burgarella, Buat, Boquien, et al., 2014 or Guinevere Kauffmann, Timothy M. Heckman, Simon D. M. White, et al., 2003.

This thesis focuses on developing a Bayesian framework and the necessary computational tools to study the physical properties of galaxies using measurements of their Spectral Energy Distribution. After an introduction to the basic ideas of galaxy emission modeling and Bayesian inference, the first chapter proposes a machine learning based Approximate Bayesian Computation scheme for the choice of Star Formation History model. The second chapter introduces a new Monte Carlo algorithm in the Adaptive Importance Sampling methodology. The third chapter reviews the CIGALE code for SED modeling and the application of Neural Network approximations of physical models. Finally the last chapter proposes a comprehensive methodology for Bayesian SED fitting, as well as its implementation and application to simulated data.

1.2. A primer on galaxy emissions

1.2.1. Spectral Energy Distribution

The spectrum of a galaxy contains the information on the physical processes acting in galaxies and the contents of these galaxies. This information is distributed over the entire electromagnetic spectrum (except for gravitational waves). A good and complete analysis asks for the widest wavelength range, at least from X-rays to radio if we wish to study the co-evolution of black holes and galaxies. Of course, in practice this is not possible for every single galaxy in the universe and not even for very large samples because spectroscopy is time consuming, and parts of the spectra are missing. A less expensive way is to observe only at a finite number of wavelengths. Those observations are called the Spectral Energy Distribution (SED) of the galaxy(Fig.1.1). The SED can be decomposed in several features:

— The continuum: Mostly due to the light emitted by the stellar populations (hundreds of millions to hundreds of billions of stars) making up the galaxies, with an (often small) contribution from the nebular emissions from the ultraviolet to the near-infrared. In the mid-infrared to the far-infrared, the emission from the dust heated by the stars and/or the active galactic nucleus dominates.

— Emission lines: spikes in luminosity superposed to the continuum. They can provide some information on the gas (atoms and molecules) and, for instance, on the metallicity properties of the galaxies (their chemical composition).

— Absorption lines : Narrow drops in luminosity at specific wavelength due to the absorption of light by atoms and molecules.

To interpret these observations and to measure fundamental physical properties of galaxies (e.g., star formation rate (SFR) and history (SFH), stellar mass, attenuation, dust mass, presence and characteristics of an active galactic nucleus, and so on), significant investments have been made in developing ever more precise and accurate models of galaxies' emission over multiple orders of magnitude in wavelength. Modeling the SED of galaxies is a difficult challenge to solve for at least two reasons: the diversity of physical phenomena acting in galaxies and the degeneracies that make SEDs to appear very similar for galaxies with very distinct characteristics. This is especially true when limited wavelength ranges are used instead of the complete SED, which is extremely difficult to impossible to collect because of the variety of telescopes on the ground and in space that are necessary to observe over the entire electromagnetic spectrum. As a result, determining the physical properties of galaxies accurately and precisely with incomplete data is a significant issue. In practice, different avenues can be taken to build physically motivated SED models and attempt to adjust them to the observations.

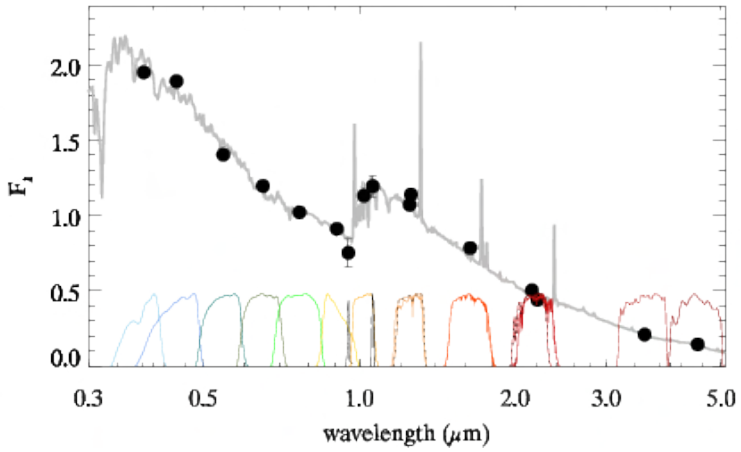

Figure 1.1. – A galaxy spectrum with a few points of measure. Each dot corresponds to the middle of the filter (the colored curves). From left to right, the 16 bands U , B, V,R, i, z, $[S_{III}]+65$, Y, NB1.06, JWFCAM, JHAWK-I, H, Ks, K, IRAC1, IRAC2. Extracted from Hatch, Muldrew, Cooke, et al., 2016.

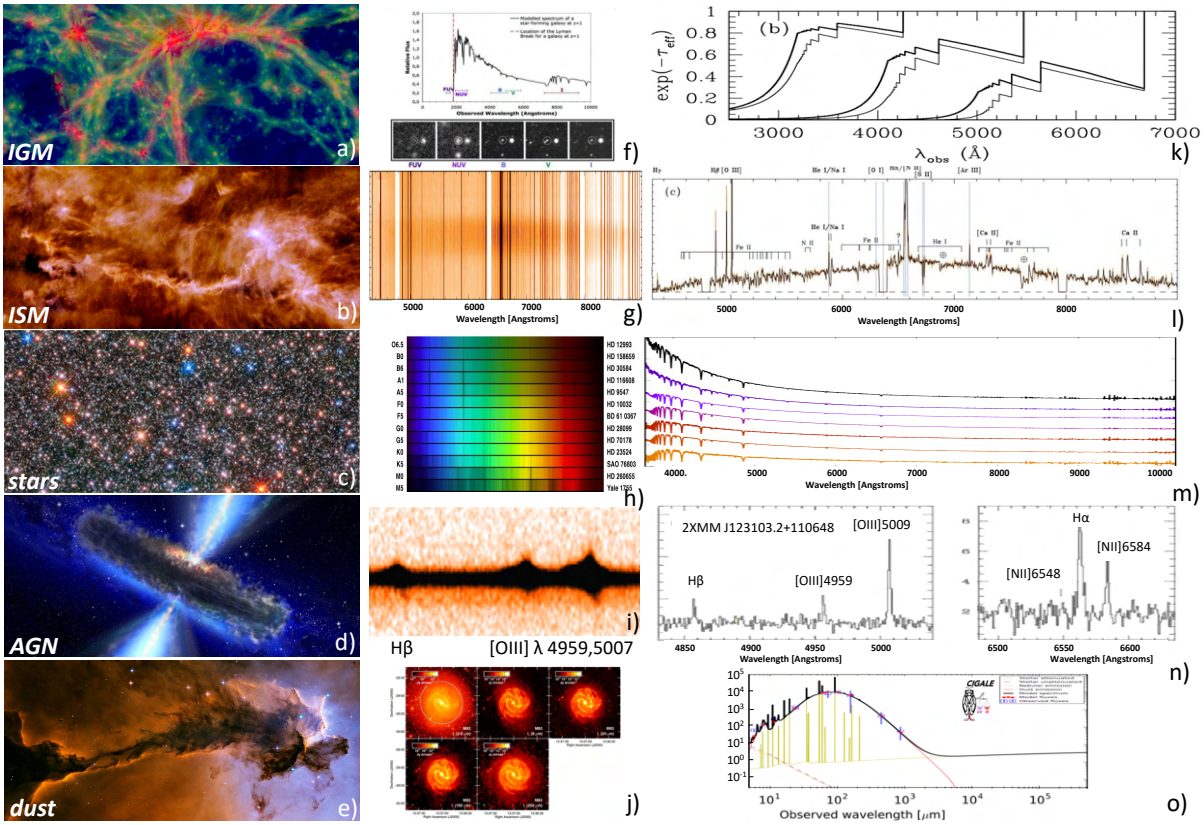

Figure 1.2. – Most physical processes at play in a galaxy contribute to the shape of the emitted SED. The left-hand part of the figure shows observed or art illustrations of the various components. The central and right-hand parts present representative emissions for each of them. If we wish to understand galaxies that are multi-facet objects, we need to be able to model each and every physical process shown here. Credits: a) from Schaye et al. (2015) by permission of Oxford University Press on behalf of the Royal Astronomical Society, b) ESA/Herschel/PACS, SPIRE/Gould Belt survey Key Programme/Palmeirim et al. (2013), c) NASA, ESA, and T. Brown (STScI), d) ESA/NASA, the AVO project and Paolo Padovani, e) NASA, ESA and the Hubble Heritage Team (STScI/AURA), g) and l) from Smith et al. (2018) by permission of Oxford University Press on behalf of the Royal Astronomical Society, h) NOAO/AURA/NSF, i) from Villar-Martin et al. (2011) by permission of Oxford University Press on behalf of the Royal Astronomical Society, j) from Jones et al. (2015) by permission of Oxford University Press on behalf of the Royal Astronomical Society, k) from Meiksin (2006) by permission of Oxford University Press on behalf of the Royal Astronomical Society, m) from Kesseli et al. (2017) ©AAS. Reproduced with permission, n) from Ho et al. (2012) ©AAS.

1.2.2. Stellar Emissions

In order to compute the light emitted by the stars of a galaxy, we need to make assumptions on several different components like, for instance, the chemical composition of the stars, their distribution in mass, evolution in time, etc.

Star Formation and evolution Stars are continuously formed from the gas present in a galaxy. The rate of this formation (measured in solar mass per year, $M_\odot / yr$) is called Star Formation Rate (SFR). Star formation mainly happens in Giant Molecular Clouds (see Fig 1.3) : dense regions composed of molecular Hydrogen H_2 protected for UV radiations by dust. If a given cloud of gas is perturbed, it will start to contract, and under specific conditions of the temperature and density will ultimately form a single star or several of them. Since a fundamental property affecting stellar emissions is the mass m of a star, we first need to assume an Initial Mass Function (IMF) describing the distribution of the masses of the stars in a given stellar population. Several such IMF have been described (Salpeter 1955,Chabrier 2003, see fig.1.4).

Figure 1.3. – Star-forming region called NGC 3324 in the Carina Nebula. Credits NASA, ESA, CSA, and STScI

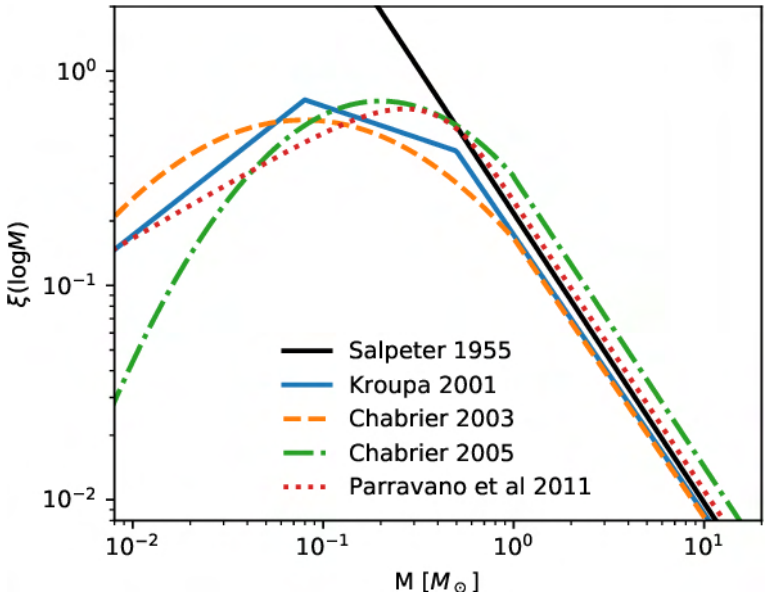

Figure 1.4. – Several proposed Initial Mass Functions. Each curve represents the number of stars (y-axis) with a given stellar mass (x-axis) in the galaxy initial population of stars. Figure adapted from Colman and Teyssier, 2020. We note that there are few massive stars (right) and that most stars are comparable to our Sun in term of their mass.

We call Single Stellar Population (SSP) a group of stars formed at the same time, following the same IMF and with the same chemical abundances (or metallicity, Z). By following the evolution of each star in this SSP along time, we can assign a spectra to each star at each time-step depending on its individual properties (mass, age, metallicity, effective temperature $T_{\rm eff}$).

However we know that most galaxies do not form all their stars simultaneously at a given time. This formation activity fluctuates during the life of the galaxy and is characterized by the SFR at each timestep. This fluctuation is called the Star Formation History (SFH) of a galaxy (see Fig.1.5 adapted from Ciesla, Elbaz, and Fensch, 2017b). Modeling the SFH is an active research area. Popular models includes simple parametric forms like

$$\mathrm{SFR}(t) \propto \exp(-t/\tau) \tag{1.1}$$

for example, that describes an exponentially decreasing SFR called the τ-exponential SFH, where t is the time. However much more complex parametric, or non parametric models have been proposed as it is unlikely that a single simple model describes accurately the evolution of all galaxies in the Universe.

Stellar Emission Assuming we know the IMF, metallicity and SFH of a given galaxy we can follow the distribution of mass and age of the stars composing it. As we also have models for the stellar evolution of single stars and their luminosity, we can model the galaxy luminosity at a given wavelength λ and time t. Let ϕ be the IMF, $F_\lambda(m,t,Z)$ the flux emitted according to the spectral library used and T the age of the galaxy. Then the galaxy emissivity is given by :

$$L_\lambda(t,Z) = \int_0^T \int_M F_\lambda(m, T-t, Z)\mathrm{SFR}(t)\phi(m)\,dm\,dt \tag{1.2}$$

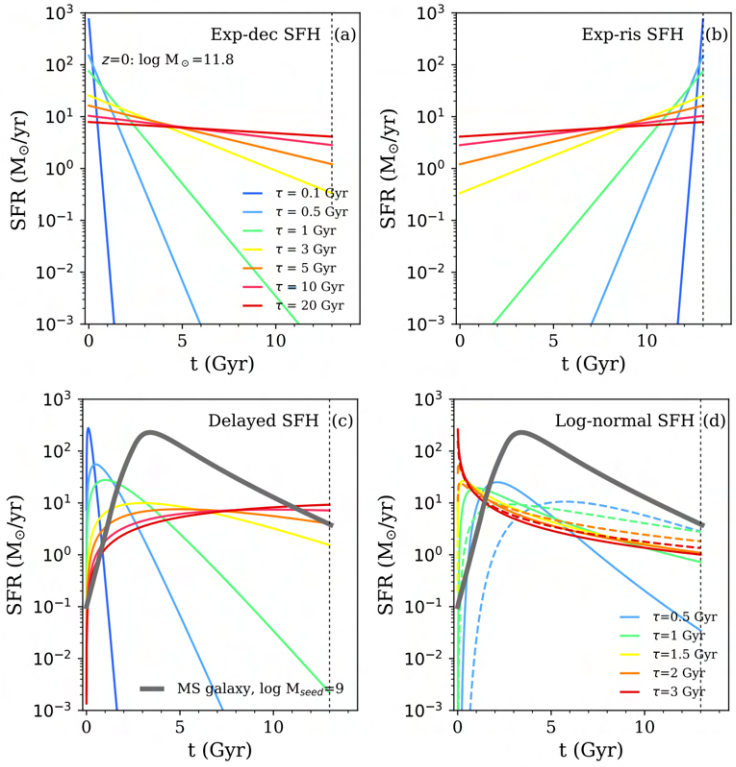

Figure 1.5. – Different SFH models. Top the “exponential" decreasing (left) and rising (right) for different values of the τ parameter. Bottom are the“Delayed" model (left), and the “log-normal" model (right). They describe different possible analytical evolutions of the SFR (y-axis) as a function of time (x-axis)

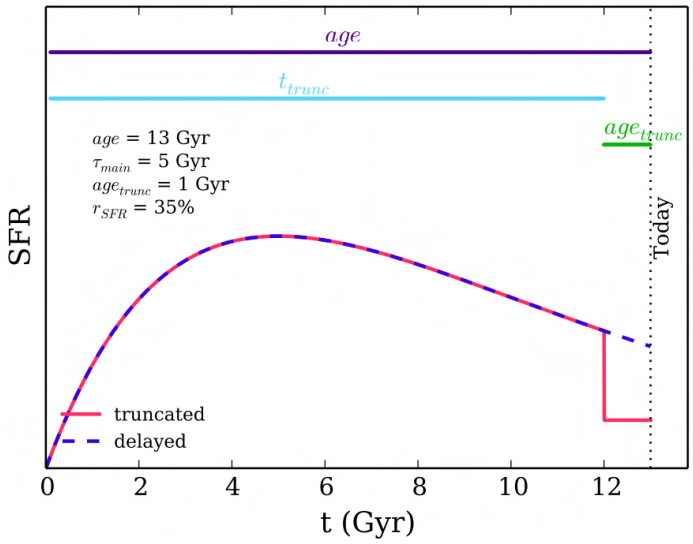

Figure 1.6. – The description of the SFH can lead to nested models. An example is the “Delayed + Trunc" model, of which “Delayed" (Bottom left of Fig.1.5) is a specific case where $age_{trunc} = 0$ or $r_{SFR} = 1$. This particular example is developped in depth in chapter 1.

1.2.3. Nebular emissions

As the stellar population of galaxy emits light, a fraction of the emission happens below 91,2 nm. This wavelength, called the Lyman break, is the limit below which photons are energetic enough to ionize the interstellar gas. In turn this ionized gas re–emits the energy in the form of a series of emission lines and a continuum. Studying those emission lines is a fundamental tool to learn about the star formation (through hydrogen lines and radio continuum) and the abundance of the different elements composing the gas (through the metal lines). As those emission lines only have a very local imprint in wavelength on the SED, studying them often requires the use of high resolution spectroscopy, although in certain cases their impact on broadband photometry measurements is substantial and needs to be taken into account for SED modeling (Boquien, Burgarella, Roehlly, et al., 2019 ; see Fig 1.7)

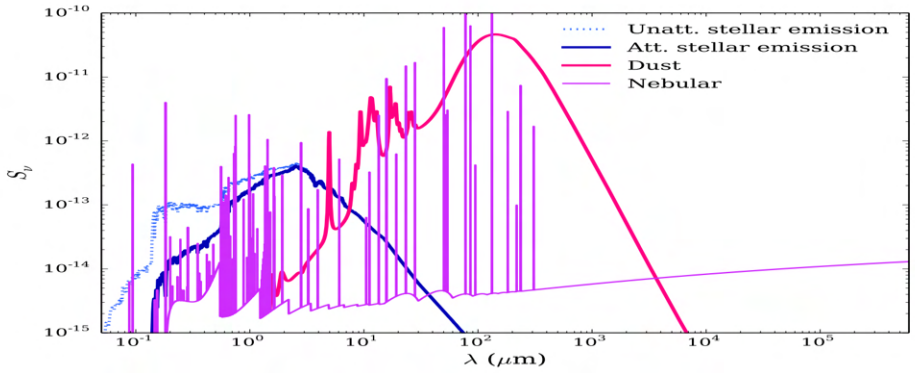

Figure 1.7. – Illustration of the contributions of emission lines due to the interstellar medium gas being ionized by the stellar light. Those emission lines are added to the dust and stellar emissions with which they interact to obtain the full emission of a galaxy.

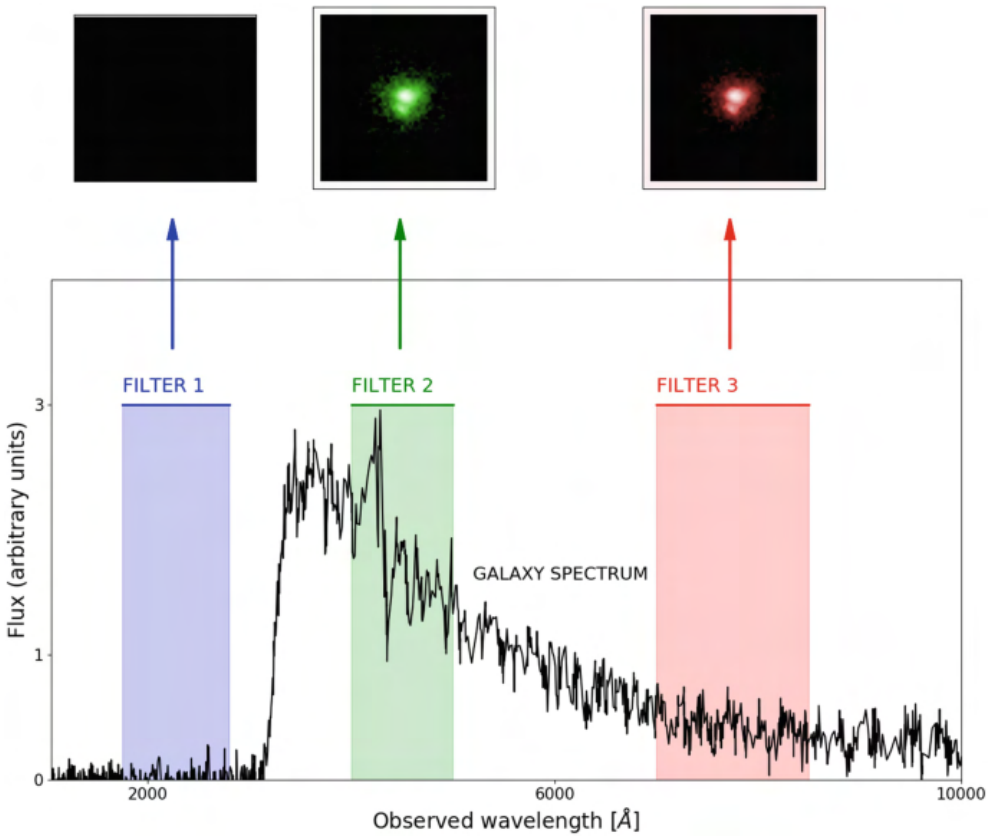

Figure 1.8. – Example of a Lyman Break Galaxy around $z = 1$. The measurements in the broadbands 1 and and 2 allows us to clearly locate the Lyman break (figure from Orlitova, 2020)

1.2.4. Dust contributions

The final piece we introduce about galaxy spectra composition is the effect of dust. Interstellar dust consist in small solid particles, often less than 1μm, with a typical size of 0.1μm. They are composed of abundant condensible elements (C, O, Mg, Si, S and F) that are put together to form grains such as silicate, carbon solids and hydrocarbons. Those particles will attenuate the stellar light at the shorter wavelengths and re-emit it at a longer wavelengths.

Dust attenuation Dust in the interstellar medium have two effects on the light emitted by the stars: reddening and obscuration. Reddening is caused by the differential absorption and scattering of the shorter wavelength (blue) light. The exact dependency between the extinction of light and the wavelength is described by a *reddening law*, which depends on the characteristics of the dust grains. This law is often characterized by the color excess E(B-V), the difference between the observed color and its intrinsic color measured in the filters B(440 nm) and V(550 nm). The attenuation law is then defined as:

$$A_\lambda = k(\lambda)E(B-V) \tag{1.3}$$

where $k(\lambda)$ is a reddening curve. The most frequently used reddening curve was empirically derived by Calzetti (D. Calzetti, Armus, Bohlin, et al., 2000), but other parametrizations have been proposed (see Fig.1.9)

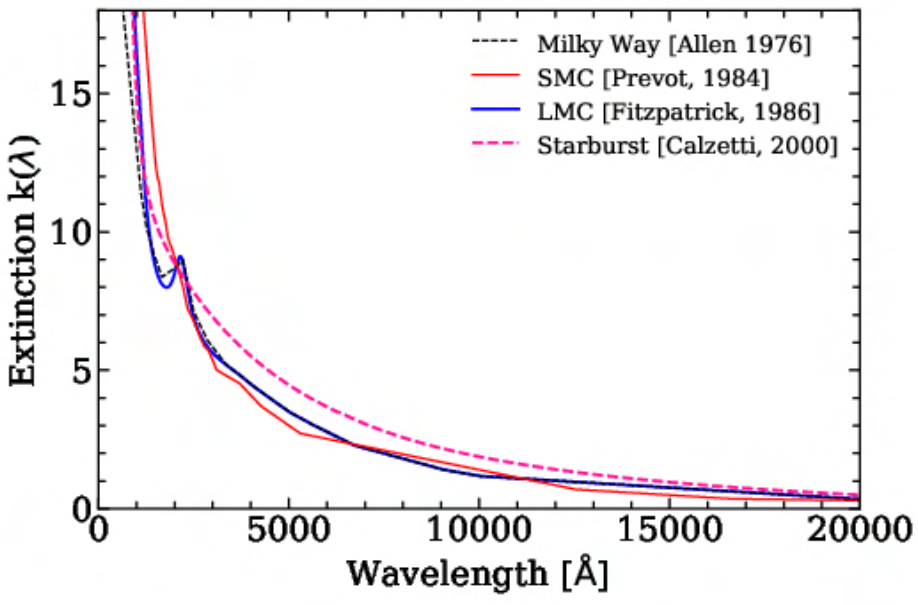

Figure 1.9. – Different reddening curves proposed to model dust attenuation.

Dust emission The mid-infrared (MIR) part of the spectrum of most galaxies that normally form stars is dominated by emission from Polycyclic aromatic hydrocarbons (PAHs) which produce a characteristic emission between $3\mu m$ and $20\mu m$ with emission peaks at $3.3, 6.2, 7.7, 8.6, 11.3$ and $12.7\mu m$. At FIR wavelengths, the emission from galaxies is generally dominated by the emission of "large" dust grains (between 10 nm and $0.1\mu m$) at low temperatures and at thermal equilibrium. This is why the peak intensity in IR of the SED of a galaxy is a good indicator of the heating of dust in the interstellar medium. The temperature of the grains depends on the intensity of the interstellar radiation field.

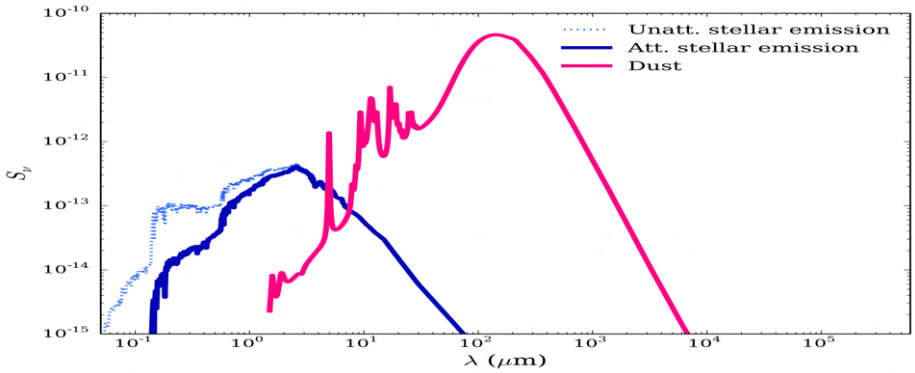

Figure 1.10. – Illustration of the influence of dust on the spectral energy distribution. Part of the emissions due to the stellar population is attenuated at low wavelength (in blue), and re-emitted by the dust at higher wavelengths(in red)

1.2.5. Redshift

Redshift is an increase in the wavelength of the observed spectrum of a distant galaxy relatively to its restframe emissions due to the dilatation of the universe on which light moves. It is used as a characterization of the distance to the source of the spectrum (and its age, see table 1.1 from Wright, 2006) . The redshift z of a galaxy can be estimated using spectroscopic measurement: by comparing the wavelength λ_{obs} of well known and easily identifiable features (such as emission lines) in the observed spectrum to the wavelength λ_0 of the same features here on Earth.

The lack of photons below the Lyman break is a very effective way to determine the redshift of a galaxy as the Lyman Break will be observed at $\lambda_z = (1+z) \times 91,2$ nm (Bouwens, Illingworth, Labbe, et al., 2011) The redshift is then easily derived as:

$$z = \frac{\lambda_{\text{obs}} - \lambda_0}{\lambda_0}$$

Table 1.1. – Emission age for a few values of z

z	Age (in Gyr)
1	6.678
2.5	9.175
4	10.049
6	10.568
10	10.977
15	11.165

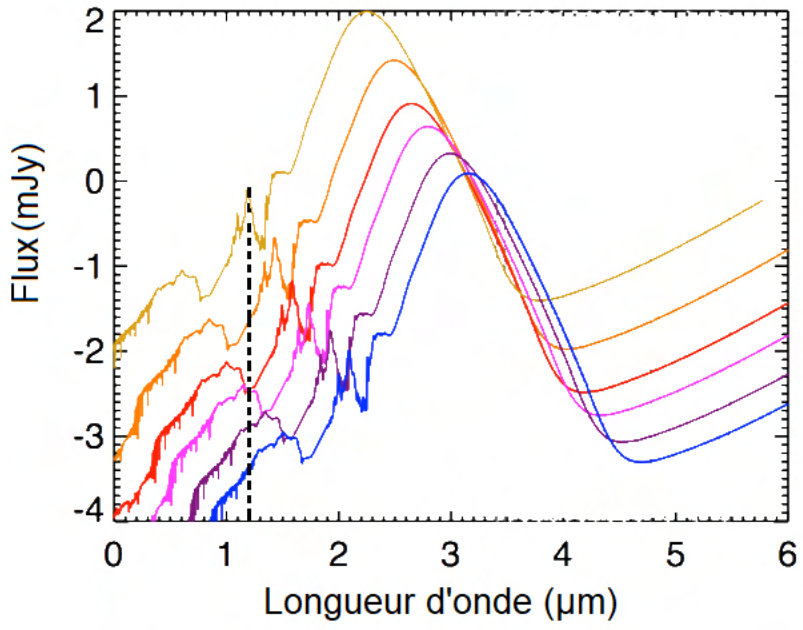

Figure 1.11. – From Caitlin M. Casey et al., 2015. SED of the same galaxy at different redshift : $z = 1$ (yellow), $z = 2.5$ (orange),$z = 4$ (red), $z = 6$ (magenta), $z = 10$ (violet), and $z = 15$ (blue). Since measurements on the SED are taken at fixed wavelengths, different features would be probed by the same instrument on objects at different redshifts.

Redshift has two related effects: it shifts the spectrum to the right (long wavelengths) multiplying the wavelengths by $1 + z$, and it decreases the observed flux, dividing the spectrum by a factor depending on the distance of the galaxy. As measuring instruments probe the spectrum at a given wavelength, redshift must be taken into account as the spectrum features observed by the instrument will depend on the redshift of the galaxy.

1.3. A primer on Bayesian statistics

Bayesian statistics is an approach to statistical analysis using Bayes' theorem to update a state of knowledge about parameters in a statistical model with the information brought by new data. The initial state of knowledge is expressed as a prior distribution. It is combined with the observed data through likelihood function to determine the posterior distribution. The latter describes the state of knowledge after the observation. This posterior distribution can then be used for making inference, taking decisions or making predictions (see e.g. Schoot, Depaoli, King, et al., 2021 ; C. Robert, 2007)
For studying a parameter of interest θ given an observation x, a Bayesian statistical model is composed of a likelihood function $p(x|\theta)$ and a prior distribution over the parameters $p(\theta)$.

1.3.1. Parameter inference

In Bayesian statistics, the focus is on estimating the entire posterior distribution of the model parameters. From a formal viewpoint we consider a model $\mathscr{M}$, parametrized by a vector θ. For example, let us consider a Gaussian linear model on the dataset $\mathscr{D} = \{(x^i, y^i)\}_{i=1}^N$, with $x^i \in \mathbb{R}^2$ and $y_i \in \mathbb{R}$. It means that

$$y^i = \beta_1 x_1^i + \beta_2 x_2^i + \varepsilon^i \tag{1.4}$$

with $\varepsilon^i \sim \mathcal{N}(0, \sigma^2)$. The corresponding likelihood is

$$p(y|x,\theta) = \left(2\pi\sigma^2\right)^{-N/2} exp\Big(-\frac{1}{2\sigma^2}||y - x\beta||^2\Big) \tag{1.5}$$

Our goal is to estimate the parameters $\theta = (\beta_1, \beta_2, \sigma^2)$. We start by setting a prior distribution $p(\theta)$ and apply Bayes' theorem to obtain the posterior

$$p(\theta|x,y) = \frac{p(y|x,\theta)p(\theta)}{p(y)}. \tag{1.6}$$

In the above equation, $p(y) = \int p(y|x,\theta)p(\theta)d\theta$ is the marginal likelihood or the model evidence.

The posterior distribution provides point estimates, such as the posterior mean, mode, variance, median, or credible intervals. These point estimates are known to be efficient, see e.g C. Robert, 2007, mainly because the posterior encodes all the information regarding θ given by the data $\mathscr{D}$. However computing those quantities is typically not directly possible as it requires computing integrals of the posterior density which are not tractable (and in practice often high-dimensional as the dimensionality is the number of parameters of the model). The most common solution is to resort to Monte Carlo methods (Christian P. Robert and Casella, 2004, C. Robert,

2007,A. Gelman, Carlin, H. Stern, et al., 2013) , to obtain a sample from the posterior distribution. Computing point estimates and credible intervals from this sample is an easy Monte Carlo routine. The previous section present the Bayesian methodology to estimate the parameters of a given model, with a given prior distribution.

1.3.2. Model choice and checking

We need to establish a way to select a model and check that it is somewhat relevant to describe the data. Yet we will not explore the different ways to choose the prior distribution(see C. Robert, 2007 or A. Gelman, Carlin, H. Stern, et al., 2013 for details on the prior ellicitation problem).
In the previous regression example, we considered a linear model with two covariates. But maybe we know of a third that could be useful ? Or a polynomial regression could be better suited ? Then we also need to choose the degree of the polynomial... More generally assume we have a collection of models $\{\mathcal{M}_i\}_i$ competing to explain our dataset $\mathcal{D}$. Once again the Bayesian approach is to assign a prior probability $p(\mathcal{M}_i)$ to each model, and to compute the posterior probability of the model conditionally to the data, namely

$$p(\mathcal{M}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_i)\,p(\mathcal{M}_i)}{p(\mathcal{D})} \tag{1.7}$$

where $p(\mathcal{D}) = \sum_i p(\mathcal{D}|\mathcal{M}_i)\,p(\mathcal{M}_i))$ is the model evidence
Unfortunately computing the model evidence again typically requires numerical integration over the parameter space. Few algorithms are reliable to obtain these quantities from posterior samples. The bridge sampling is the most popular and reliable, though a bit tricky to implement when the dimensions of the models differ. The simpler harmonic mean estimator is still popular, though not reliable, see e.g Marin and C. Robert, 2009. In addition to computing the posterior probability of a model given the data, practical Bayesian model choice sometimes rely on the computation of one Bayes factor (Kass and Raftery, 1995) to quantify the preference for one model over an other. If we consider two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, with prior probability over the models $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$, we can define :

$$\begin{aligned} B_{1/2} &= \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} \\ &= \frac{p(\mathcal{M}_1|\mathcal{D})}{p(\mathcal{M}_2|\mathcal{D})} \times \frac{p(\mathcal{M}_2)}{p(\mathcal{M}_1)} \end{aligned}$$

as the Bayes factor in favor of $\mathcal{M}_1$. Although setting a universal threshold as a decision rule for every problem is impossible, Jeffreys, 1998 proposed the frequently used interpretation scale given in table 1.2.

A feature of Bayesian model choice is that it gives a preference towards simpler models, in line with Occam's razor (see e.g C. Robert, 2007). Indeed as the marginal likelihood can be seen as a probability distribution over the data space conditionally to the model, it must normalize to 1 over this space. Since more complex models can

Table 1.2. – Jeffreys scale

Grade	Evidence against $\mathscr{M}_2$	Bayes Factor
1	Barely worth mentioning	10^0 to $10^{\frac{1}{2}}$
2	Substantial	$10^{\frac{1}{2}}$ to 10^1
3	Strong	10^1 to $10^{\frac{3}{2}}$
4	Very strong	$10^{\frac{3}{2}}$ to 10^2
5	Decisive	$> 10^2$

generate a greater diversity of observations, the supports of their distributions are wider in the data space, penalizing the probability density of any one observation. When shifting from a simpler to a more complex model, the probability density of some datasets explained by the simpler model must decrease in order to increase the amount of datasets explained by the more complex model. Although Bayesian model choice is sensitive to the choice of the prior distribution, it therefore naturally handles the notion of complexity otherwise requiring criterions such as AIC (Akaike, 1973) or BIC (Schwarz, 1978).

Finally, once we are able to select a model among competing ones and compute the posterior distribution, we should be able to investigate whether the model and the estimated parameters are able to properly explain the observed data. Intuitively, if the model is a relevant way to explain the observed data, new data simulated from the model should be compatible with the observed ones (see e.g A. Gelman, Carlin, H. Stern, et al., 2013). Posterior predictive checks are techniques to assess this compatibility.
Assuming $\mathscr{D}$ is our observed data, we computed the posterior distribution

$$p(\theta|\mathscr{D}) = \frac{p(\mathscr{D}|\theta)\,p(\theta)}{p(\mathscr{D})}$$

and from this distribution, we can construct the *posterior predictive distribution*

$$p(\mathscr{D}^{\text{rep}}|\mathscr{D}) = \int p(\mathscr{D}^{\text{rep}}|\theta)\,p(\theta|\mathscr{D}^{\text{rep}})\,d\theta$$

i.e the distribution of the data $\mathscr{D}^{\text{rep}}$ we expect given the model and our knowledge on θ based on $\mathscr{D}$. From this posterior predictive distribution, we have two simple ways to assess the compatibility with the observed data :

— Draw a sample from $p(\mathscr{D}^{\text{rep}}|\mathscr{D})$and graphically assess that the observed $\mathscr{D}$ lies in the simulated sample

— Define a test statistic $T(\mathscr{D}^{\text{rep}},\theta)$ and check that $T(\mathscr{D}^{\text{rep}},\theta)$ are not significantly different from what we expected .

2. Bayesian Model Choice for Star Formation History model Selection

Sommaire

Abstract

Although galaxies are found to follow a tight relation between their star formation rate and stellar mass, they are expected to exhibit complex star formation histories (SFH), with short-term fluctuations. The goal of this pilot study is to present a method that will identify galaxies that are undergoing a strong variation of star formation activity in the last tens to hundreds Myr. In other words, the proposed method will determine whether a variation in the last few hundreds of Myr of the SFH is needed to properly model the SED rather than a smooth normal SFH. To do so, we analyze a sample of COSMOS galaxies with $0.5 < z < 1$ and $\log M_* > 8.5$ using high signal-to-noise ratio broad band photometry. We apply Approximate Bayesian Computation, a state-of-the-art statistical method to perform model choice, associated with machine learning algorithms to provide the probability that a flexible SFH is preferred based on the observed flux density ratios of galaxies. We present the method and test it

on a sample of simulated SEDs. The input information fed to the algorithm is a set of broadband UV to NIR (rest-frame) flux ratios for each galaxy. The choice of using colors is made to remove any difficulty linked to normalization when using classification algorithms. The method has an error rate of 21% in recovering the right SFH and is sensitive to SFR variations larger than 1 dex. A more traditional SED fitting method using CIGALE is tested to achieve the same goal, based on fits comparisons through Bayesian Information Criterion but the best error rate obtained is higher, 28%. We apply our new method to the COSMOS galaxies sample. The stellar mass distribution of galaxies with a strong to decisive evidence against the smooth delayed-τ SFH peaks at lower M_* compared to galaxies where the smooth delayed-τ SFH is preferred. We discuss the fact that this result does not come from any bias due to our training. Finally, we argue that flexible SFHs are needed to be able to cover that largest SFR-M_* parameter space possible.

2.1. Introduction

The tight relation linking the star formation rate (SFR) and stellar mass of star-forming galaxies,the so-called main sequence (MS), opened a new window in our understanding of galaxy evolution Elbaz, Daddi, Le Borgne, et al., 2007; Noeske, B. J. Weiner, Faber, et al., 2007. It implies that the majority of galaxies are likely to form the bulk of their stars through steady-state processes rather than violent episodes of star formation. However, this relation has a scatter of ~0.3 dex Schreiber, Pannella, Elbaz, et al., 2015 that is found to be relatively constant at all masses and over cosmic time Guo, Zheng, and Fu, 2013; Ilbert, Arnouts, Le Floc'h, et al., 2015; Schreiber, Pannella, Elbaz, et al., 2015. One possible explanation of this scatter could be its artificial creation by the accumulation of errors in the extraction of photometric measurements and/or in the determination of the SFR and stellar mass in relation with model uncertainties. However, several studies have found a coherent variation of physical galaxy properties such as the gas fraction Magdis, Daddi, Béthermin, et al., 2012, Sersic index and effective radius S. Wuyts, Förster Schreiber, van der Wel, et al., 2011, and U-V color e.g., Salmi, Daddi, Elbaz, et al., 2012, suggesting that the scatter is more related to the physics than to measurement and model uncertainties. Furthermore, oscillations of the SFR resulting from a varying infall rate and compaction of star-formation have been proposed to explain the MS scatter Sargent, Daddi, Béthermin, et al., 2014; Scoville, Sheth, Aussel, et al., 2016; Tacchella, Avishai Dekel, Carollo, et al., 2016 and even suggested by some simulations e.g., A. Dekel and Burkert, 2014.

To decipher if the scatter is indeed due to star formation history (SFH) variations, one must be able to put constraint on the recent star formation history (SFH) of galaxies, to reconstruct their path along the MS. This information is embedded in the spectral energy distribution (SED) of galaxies. However, recovering it through SED modeling is complex and subject to many uncertainties and degeneracies. Indeed, galaxies are expected to exhibit complex SFHs, with short-term fluctuations, requiring sophisticated SFH parametrizations to model them e.g., Lee, Ferguson, Somerville,

et al., 2010; Pacifici, Kassin, B. Weiner, et al., 2013; Behroozi, Wechsler, and Conroy, 2013; Pacifici, S. Oh, K. Oh, et al., 2016; Leja, Carnall, Johnson, et al., 2019. The implementation of these models is complex and large libraries are needed to model all galaxies properties. Numerous studies have, instead, used simple analytical forms to model galaxies SFH e.g., Papovich, Dickinson, and Ferguson, 2001; C. Maraston, Pforr, Renzini, et al., 2010; Pforr, C. Maraston, and Tonini, 2012; Gladders, Oemler, Dressler, et al., 2013; Simha, Weinberg, Conroy, et al., 2014; Buat, Heinis, Boquien, et al., 2014; Boquien, Buat, and Perret, 2014; Ciesla, Charmandaris, Georgakakis, et al., 2015; Abramson, Gladders, Dressler, et al., 2016; Ciesla, Boselli, Elbaz, et al., 2016; Ciesla, Elbaz, and Fensch, 2017a. However, SFH parameters are known to be difficult to constrain from broadband SED modeling e.g., C. Maraston, Pforr, Renzini, et al., 2010; Pforr, C. Maraston, and Tonini, 2012; Buat, Heinis, Boquien, et al., 2014; Ciesla, Charmandaris, Georgakakis, et al., 2015; Ciesla, Elbaz, and Fensch, 2017a; Carnall, Leja, Johnson, et al., 2019.

Ciesla, Boselli, Elbaz, et al., 2016 and Boselli, Roehlly, Fossati, et al., 2016 have shown on a sample of well-known local galaxies benefiting from a wealth of ancillary data, that a drastic and recent decrease of the star formation activity of galaxies can be probed as long as a good UV to NIR rest frame coverage is available. They showed that the intensity of the variation of SF activity can be relatively well constrained from broadband SED fitting. Spectroscopy is however needed to bring information on the time when the change in star formation activity occurred Boselli, Roehlly, Fossati, et al., 2016. These studies were made on well-known sources of the Virgo cluster, for which the quenching mechanism - ram pressure stripping - is known and HI observations allow a direct verification of the SED modeling results. To go a step further, Ciesla, Elbaz, Schreiber, et al., 2018 have blindly applied the method on the GOODS-South sample aiming at identifying sources that underwent a recent and drastic decrease of their star-formation activity. They compared the quality of the results from SED fitting using two different SFH and obtained a sample of galaxies where a modeled recent and strong decrease of SFR produced significantly better fits of the broad band photometry. In this work, we aim at improving the method of Ciesla, Elbaz, Schreiber, et al., 2018 gaining in power by applying to a subsample of COSMOS galaxies a state-of-the-art statistical method to perform the SFH choice: the Approximate Bayesian Computation ABC, see, e.g. Marin, Pudlo, Christian P Robert, et al., 2012; Sisson, Fan, and Beaumont, 2018. Based on the observed SED of a galaxy, we want to choose the most appropriate SFH between a finite set. The main idea behind ABC is to rely on many simulated SEDs generated from all the SFHs in competition using parameters drawn from the prior.

The paper is organized as follows: Sect. 2.2 describes the astrophysical problem and presents the sample. In Sect. 2.3 we present the statistical approach as well as the results obtained from a catalog of simulated SEDs of COSMOS-like galaxies. In Sect. 2.4 we compare the results of this new approach with more traditional SED modeling methods, and apply it to real COSMOS galaxies in Sect. 2.5. Our results are discussed in Sect. 2.6.

2.2. Constraining the recent star formation history of galaxies using broad-band photometry

2.2.1. Building upon the method of Ciesla, Elbaz, Schreiber, et al., 2018

The main purpose of the study presented in Ciesla, Elbaz, Schreiber, et al., 2018 was to probe variations in SFH that occurred in very short timescales, i.e. on hundreds of Myrs. Therefore, a large-number statistics was needed to be able to catch galaxies at the moment when these variations happened. They aimed at identifying galaxies that have recently undergone a rapid (<500 Myr) and drastic downfall of their SFR (more than 80%) from broadband SED modeling, since large photometric samples can provide the statistics needed to pinpoint these objects.

To perform their study, they took advantage of the versatility of the SED modeling code CIGALE [1] Boquien, Burgarella, Roehlly, et al., 2019. CIGALE is a SED modeling software package that has two functions: a modeling function to create SEDs from a set of given parameters and a SED fitting function to derive the physical properties of galaxies from observations. Galaxies SEDs are computed from UV-to-radio taking into account the balance between the energy absorbed by dust in the UV-NIR and remitted in IR. To build the SEDs, CIGALE uses a combination of modules including the star formation history assumption, either analytical, stochastic, or outputs from simulations e.g., Boquien, Buat, and Perret, 2014; Ciesla, Charmandaris, Georgakakis, et al., 2015; Ciesla, Elbaz, and Fensch, 2017a, the stellar emission from stellar population models Bruzual and S. Charlot, 2003; C. Maraston, 2005, the nebular lines, and the attenuation by dust e.g., D. Calzetti, Armus, Bohlin, et al., 2000; S. Charlot and Fall, 2000.

Ciesla, Elbaz, Schreiber, et al., 2018 compared the results of SED fitting on a sample of GOODS-South galaxies using two different SFHs: one normal delayed-τ SFH and one flexible SFH modeling a truncation of the SFH. The normal delayed-τ SFH is given by the equation:

$$\mathrm{SFR}(t) \propto t \times exp(-t/\tau_{main}) \tag{2.1}$$

where SFR is the star formation rate, t the time, and τ_{main} is the e-folding time. Examples of delayed-τ SFHs are shown in Fig. 2.1 for different values of τ_{main}. The flexible SFH is an extension of the delayed-τ model:

$$\mathrm{SFR}(t) \propto \begin{cases} t \times exp(-t/\tau_{main}), & \text{when } t \leq t_{flex} \\ r_{\mathrm{SFR}} \times \mathrm{SFR}(t = t_{flex}), & \text{when } t > t_{flex} \end{cases}, \tag{2.2}$$

where t_{flex} is the time at which the star formation is instantaneously affected, and r_{SFR} is the ratio between $\mathrm{SFR}(t > t_{flex})$ and $\mathrm{SFR}(t = t_{flex})$:

1. https://cigale.lam.fr/

$$r_{\mathrm{SFR}} = \frac{\mathrm{SFR}(t > t_{flex})}{\mathrm{SFR}(t_{flex})}. \tag{2.3}$$

A representation of flexible SFHs is also shown in Fig. 2.1. The normal delayed-τ SFH is at first order a particular case of the flexible SFH for which $r_{\mathrm{SFR}} = 1$.

To differentiate between the two models, Ciesla, Elbaz, Schreiber, et al., 2018 estimated the Bayesian Information Criterion (BIC, see Sect. 2.3.2) linked to the two models and put conservative limits on the difference between the two BIC to select the most suited model. They showed that a handful of sources were better fitted using the flexible SFH, that assumes a recent instantaneous break in the SFH, compared to the more commonly used delayed-τ SFH. In fact, they discussed that these galaxies have indeed physical properties that are different from the main population and characteristic of sources in transition.

The limited number of sources identified in the study of Ciesla, Elbaz, Schreiber, et al., 2018 (102 out of 6,680) was due to their will to be conservative in their approach and find a clean sample of sources that underwent a rapid quenching of star formation. Indeed, they imposed that the instantaneous decrease of SFR was more than 80% and that the BIC difference was larger than 10. These criteria prevent a complete study of rapid variations in the SFH of galaxies as many of them would be missed. Furthermore, only decrease of SFR were considered and not the opposite, that is star formation bursts. Finally, their method is time consuming as one has to run the CIGALE code twice, once per SFH model considered, to perform the analysis. To go beyond this drawbacks and improve the method of Ciesla, Elbaz, Schreiber, et al., 2018, we consider in the present pilot study a statistical approach, the Approximate Bayesian Computation, combined with classification algorithm to improve both the accuracy and the efficiency of their method.

2.2.2. The sample

In this pilot work, we use the wealth of data available on the COSMOS field. The choice of this field is driven by the good spectral coverage of the data and the large statistics of sources available.

We draw a sample from the COSMOS catalog of C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a. A first cut is made to restrict ourselves to galaxies with a stellar mass C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a higher than $10^{8.5}\,\mathrm{M}_{\odot}$. Then, we restrict the sample to a relatively narrow range of redshift to minimize its impact on the SED and focus our method to the SFH effect on the SED. We thus select galaxies with redshift between 0.5 and 1, assuring sufficient statistics in our sample. We use the broad bands of the COSMOS catalog, listed in Table 2.1. For galaxies with

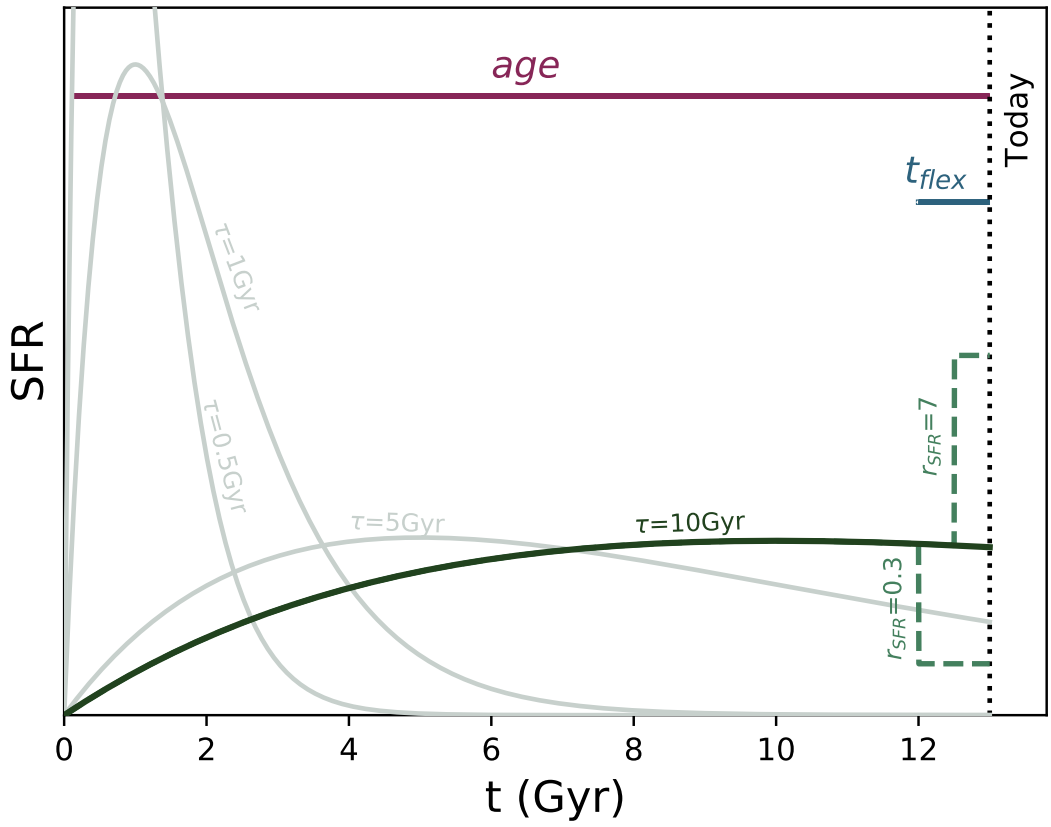

Figure 2.1. – Examples of delayed-τ SFHs considered in this work (star formation rate as a function of cosmic time). Different SFHs using τ_{main} =0.5, 1, 5, and 10 Gyr are shown to illustrate the impact of this parameter (light green and dark green solid lines). An example of delayed-τ SFH with flexibility is shown in solid dark green with the flexibility in green dashed lines for ($age_{\rm flex}$=1 Gyr & $r_{\rm SFR}$ =0.3) and ($age_{\rm flex}$=0.5 Gyr & $r_{\rm SFR}$ =7).

redshifts between 0.5 and 1, *Spitzer*/IRAC3 probes the 2.9-3.9 μm wavelength range rest frame and *Spitzer*/IRAC4 probes the 4-5.3 μm range rest frame. These wavelength ranges correspond to the transition between stellar and dust emission. To keep this pilot study simple we only consider the UV-to-NIR part of the spectrum, unaffected by dust emission.

One aspect of the ABC method that is still to be developed is how to handle missing data. In our astrophysical application, we identify several types of missing data. First there is the impact of redshifting that is the fact that a galaxy is undetected at wavelength shorter than the Lyman break at its redshift. Here, the absence of detection provides an information on the galaxy coded in its SED. Another type of missing data is linked to the definition of the photometric surveys: the spatial coverage is not exactly the same in every bands and the different sensitivity limits yields to undetected galaxies due to the faintness of their fluxes. To keep the statistical problem simple

Table 2.1. – COSMOS broad bands used in this work.

Instrument	**Band**	$\boldsymbol{\lambda}$ ($\boldsymbol{\mu}$**m**)
GALEX	FUV	0.153
GALEX	NUV	0.229
CFHT	u'	0.355
SUBARU	B	0.443
SUBARU	V	0.544
SUBARU	r	0.622
Suprime Cam	i'	0.767
Suprime Cam	z'	0.902
VISTA	Y	1.019
VISTA	J	1.250
VISTA	H	1.639
VISTA	Ks	2.142
Spitzer	IRAC1	3.6
Spitzer	IRAC2	4.5

in this pilot study, we remove galaxies that are not detected in all bands. This strong choice is motivated by the fact that the ABC method that we use in this pilot study has not been tested and calibrated in the case of missing data such as extragalactic field surveys can produce. The impact of missing data on this method would require an important work of statistical research which is beyond the scope of this paper.

As an additional constraint, we select galaxies with a SNR equal or greater than 10. However, given the importance of the NUV band Ciesla, Boselli, Elbaz, et al., 2016; Ciesla, Elbaz, Schreiber, et al., 2018 and the faintness of the fluxes compared to the other bands, we relax our criteria to a SNR of 5 for this band. The first motivation for this cut is again to keep our pilot study simple, but we show in Appendix A that indeed this SNR cut is relevant. In the following, we will consider a final sample composed of 12,380 galaxies for which the stellar mass distribution as a function of redshift is shown in Fig. 2.2 (top panel) and the distribution of the rejected sources in the bottom panel of the same figure.

The stellar mass distribution, from C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a, of the final sample is shown in Fig. 2.3. As a sanity check, we verify that above $10^{9.5}\,M_{\odot}$, the stellar mass, star formation rate, and specific star formation rate distributions are similar. Our selection criteria mostly affect low mass galaxies which is expected since we made SNR cuts.

Given the wide ranges of redshift, stellar masses, and SED shapes considered in our study, there is a normalization aspect that needs to be taken into account. Indeed, this diversity in galaxies' properties translates into a large distribution of fluxes in a given photometric band, spanning over several orders of magnitude: 8 orders of

magnitudes in the FUV band and 6 in the Ks band, for instance. This parameter space is very challenging for classification algorithms. To avoid this problem, we compute flux ratios. First we combine each flux with the closest one in terms of wavelength. This set of colors provides an information on the shape of the SED but effects of the SFH are also expected on wider scales in terms of wavelength. As discussed in Ciesla, Elbaz, Schreiber, et al., 2018, discrepancy between the UV and NIR emission assuming a smooth delayed-τ SFH is the signature that we are looking for indicating a possible change in the recent SFH. To be able to probe these effects, we also normalize each photometric band to the Ks flux and add this set of colors to the previous one. Finally, we set the flux ratios FUV/NUV and FUV/Ks to be 0 when $z > 0.68$ to account for the missing FUV flux density due to the Lyman break at these redshifts.

2.3. Statistical approach

In this section, we present the statistical approach that we will use to infer the most suitable SFH from photometric data. This new approach is applied to the sample described in Sect. 2.2.2 as a pilot study but can be applied to other datasets and to test other properties than the SFH.

2.3.1. Statistical modeling

As explained in the previous section, we want to distinguish between two SFH models: the first one is the smooth delayed-τ SFH, or SFH model $m = 0$, and the second is the same with a flexibility in the last 500 Myr, or SFH model $m = 1$, as presented in Sect. 2.2.1. The smooth delayed-τ SFH is thus a specific case of the flexible SFH obtained when there is no burst nor quenching ($r_{\rm SFR} = 1$).

Let $x_{\rm obs}$ denote the broadband data collected about a given galaxy. The statistical issue of deciding which SFH better fits the data can be seen as the Bayesian testing procedure distinguishing between both hypotheses

$$H_0 : r_{\rm SFR} = 1 \quad \text{vs} \quad H_1 : r_{\rm SFR} \neq 1.$$

The procedure will decide in favor of a possible change in the recent history when $r_{\rm SFR}$ is significantly different from 1 based on the data $x_{\rm obs}$. Conducting a Bayesian testing procedure based on the data $x_{\rm obs}$ of a given galaxy is exactly the same thing as the Bayesian model choice distinguishing between two nested statistical models C. Robert, 2007.

The first statistical model ($m = 0$), that is the delayed-τ SFH, is composed as follow: Let θ_0 denote the vector of all parameters necessary to compute the mock SED, denoted SED(θ_0). In particular θ_0 includes the parameters of the SFH. We denote $p(\theta_0|m = 0)$ the prior distribution over the parameter space for this statistical model. Likewise for the second SFH model : let $\theta_1 = (\theta_0, r_{\rm SFR}, t_{\rm flex})$ be the vector of all parameters for the delayed-τ+flex SFH. This vector includes the same parameters as for the previous SFH, plus two added parameters $r_{\rm SFR}$ and $t_{\rm flex}$. Let $p(\theta_1|m = 1)$ be the prior

distribution over the parameter space for the second model. We furthermore add a prior probability on the SFH index, $p(m=1)$ and $p(m=0)$, which are both 0.5 if we want to remain noninformative.

Finally, we assume a Gaussian noise. Thus, the likelihood $p(x_{\rm obs}|\theta_m, m)$ of θ_m given $x_{\rm obs}$ under the statistical model m is a multivariate Gaussian distribution, centered on $\mathrm{SED}(\theta_m)$ with a diagonal covariance matrix. The standard deviations are set to $0.1 \times \mathrm{SED}(\theta_m)$ because of the assumed value of SNR in the observations. In particular, it means that, up to constant, the loglikelihood is the negative χ^2-distance between the observed SED and the mock $\mathrm{SED}(\theta_m)$:

$$p(x_{\rm obs}|\theta_m, m) \propto \exp\left(-\frac{1}{2}\chi^2\Big(x_{\rm obs}, \mathrm{SED}(\theta_m)\Big)\right), \quad \text{where}$$
$$\chi^2\Big(x_{\rm obs}, \mathrm{SED}(\theta_m)\Big) = \sum_{j=1}^{J} \frac{\Big(x_{\rm obs}(\lambda_j) - \mathrm{SED}(\theta_m, \lambda_j)\Big)^2}{\Big(0.1\mathrm{SED}(\theta_m, \lambda_j)\Big)^2}. \tag{2.4}$$

2.3.2. Bayesian model choice

Bayesian model choice C. Robert, 2007 relies on the evaluation of the posterior probabilities $p(m|x_{\rm obs})$ which, using Bayes formula, is given by

$$p(m|x_{\rm obs}) = \frac{p(m)p(x_{\rm obs}|m)}{\sum_{m'} p(m')p(x_{\rm obs}|m')}, \tag{2.5}$$

where

$$p(x_{\rm obs}|m) = \int p(x_{\rm obs}|\theta_m, m)p(\theta_m|m)d\theta_m \tag{2.6}$$

is the likelihood integrated over the prior distribution of the m-th statistical model. Seen as a function of $x_{\rm obs}$, $p(x_{\rm obs}|m)$ is called the evidence or the integrated likelihood of the m-th model.

Bayesian model choice procedure innately embodies Occam's razor. This principle consists in choosing the simplest model as long as it is sufficient to explain the observation [2]. In this study, the two parametric SFHs are nested: when the parameter $r_{\rm SFR}$ of an SFH $m=1$ (flex + delayed-τ) is set to 1, we have an SFH that is also in the model $m=0$ (delayed-τ). Because of Occam's razor, if we choose the SFH with highest posterior probability when analyzing an observed SED $x_{\rm obs}$ that can be explained by

2. Indeed, the evidence $p(x_{\rm obs}|m)$ is a normalized probability density, that represents the distribution of datasets drawn from the m-th model, whatever the value of the parameter θ_m from its prior distribution. If models $m=0$ and $m=1$ are nested, the region of the data space of non-negligible probability under model $m=0$ has also a non-negligible probability under model $m=1$. Moreover, since model $m=1$ can fit to much more datasets, the probability density $p(x_{\rm obs}|m=1)$ is much more diffuse than the density $p(x_{\rm obs}|m=0)$. Hence, we expect for datasets x that can be explained by both models $m=0,1$ that $p(x|m=1) \leq p(x|m=0)$. If the prior probabilities $p(m=0)$ and $p(m=1)$ of both models are equal, it implies that, for datasets $x_{\rm obs}$ that can be explained by both models, $p(m=1|x) \leq p(m=0|x)$.

both SFHs, we choose the simplest model $m = 0$.

To analyse the dataset x_{obs}, it remains to compute the posterior probabilities. In our situation, the evidence of the statistical model m is intractable. It means that it cannot be easily evaluated numerically. Indeed, the function that computes SED(θ_m) given m and θ_m is fundamentally a black-box numerical function.

There are two methods to solve this problem. First, we can use a Laplace approximation of the integrated likelihood. The resulting procedure is the one that choose the SFH with the smallest Bayesian Information Criterion (BIC). Denoting $\widehat{\theta}_m$ the maximum likelihood estimate under the SFH m, χ^2 the non-reduced χ^2-distance of the fit, k_m the degree of freedom of model m, and n the number of observed photometric bands, the BIC of SFH m is given by

$$\begin{aligned} \text{BIC}_m &= -2 \max_{\theta_m} \ln p(x_{\text{obs}}|\theta_m, m) + k_m \times \ln(n), \\ &= \chi^2\Big(\text{SED}(\widehat{\theta}_m), x_{\text{obs}}\Big) + k_m \times \ln(n). \end{aligned} \tag{2.7}$$

Choosing the model with the smallest BIC is therefore an approximate method to find the model with the highest posterior probability. The results of Ciesla, Elbaz, Schreiber, et al. (2018) based on BIC are justified on this ground. But the Laplace approximation assumes that the number of observed photometric bands n is large enough. Moreover, determining the degree of freedom k_m of a statistical model can be a complex question. For all these reasons, we expect to improve the method of Ciesla, Elbaz, Schreiber, et al. (2018) based on BIC in the present paper.

Clever Monte Carlo algorithms to compute the evidence, Equation (2.6), of each statistical model will give us a much sharper approximation of the posterior probabilities of each SFH . We decided to rely on Approximate Bayesian Computation ABC, see e.g. Marin, Pudlo, Christian P Robert, et al., 2012; Sisson, Fan, and Beaumont, 2018 to compute $p(m|x_{\text{obs}})$.We could have considered other methods Vehtari, Ojanen, et al., 2012 such as bridge sampling, reversible jump MCMC, nested sampling, etc. But these methods require separate runs of the algorithm to analyze each galaxy, and probably more than a few minutes per galaxy. We expect to design a faster method here with ABC.

Finally,to interpret the results, we rely on the Bayes factor of the delayed-τ+flex SFH ($m = 1$) against the delayed-τSFH ($m = 0$)given by

$$\text{BF}_{1/0}(x_{\text{obs}}) = \frac{p(x_{\text{obs}}|1)}{p(x_{\text{obs}}|0)} = \frac{p(1|x_{\text{obs}})}{p(0|x_{\text{obs}})} = \frac{p(1|x_{\text{obs}})}{1 - p(1|x_{\text{obs}})}.$$

The computed value of the Bayes factor is compared to standard thresholds established by Jeffreys see, e.g., C. Robert, 2007 in order to evaluate the strength of the evidence in favor of delayed-τ+flex SFH if $\text{BF}_{1/0}(x_{\text{obs}}) \geq 1$. Depending on the value of the Bayes factor, Bayesian statisticians are used to say that the evidence in favor of model $m = 1$ is either *barely worth mentioning* (from 1 to $\sqrt{10}$) or *substantial* (from $\sqrt{10}$ to 10) or *strong* (from 10 to $10^{3/2}$) or *very strong* (from $10^{3/2}$ to 100) or *decisive* (larger than 100).

Table 2.2. – Basic ABC model choice algorithm that aims at computing the posterior probabilities of statistical models in competition to explain the data.

Input: - $x_{\rm obs}$, the observed SED we want to analyse
- $p(m)$, prior probability of the m-th statistical model
- $p(\theta_m|m)$, prior distribution of parameter θ_m of the m-th statistical model
- $p(x|\theta_m, m)$, probability density of a SED x given the m-th statistical model, and the parameter θ_m, see Eq. (2.4)
- N, number of simulations from the prior
- $S(x)$, a function that computes the summary statistics of a SED x

Output:
An approximation $\widehat{p}(m|x_{\rm obs})$ of the posterior probability of the m-th statistical model given the observed data for all m.

1 For $i = 1$ to N
2 Generate m^i from the prior $p(m)$
3 Generate θ_m^i from the prior $p(\theta_m|m)$
4 Generate x^i from the model $p(x|\theta_m, m)$
5 Compute $S(x^i)$ and store $(m^i, \theta_m^i, S(x^i))$
6 End For
7 Compute $\widehat{p}(m|x_{\rm obs})$ with Eq. (2.8) for all m

---

2.3.3. The Approximate Bayesian Computation method

To avoid the difficult computation of the evidence, Equation (2.6), of model m and get a direct approximation of $p(m|x_{\rm obs})$, we resort to the family of methods named ABC model choice Marin, Pudlo, Estoup, et al., 2018.

The main idea behind the ABC framework is that we can avoid the evaluation of the likelihood and directly estimate a posterior probability by relying on N random simulations (m^i, θ_m^i, x^i), $i = 1, \ldots, N$ from the joint distribution $p(m)p(\theta_m|m)p(x|\theta_m, m)$. Here simulated (m^i, θ_m^i, x^i) are obtained as follow: first, we draw a SFH m^i at random, with the prior probability $p(m^i)$; then we draw θ_m^i according to the prior $p(\theta_m^i|m^i)$; finally we compute the mock SED(θ_m^i) with CIGALE and add a Gaussian noise to the mock SED to get x^i. This last step is equivalent to sampling from $p(x^i|\theta_m^i, m^i)$ given in (2.4). Basically, the posterior distribution $p(m|x_{\rm obs})$ can be approximated by the frequency of SFH m among the simulations close enough to $x_{\rm obs}$.

To measure how close x is from $x_{\rm obs}$, we introduce the distance between vectors of summary statistics $d\big(S(x), S(x_{\rm obs})\big)$ and we set a threshold ε: simulations (m, θ_m, x) that satisfy $d\big(S(x), S(x_{\rm obs})\big) \leq \varepsilon$ are considered "close enough" to $x_{\rm obs}$. The summary statistics $S(x)$ are primarily introduced as a way to handle feature extraction, whether it is for dimensionality reduction or for data normalization. For the present study, the components of the vector $S(x)$ are flux ratios from the SED x, chosen for normalization

purposes. Mathematically speaking, $p(m = 1|x_{\text{obs}})$ ies thus approximated by

$$\widehat{p}(m|x_{\text{obs}}) = \frac{\sum_{i=1}^{N} \mathbf{1}\{m^i = m\}\mathbf{1}\Big\{d\big(S(x^i), S(x_{\text{obs}})\big) \leq \varepsilon\Big\}}{\sum_{i=1}^{N} \mathbf{1}\Big\{d\big(S(x^i), S(x_{\text{obs}})\big) \leq \varepsilon\Big\}}. \tag{2.8}$$

The resulting algorithm, named basic ABC model choice, is given in Table 2.2. Finally, note that, if k is the number of simulations close enough to x_{obs}, the last step of Table 2.2 can be seen as a k-nearest neighbor (k-nn) method predicting m based on the features (or covariates) $S(x)$.

The k-nn can be replaced by other machine learning algorithms to obtain sharper results. Indeed, the k-nn is known to perform poorly when the dimension of $S(x)$ is larger than 4. For instance, Pudlo, Marin, Estoup, et al. (2016) decided to rely on the method called Random Forest Breiman, 2001. The machine learning based ABC algorithm is given in Table 2.3. All machine learning models given below are classification methods. In our context, they aim at separating the simulated datasets x depending on the SFH ($m = 0$ or 1) that was used to generate them. The machine learning model is fitted on the catalog of simulations (m^i, θ_m^i, x^i), that is to say, it learns how to predict m based on the value of x. To this purpose, we fit a function $\widehat{p}(m = 1|x)$ and perform the classification task on a new dataset x' by comparing the fitted $\widehat{p}(m = 1|x')$ to 1/2: if $\widehat{p}(m = 1|x') > 1/2$, the dataset x' is classified as generated by SFH $m = 1$; otherwise, it is classified as generated by SFH $m = 0$. The function $\widehat{p}(m = 1|x')$ depends on some internal parameters not explicitly shown in the notation. For example, this function can be computed with the help of a neural network. A neuron here is a mathematical function that receives inputs and produces an output based on a weighted combination of the inputs; each neuron processes the received data and transmits its output downstream in the network. Generally, the internal parameters (ϕ, ψ) are of two kinds: the coordinates of ϕ are optimized on data with a specific algorithm, and the coordinates of ψ are called tuning parameters (or hyperparameters). For instance, with neural networks ψ represents the architecture of the network and the amount of dropout; ϕ represents the collection of the weights in the network.

The gold standard machine learning practice is to split the catalog of data into three parts: the training catalog and the validation catalog, that are both used to fit the machine learning models, and the test catalog that is used to compare the algorithms fairly and get a measure of the error committed by the models. Actually each fit requires two catalogs (training and validation) because modern machine learning models are fitted to the data with a two step procedure. We detail the procedure for a simple dense neural network and refer to Appendix B for the general case. The hyperparameters we consider are the number of hidden layers, the number of nodes in each layers, and the amount of dropout. We fix a range of possible values for each hyperparameters (see table 2.5). We select a possible combination of hyperparameters ψ, and train the obtained neural network on the training catalog. Once the weights ϕ

Table 2.3. – Machine learning based ABC model choice algorithm that aims at computing the posterior probability of two statistical models in competition to explain the data

---

Input and output: same as Table 2.2

1 Generate N simulations (m^i, θ_m^i, x^i) from the joint distribution $p(m)\,p(\theta_m|m)\,p(x|\theta_m, m)$

2 Summarize all simulated datasets (photometric SED) x^i with $S(x^i)$ and store all simulated $(m^i, \theta_m^i, S(x^i))$ into a large catalog

3 Split the catalog into three parts: training, validation and test catalogs

4 Fit each machine learning method on the training and validation catalogs to approximate $p(m = 1|S(x))$ with $\widehat{p}_{\widehat{\psi}}(m = 1|x)$

5 Choose the best machine learning method by comparing their classification errors on the test catalog

6 Return the approximation $\widehat{p}(m = 1|x_{\text{obs}})$ computed with the best method

---

are optimized on the training catalog, we evaluate the given neural network on the validation catalog and associate the obtained classification error with the combination of hyperparameters used. We follow the same training and evaluating procedure for several hyperparameters combinations ψ, and we select the one obtaining the lowest classification error. At the end of the process, we evaluate the classification error on the test catalog using the selected combination of hyperparameters $\widehat{\psi}$

The test catalog is willingly left out during the training and the tuning of the machine learning methods. Indeed, the comparison of the accuracy of the approximation returned by each machine learning method on the test catalog ensures a fair comparison between the methods, on data unseen during the fit of $\widehat{p}_{\widehat{\psi}}(m|x)$.

In this pilot study, we tried different machine learning methods and compared their accuracy:

— logistic regression and linear discriminant analysis Friedman, Hastie, and Tibshirani, 2001, that are almost equivalent linear models, and serve only as baseline methods,

— neural networks with 1 or 3 hidden layers, the core of deep learning methods, that have proved to get sharp results on various signal datasets (images, sounds)

— classification tree boosting with XGBoost, see Chen and Guestrin, 2016, which is considered as state-of-the-art methods in many applied situations, and is often the most accurate algorithm when correctly calibrated on a large catalog.

We did not try Random Forest since it cannot be run on a simulation catalog of size as large as the one we are relying on in this pilot study ($N = 4\times10^6$). Indeed the motivation of the proposed methodology is to bypass the heavy computational burden of MCMC based algorithms to perform statistical model choice. In this study, Random Forest was not able to fulfill this aim unlike the classification methods given above.

Table 2.4. – Prior range of the parameters used to generate the simulation table of SEDs with redshift between 0.5 and 1.

Parameter	Value
Delayed-τ SFH	
age (Gyr)	$[0.5; 9]$
τ_{main} (Gyr)	$[0.1; 10]$
Flexible delayed-τ SFH	
age (Gyr)	$[0.5; 9]$
τ_{main} (Gyr)	$[0.1; 10]$
age_{flex} (Myr)	10, 100, 450
$\log r_{\mathrm{SFR}}$	[-6; 6]
Dust attenuation	
A_V	$[0.1; 4]$

2.3.4. Building synthetic photometric data

To compute or fit galaxies' SEDs with CIGALE, one has to provide a list of prior values for each model's parameters. The comprehensive module selection in CIGALE allows to specify entirely the SFH and how the mock SED is computed. The list of prior values for each module's parameters specifies the prior distribution $p(\theta_m|m)$. CIGALE uses this list of values or ranges to sample from the prior distribution by picking values on θ_m on a regular grid. This has the inconvenient of: being very sensitive to the number of parameters (if d is the number of parameters, and if we assume 10 different values for each parameter, the size of the grid is 10^d); producing simulations that are generated with some parameters that are equals. Instead, in this study, we advocate in favor of drawing values of all parameters at random from the prior distribution, which is uniform over the specified ranges or list of values. The ranges for each model parameters are chosen to be consistent with those used by Ciesla, Elbaz, Schreiber, et al., 2018. In particular, the catalog of simulations drawn at line 1 in Table 2.3 follow this rule. Each SFH (the simple delayed-τ or the delayed-τ + flex) is then convolved with the stellar population models of Bruzual and S. Charlot, 2003. The attenuation law described in S. Charlot and Fall, 2000 is then applied to the SED. Finally CIGALE convolves each mock SED into a COSMOS-like set of filters described in Table 2.1.

2.4. Application to synthetic photometric data

We first applied our methodology on simulated photometric data to evaluate its accuracy. The main interest of such synthetic data is that we control all parameters (flux densities, colors, physical parameters). The whole catalog of simulations was composed of 4×10^6 simulated datasets. We split this catalog at random into three

Table 2.5. – Calibration and test of machine learning methods

Method	Tuning parameter	Explored range	Best value	Error rate (%)
Logistic regression	$\emptyset$			30.27
Linear Discriminant Analysis	$\emptyset$			30.43
k-nearest neighbors	k	[3600, 180000]	5000	23.79
1-layer neural network	dropout	[0.1, 0.5]	0.2	22.51
	nodes in each layer	[16, 256]	128	—
3-layer neural network	dropout	[0.1, 0.5]	0.2	21.06
	nodes in each layer	[16, 256]	128	—
Tree boosting (XGBoost)	number of trees (nround)	[100, 1000]	400	20.98
	depth of each tree (max_depth)	[4, 15]	12	—
	learning rate (eta)	[0.01, 0.2]	0.1	—

The best value of each tuning parameter was found by comparing error rates on the validation catalog.
The error rate given in the last column is computed on the test catalog.

parts, as explained in Sect. 2.3.3, and add an extra catalog for comparison with CIGALE:

— 3.6×10^6 sources (90%) to compose the training catalog,

— 200,000 sources (5%) to compose the validation catalog,

— 200,000 sources (5%) to compose the test catalog,

— 30,000 additional sources to compose the extra catalog for comparison with CIGALE.

The size of the extra catalog is much smaller to limit the amount of computation time required by CIGALE to run its own algorithm of SED fitting.

2.4.1. Calibration and evaluation of the machine learning methods on the simulated catalogs

In this section, we present the calibration of the machine learning techniques and their error rates on the test catalog. We then try to interpret the results given by our methodology.

As described in Sect. 2.3.3, we trained and calibrated the machine learning methods on the training and validation catalog. The results are given in Table 2.5. Neither Logistic regression nor Linear Discriminant Analysis have tuning parameters that need to be calibrated on the validation catalog. The error rate of these techniques are about 30% on the test catalog. But the modern machine learning methods (k-nearest neighbors, neural networks and tree boosting) have been calibrated on the validation catalog. The best value of the explored range for ψ were found by comparing error rates on the validation catalog and are given in Table 2.5. The error rates of these

methods on the test catalog vary between 24% and 20%. Thus, it is clear that there is a significant gain to use non-linear methods. But we see no obvious use in training a more complex algorithm (such as a deeper neural network) for this problem, although it could become useful when increasing the number of photometric bands and the redshift range. Finally, we favor XGBoost for our study. Indeed, while neural networks could probably be tuned more precisely to match or exceed its performances, we find XGBoost easier to tune and to interpret.

Machine learning techniques that fit $\widehat{p}_{\widehat{\psi}}(m|x)$ are often affected by some bias and may require some correction Niculescu-Mizil and Caruana, 2012.Such classification algorithms compare the estimated probabilities of m given x and return the most likely m given x. The output m can be correct even if the probabilities are biased towards 0 for small probabilities or towards 1 for large probabilities. A standard reliability check shows no such problem for our XGBoost classifier. To this aim, the test catalog is divided into 10 bins: the first bin is composed of simulations with a predicted probability $\widehat{p}(m=1|x_{\rm obs})$ between 0 and 0.1, the second with $\widehat{p}(m=1|x_{\rm obs})$ between 0.1 and 0.2... The reliability check procedure ensures that the frequency of the SFH $m=1$ among the k-th bin falls within the range $[(k-1)/10; k/10]$, because the $\widehat{p}(m=1|x_{\rm obs})$ predicted by XGBoost are between $(k-1)/10$ and $k/10$.

We studied the ability of our methodology to distinguish the SFH of the simulated sources of the test catalog. The top panel of Fig. 2.4 shows the distribution of $\widehat{p}(m=1|x_{\rm obs})$ when x varies in the test catalog. Naively, a perfect result would have half of the sample with $p=1$ and the other half with $p=0$. In fact, when $m=0$, the SFH $m=1$ is also suitable since the models are nested. In this case, Occam's razor favors the model $m=0$, and $\widehat{p}(m=1|x_{\rm obs})$ must be less than 0.5, see Sect. 2.3.2. On the contrary, for the SEDs solely explained by the SFH model $m=1$, $\widehat{p}(m=1|x_{\rm obs})$ is close to 1.

The distribution (Fig. 2.4, bottom left panel) has two peaks, one centered around $p=0.2$ and one between 0.97 and 1. This peak at 0.2, and not 0, is expected when one of the model proposed to the choice is included in the second model. In the distribution of the $\widehat{p}(m=1|x_{\rm obs})$, 20% of the sources have a value higher than 0.97 and 52% lower than 0.4. In the right panels of Fig. 2.4, we show the distribution of $r_{\rm SFR}$ for the galaxies x with $\widehat{p}(m=1|x_{\rm obs})>0.97$. With a perfect method, galaxies with $r_{\rm SFR}\neq 1$ should have $\widehat{p}(m=1|x_{\rm obs})=1$. Here we see indeed a deficit of galaxies around $p=1$, however the range of affected $r_{\rm SFR}$ goes from 0.1 to 10. Therefore, the method is not able to identify galaxies having a variability of its SFR if this variability is only 0.1 to 10 times the SFR before the variability began. In other words, the method is sensitive to $|\log r_{\rm SFR}|>1$. This is confirmed by the distribution of $r_{\rm SFR}$ for galaxies with $p<0.40$ (Fig. 2.4, bottom panel). However, there are sources with a $|\log r_{\rm SFR}|>1$ associated with low values of $\widehat{p}(m=1|x_{\rm obs})$. The complete distribution of $r_{\rm SFR}$ as a function of $\widehat{p}(m=1|x_{\rm obs})$ is shown in Fig. 2.4.

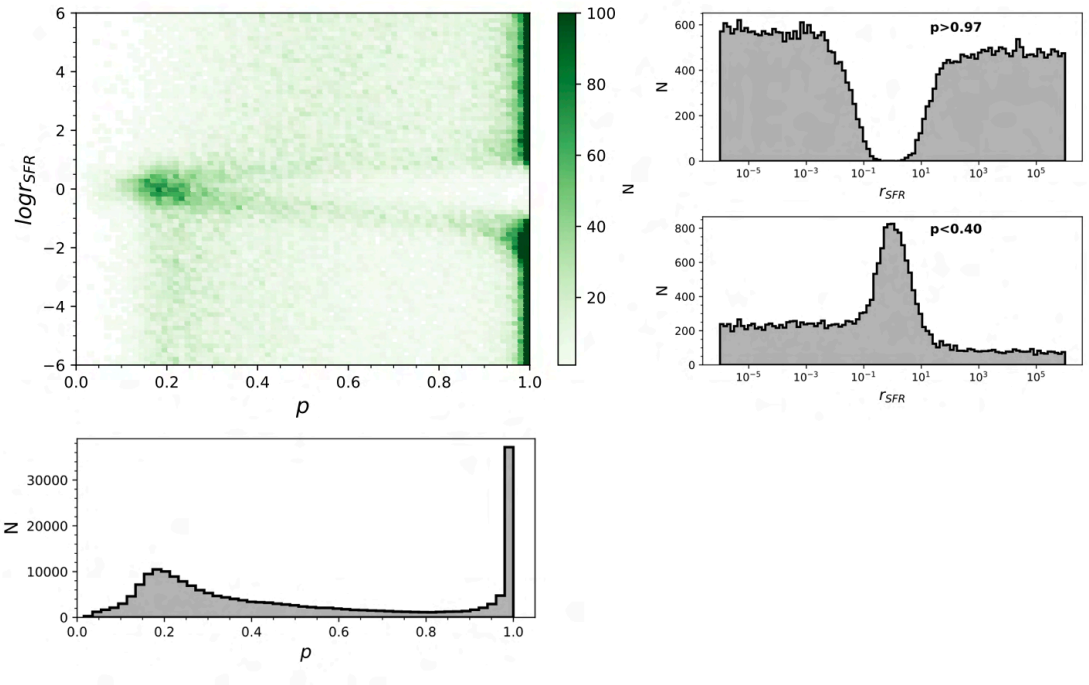

Figure 2.4. – Study of the statistical power of $\widehat{p}(m = 1|x_{\rm obs})$ to detect short-term variations with respect to the value of $r_{\rm SFR}$. **Top left panel:** Joint distribution of $\widehat{p}(m = 1|x_{\rm obs})$ and $r_{\rm SFR}$. **Bottom left panel:** Distribution of $\widehat{p}(m = 1|x_{\rm obs})$ obtained with x coming from the test catalog. **Right panels:** Marginal distributions of $r_{\rm SFR}$ for mock sources with $\widehat{p}(m = 1|x_{\rm obs}) > 0.97$ (top right panel) and for mock sources with $\widehat{p}(m = 1|x_{\rm obs}) < 0.4$ (bottom right panel).

2.4.2. Importance of particular flux ratios

We try to find which part of the dataset x influences the most on the choice of SFH given by our method. The analyse of x relies entirely on the summary statistics $S(x)$, the flux ratios. Hence, we tried to understand which flux ratios are most discriminant for the model choice. We wanted to check that the method is not based on a bias of our simulations and wanted to assess which part of the data could be removed without losing crucial information.

We use different usual metrics e.g. Friedman, Hastie, and Tibshirani, 2001; Chen and Guestrin, 2016 to assess the importance of each flux ratio in the machine learning estimation of $\widehat{p}(m = 1|x)$. Those metrics are used as indicators of the relevance of each flux ratio for the classification task. As expected, the most important flux ratios for our problem involve the bands at shortest wavelength (FUV at $z < 0.68$ and NUV above, as FUV is no longer available), normalized by either Ks or u. This is expected as these bands are known to be sensitive to SFH e.g., Arnouts, Le Floc'h, Chevallard, et al., 2013. We see no particular pattern in the estimated importance of the other flux ratios. They are all used for the classification and removing any of them decreases classification

Table 2.6. – Input parameters used in the SED fitting procedures with CIGALE.

Parameter	Value
delayed-τ SFH	
age (Gyr)	[0.5;9], 15 values linearly sampled
τ_{main} (Gyr)	[0.1;10], 15 values linearly sampled
Flexible delayed-τ SFH	
age (Gyr)	[0.5;9], 15 values linearly sampled
τ_{main} (Gyr)	[0.1;10], 15 values linearly sampled
age_{flex} (Myr)	10, 100, 450
$\log r_{\mathrm{SFR}}$	[−6;6], 12 values linearly sampled
Dust attenuation	
A_V^{ISM}	[0.1;4], 10 values linearly sampled

accuracy, except for IRAC1/Ks whose importance is consistently negligible across every considered metric.

We also test if the UVJ selection used to classify galaxies according to their star formation activity e.g., Stijn Wuyts, Labbé, Franx, et al., 2007; Williams, Quadri, Franx, et al., 2009 is able to probe the kind of rapid and recent SFH variations we are investigating in this study. We train an XGBoost classification model using only u/V and V/J in order to evaluate the benefits of using all available flux ratios. This results in a severe increase in classification error, going from 21.0% using every flux ratios to 35.8%.

2.4.3. Comparison with SED fitting methods based on BIC

In this section, we compare the results obtained with the ABC method to those obtained with a standard SED modeling. The goal of this test is to understand and quantify the improvement that the ABC method brings in terms of accuracy of the results. We use the simulated catalog of 30,000 sources, described in the beginning of this section, for which we control all parameters.

The ABC method is also used on this extra catalog. This test is very similar to the training procedure described in Sect. 2.4.1. Indeed, with this extra catalog, the ABC method has an error rate of 21.2% compared to 21.0% with the previous test sample.

CIGALE is run on the test catalog as well. The set of modules is the same as those used to create the mock SEDs, however the parameters used to fit the test catalog do not include the input parameters that were randomly chosen. This test is intentionally thought to be simple and represent an ideal case scenario. The error rate that will be obtained with CIGALE will therefore represent the best result achievable.

To decide whether a flexible SFH was preferable to a normal delayed-τ SFH using CIGALE, we adopt on whether a flexible SFH is preferred to a normal delayed-τ SFH, we adopt the method of Ciesla, Elbaz, Schreiber, et al., 2018 described in Sect. 2.2.1.

The quality of fit using each SFH is tested through the use of the Bayesian Information Criterion (BIC).

In detail, the method that we use is the following: First, we make a run with CIGALE using a simple delayed-τ SFH which parameters are presented in Table 2.6. A second run is then performed with the flexible SFH. We compare the results and quality of the fits using one SFH or the other. The two models have different number of degrees of freedom. To take this into account, we compute the BIC presented in Sect. 2.3.2 for each SFH.

We then calculate the difference between $\mathrm{BIC}_{delayed}$ and BIC_{flex} (ΔBIC) and use the threshold defined by Jeffreys (Sect. 2.3.2) valid either for the BF and the BIC and also used in Ciesla, Elbaz, Schreiber, et al., 2018: a ΔBIC larger than 10 is interpreted as a strong difference between the two fits Kass and Raftery, 1995, with the flexible SFH providing a better fit of the data than the delayed-τ SFH.

We apply this method to the sample containing 15k sources modeled with a delayed-τ SFH and 15k modeled using a delayed-τ+flexibility. With this criteria, we find that the error rate of CIGALE, in terms of identifying SEDs built with a delayed-τ+flex SFH, is 32.5%. This rate depends on the ΔBIC threshold chosen and increases with the value of the threshold as shown in Fig. 2.5. The best value, 28.7%, is lower than the error rate obtained from a logistic regression or a LDA (see Table 2.5) but significantly higher than the error rate obtained from our procedure using XGBoost (21.0%) In this best case scenario test for CIGALE, a difference of 7.7% is substantial and implies that the ABC method tested in this study provides better results than a more traditional one using SED fitting. When considering sources with ΔBIC>10, i.e. sources for which the method using CIGALE estimates that there is a strong evidence for the flexible SFH, 95.4% are indeed SEDs simulated with the flexible SFH. Using our procedure with XGBoost, and the Bayes factor corresponding threshold of 150 Kass and Raftery, 1995, we find that 99.7% of the sources' SFH are correctly identified. The ABC method provides a cleaner sample than the CIGALE ΔBIC based method.

2.5. Application on COSMOS data

We now apply our method to the sample of galaxies drawn from the COSMOS catalog, which selection is described in Sect. 2.2.2. As a result, we show the $\widehat{p}(m = 1|x_{\mathrm{obs}})$ distribution obtained for this sample of observed galaxies in Fig. 2.6. We remind that the 0 value indicates that the delayed-τ SFH is preferred whereas $\widehat{p} = 1$ indicates that the flexible SFH is more adapted to fit the SED of the galaxy. As a guide, we indicate the different grades of the Jeffreys scale and provide the number of sources in each grade in Table 2.7. The flexible SFH better models the observations of 16.4% of our sample than the delayed-τ SFH. However, it also means that for most of the dataset (83.6%), there is no strong evidence for the necessity to increase the complexity of the SFH, a delayed-τ is sufficient to model the SED of these sources.

To investigate the possible differences in terms of physical properties of galaxies according to their Jeffreys grade, we divide the sample of galaxies in two groups. The

Table 2.7. – Jeffreys scale and statistics of our sample.

Grade	Evidence against delayed-τ SFH	Number	%
1	Barely worth mentioning	1,187	9.6
2	Substantial	466	3.8
3	Strong	209	1.7
4	Very strong	90	0.7
5	Decisive	77	0.6

first group corresponds to galaxies with $\widehat{p}(m = 1|x_{\rm obs}) < 0.5$, galaxies for which there is no evidence for the need of a recent burst or quenching in the SFH, a delayed-τ SFH is sufficient to model the SED of these sources. We select the galaxies of the second group imposing $\widehat{p}(m = 1|x_{\rm obs}) > 0.75$, i.e. Jeffreys scale grades of 3, 4, or 5: from strong to decisive evidence against the normal delayed-τ. In Fig. 2.7 (top panel), we show the stellar mass distribution of both subsamples. Although the stellar masses obtained with either the smooth delayed-τ or the flexible SFH are consistent with each other, for each galaxies we use the most suitable stellar mass: if the galaxy has $\widehat{p}(m = 1|x_{\rm obs}) < 0.5$ the stellar mass obtained from the delayed-τ SFH is used, and if the galaxy has $\widehat{p}(m = 1|x_{\rm obs}) > 0.75$ the stellar mass obtained with the flexible SFH is used. The stellar mass distribution of galaxies with a delayed-τ SFH is similar to the distribution of the whole sample, as shown in the middle panel of Fig. 2.7. However, the stellar mass distribution of galaxies needing a flexibility in their recent SFH shows a deficit of galaxies with stellar masses between $10^{9.5}$ and $10^{10.5}\,{\rm M}_{\odot}$ compared to the distribution of the full sample. We note that at masses larger than $10^{10.5}\,{\rm M}_{\odot}$ the distribution are identical, despite a small peak at $10^{11.1}\,{\rm M}_{\odot}$. To check if this results is not due to our SED modeling procedure and the assumptions we adopted, we show in the middle panel of Fig. 2.7 the same stellar mass distributions using this time the values published by C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a. The two stellar mass distributions, with the one of galaxies with $\widehat{p}(m = 1|x_{\rm obs}) > 0.75$ peaking at a lower mass, are recovered. This implies that these differences between the distributions are independent from the SED fitting method employed to determine the stellar mass of the galaxies. We note that when the algorithm has been trained, only ratios of fluxes were provided to remove the normalization factor out of the method and the mock SEDs from which the flux ratios were computed were all normalized to $1\,{\rm M}_{\odot}$. The stellar mass is at first order a normalization through, for instance, the L_K-M_* relation e.g., Gavazzi, Pierini, and Boselli, 1996. Using flux ratios, the algorithm had no information linked to the stellar mass of the mock galaxies. Nevertheless, applied to real galaxies the result of our procedure yields two different stellar mass distributions between galaxies identified as having smooth SFH and galaxies undergoing a more drastic episode (star formation burst or quenching).

In the bottom panel of Fig. 2.7, we show the distribution in specific star formation rate (sSFR, sSFR ≡ SFR/M_*) for the same two samples. The distribution of galaxies with $\widehat{p}(m = 1|x_{\rm obs}) < 0.5$ is narrow ($\sigma = 0.39$) and has one peak at $\log$sSFR $= -0.32$ (Gyr^{-1}), clearly showing the MS of star forming galaxies. Galaxies with high probability to have a recent strong variation of their SFH form a double-peaked distribution with one peak above the MS formed by galaxies with $\widehat{p}(m = 1|x_{\rm obs}) > 0.75$ ($\log$sSFR $= 0.66$), corresponding to galaxies having experienced a recent burst, and a second peak at lower sSFRs than the MS, corresponding to sources having undergone a recent decrease of their star formation activity ($\log$sSFR $= -1.38$). In the sample of galaxies with $\widehat{p}(m = 1|x_{\rm obs}) > 0.75$, 28% of these sources are in the peak of galaxies experiencing a burst of star formation activity and 72% seem to undergo a rapid and drastic decrease of their SFR. One possibility to explain this assymetry could be a bias produced by the algorithm, as shown in Fig. 2.4, more sources with $\widehat{p}(m = 1|x_{\rm obs}) > 0.97$ tend to be associated with low values of $r_{\rm SFR}$ than to $r_{\rm SFR} > 1$. However, in the case of the extra catalog, this disparity is 47% and 53% for high and low $r_{\rm SFR}$, respectively.

The distribution of the two samples in terms of sSFR indicates that, to be able to reach the sSFR of galaxies that are outside the MS, one had to take into account a flexibility in the SFH of galaxies when performing the SED modeling. This is needed to recover as much as possible the parameter space in SFR and M_*.

2.6. Conclusions

In this pilot study, we have proposed to use a state-of-the-art statistical method using machine learning algorithm, the Approximate Bayesian Computation, to determine the best-suited SFH to be used to measure the physical properties of a subsample of COSMOS galaxies. These galaxies have been selected in mass ($\log M_*$ >8.5) and in redshift ($0.5 < z < 1$). Furthermore, we impose that the galaxies should be detected in all UV-to-NIR bands with a SNR higher than 10. We verified that these criteria do not bias the sSFR distribution of the sample.

To model these galaxies, we considered a smooth delayed-τ SFH with or without a rapid and drastic change in the recent SFH, that is in the last few hundreds Myr. We have built a mock galaxies SED using the SED fitting code CIGALE. The mock SEDs have been integrated into the COSMOS set of broad band filters. To avoid large dynamical ranges of fluxes which is to be avoided when using classification algorithms, we compute flux ratios.

Different classification algorithms have been tested with XGBoost providing the best results with a classification error of 20.98%. As output, the algorithm provides the probability that a galaxy is better modeled using a flexibility in the recent SFH. The method is sensitive to variations of SFR that are larger than 1 dex.

We have compared the results from the ABC new method with SED fitting using CIGALE. Following the method proposed by Ciesla, Elbaz, Schreiber, et al., 2018, we compare the results of two SED fits, one using the delayed-τ SFH and the other one adding a flexibility in the recent history of the galaxy. The Bayesian Information

Criterion are computed and compared to determine which SFH provides a better fit. The BIC method provides a high error rate, 28%, compared to the 21% obtained with the ABC method. Moreover, since the BIC method requires two SED fits per analyze of a source, it is much slower than the proposed ABC method: we were not able to compare them on the test catalog of 200,000 sources and we had to introduce a smaller simulated catalog of size 30,000 to compute their BIC in a reasonable amount time.

We use the result of the ABC method to determine the stellar mass and SFRs of the galaxies using the best-suited SFH for each of them. We compare two samples of galaxies: the first one is galaxies with $\widehat{p}(m = 1|x_{\text{obs}}) < 0.5$, that are galaxies for which the smooth delayed-τ SFH is preferred, the second one is galaxies with $\widehat{p}(m = 1|x_{\text{obs}}) > 0.75$, galaxies for which there is a strong to decisive evidence against the smooth delayed-τ SFH. The stellar mass distribution of these two samples is different. The mass distribution of galaxies for which the delayed-τ SFH is preferred is similar to the distribution of the whole sample. However, the mass distribution of galaxies needing a flexible SFH shows a deficit between $10^{9.5}$ and $10^{10.5}\,\mathrm{M}_{\odot}$. Their distribution is however similar to the whole sample's above $\mathrm{M}_{*} = 10^{10.5}\,\mathrm{M}_{\odot}$. Furthermore, the results of this study also implies that a flexible SFH is needed to cover the largest parameter space in terms of stellar mass and SFR possible, as seen from the sSFR distributions of galaxies with $\widehat{p}(m = 1|x_{\text{obs}}) > 0.75$.

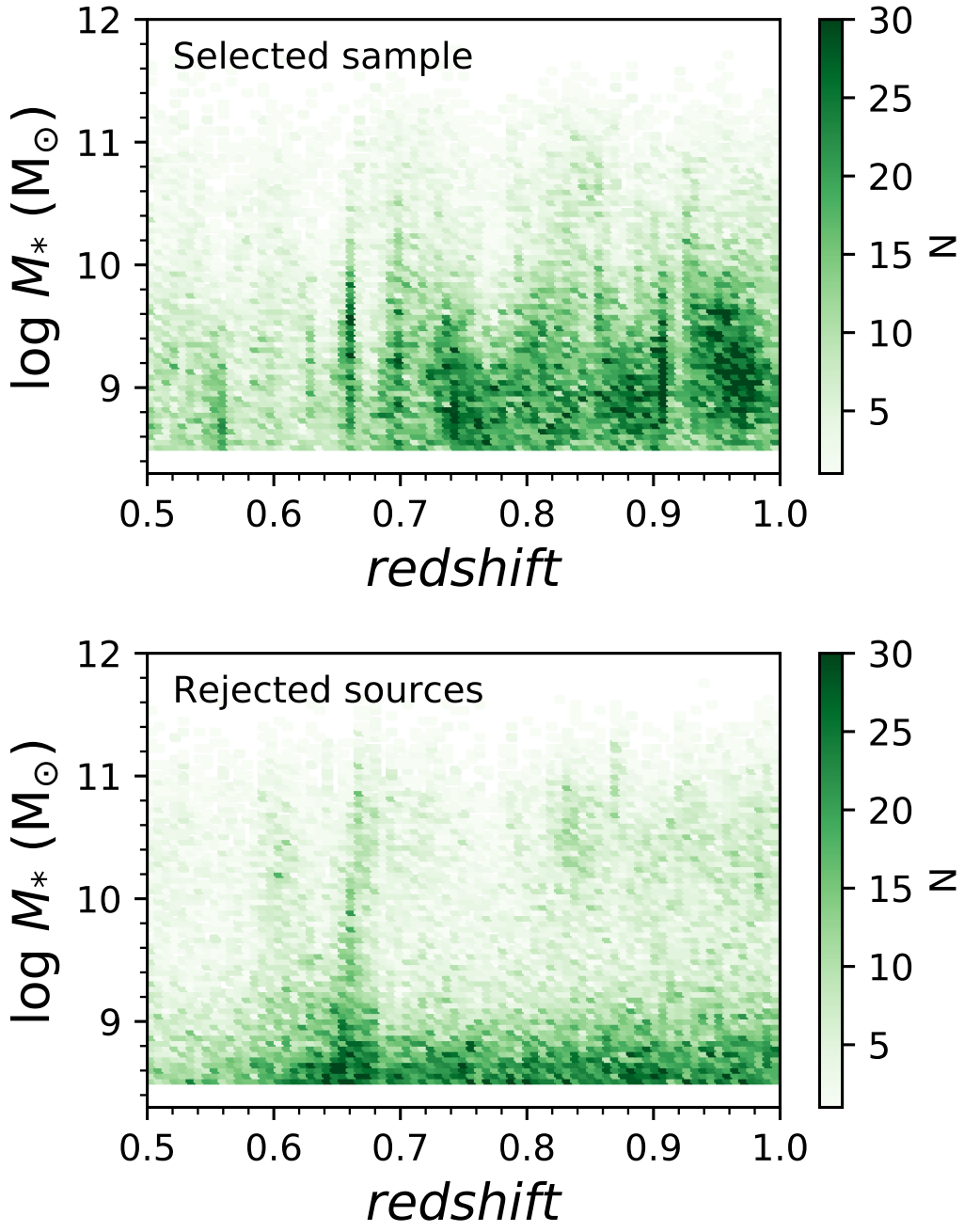

Figure 2.2. – Stellar mass from C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a as a function of redshift for the final sample (**top panel**) and for the rejected galaxies following our criteria (**bottom panel**).

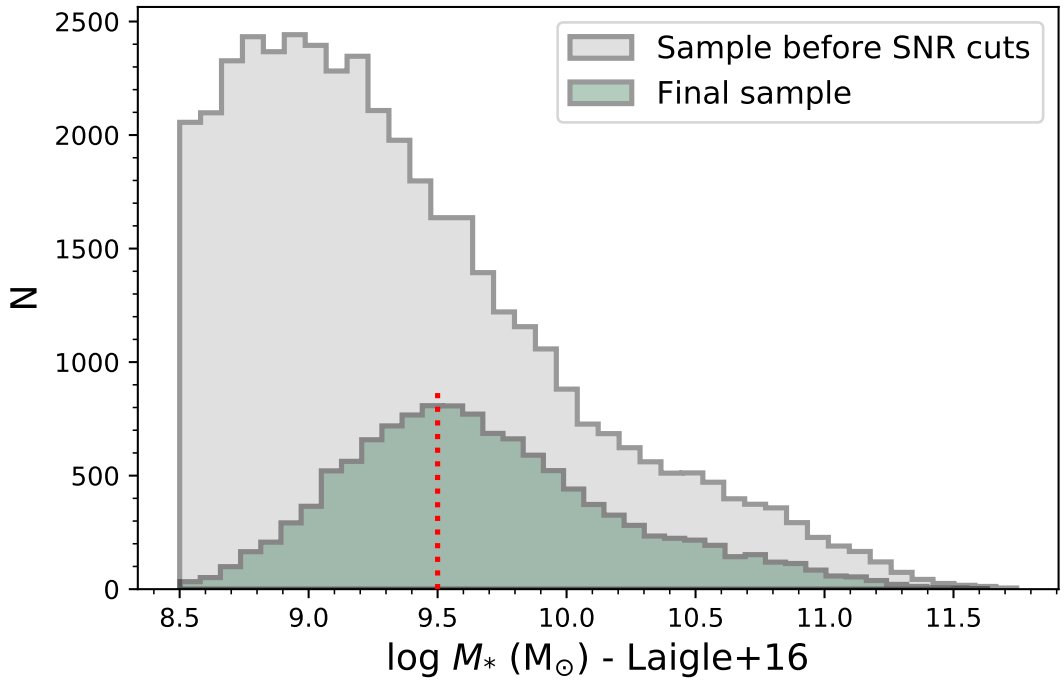

Figure 2.3. – Distribution of stellar mass for the sample before the SNR cut (grey) and the final sample (green). The red dotted line indicated the limit above which our final sample is considered as complete. The stellar masses indicated here are from C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a.

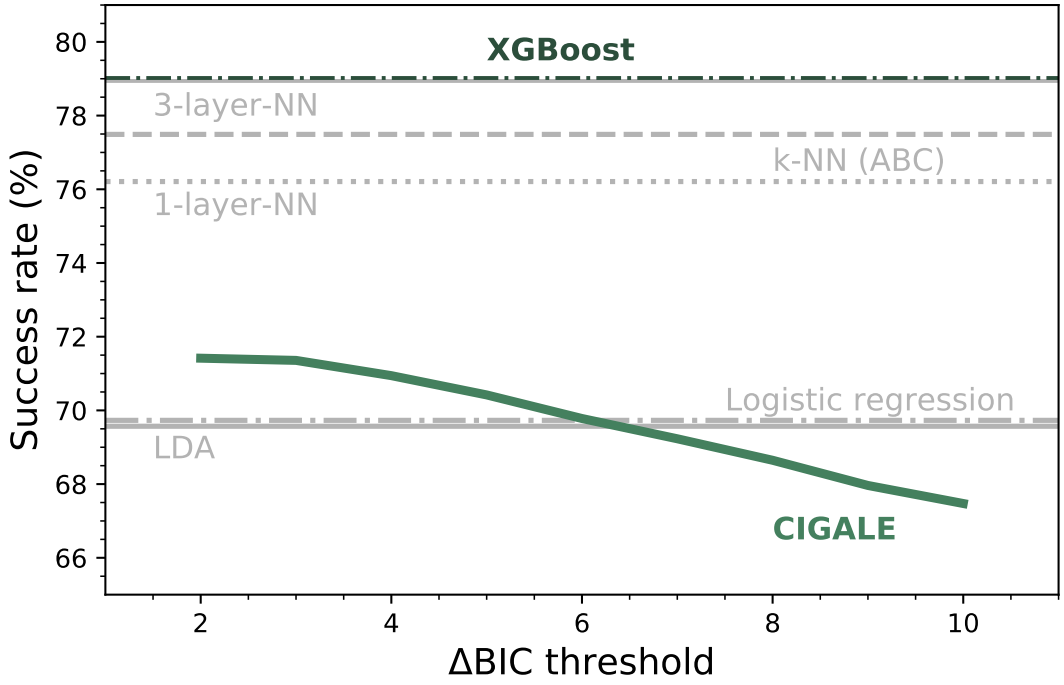

Figure 2.5. – Error rate obtained with CIGALE as a funtion of ΔBIC chosen threshold. For comparison we show the error rates obtained by the classification methods tested in Sect. 2.3.

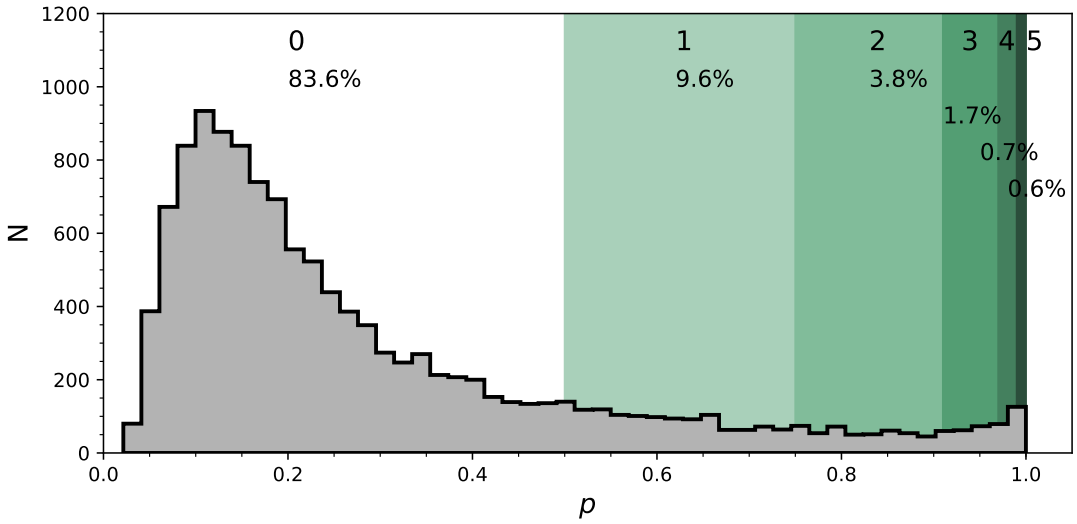

Figure 2.6. – Distribution of the predictions $\widehat{p}(m = 1|x_{\rm obs})$ produced by our algorithm on the selected COSMOS data. Sources with a $\widehat{p}(m = 1|x_{\rm obs})$ close to 1 tend to prefer the delayed-τ+flex SFH while sources with lower $\widehat{p}(m = 1|x_{\rm obs})$ favors a simple delayed-τ SFH. The green regions numbered from 1 to 5 indicate the Jeffreys scale of the Bayes factor, 1: Barely worth mentioning, 2: Substantial, 3: Strong, 4: Very strong, and 5: Decisive (detailed at the end of Sect. 2.3.2). The percentage of sources in each grade is provided on the Figure and in Table 2.7.

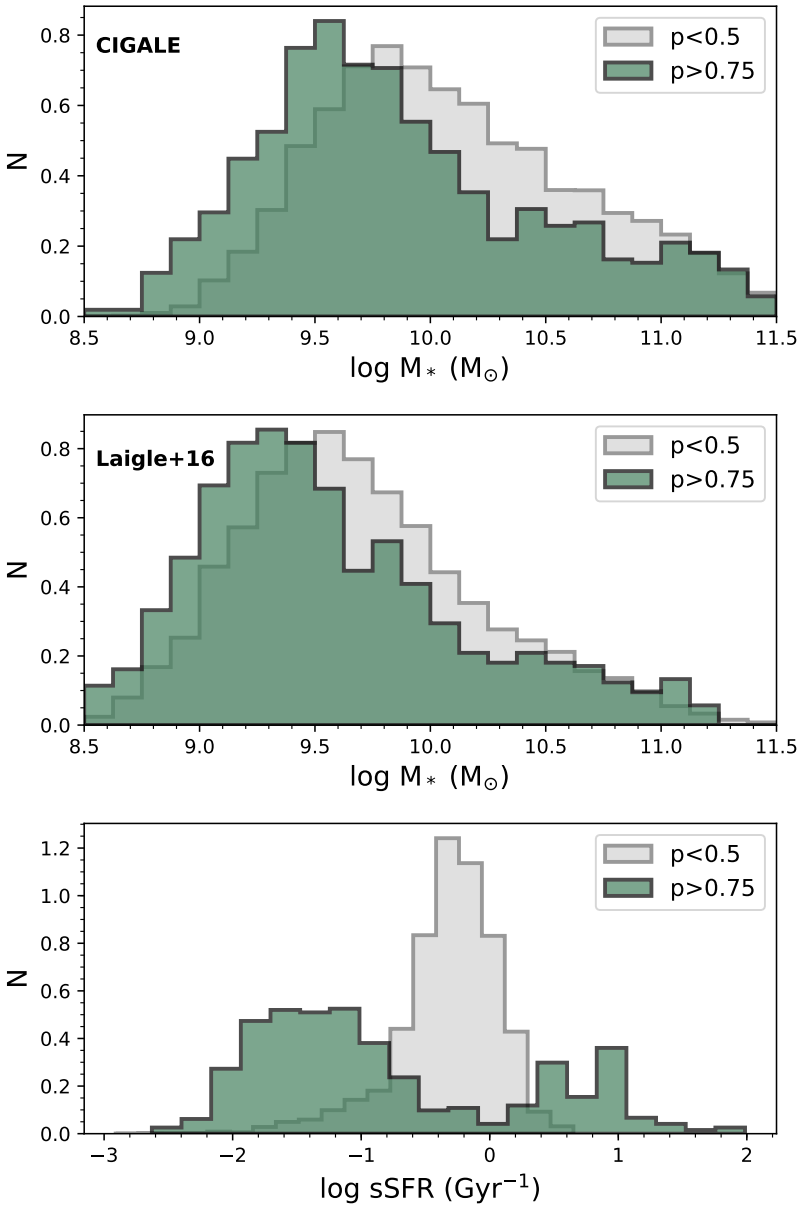

Figure 2.7. – **Top panel:** Comparison of stellar mass distribution, obtained with CIGALE, for the sample of galaxies with $\widehat{p}(m = 1|x_{\rm obs}) >= 0.75$ (green) and galaxies with $\widehat{p}(m = 1|x_{\rm obs}) < 0.5$ (grey). **Middle panel:** Comparison of stellar mass distribution, obtained by C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang- Jensen, et al., 2016a, for the sample of galaxies with $\widehat{p}(m = 1|x_{\rm obs}) >= 0.75$ (green) and galaxies with $\widehat{p}(m = 1|x_{\rm obs}) < 0.5$ (grey). **Bottom panel:** Comparison of sSFR distribution for the sample of galaxies with $\widehat{p}(m = 1|x_{\rm obs}) >= 0.75$ (green) and galaxies with $\widehat{p}(m = 1|x_{\rm obs}) < 0.5$ (grey).

3. Tempered, Anti-truncated Multiple Importance Sampling

Sommaire

3.1. Introduction

Importance sampling is a Monte Carlo method that predates Markov Chain Monte Carlo (MCMC). It was and is still used to sample distributions. Importance sampling targets $\pi(x)$ with draws from the proposal distribution $q(x)$. A draw x is weighted with $\pi(x)/q(x)$ to correct the discrepancy between q and π. When $\pi \ll q$, these algorithms are unbiased. Moreover, when the density of the target $\pi(x)$ is known up to a constant, we normalized the weights by their sum, which introduce a small bias that has been well studied (see, e.g. Christian P Robert, Casella, and Casella, 1999). Unlike MCMC, importance sampling is an embarrassingly parallel algorithm that can easily be distributed on CPU cores or clusters. Moreover, importance sampling does not require to sort the wheat from the chaff by finding the limit of the warm-up or burn-in period. And, since it is not based on local moves, it may be able to discover the different modes of the target. It has therefore received a recent interest, in particular when considering algorithms that calibrates the tuning parameters of the algorithm to the target (M. F. Bugallo, V. Elvira, Martino, et al., 2017).

The efficiency of importance sampling depends heavily on the choice of the proposal. Many adaptive algorithms (M.-S. Oh and Berger, 1992; M.-S. Oh and Berger, 1993) have been proposed to calibrate the proposal based on past samples from the target. Thus a temporal dimension is introduced in these algorithms to adapt the tuning parameters of the proposal distribution: at time t, draws x are sampled from a distribution $q_t(x) = q(x|\theta_t)$ whose parameter θ_t is adapted on past results. However these algorithms suffer from numerical instability and sensibility to the first proposal used at initialization. For instance, Liu (2001, Section 2.6) claimed that such algorithms were unstable. Indeed estimating large covariance matrices from weighted samples can lead to ill-conditioned estimation problems (see, e.g., El-Laham, Victor Elvira, and M. Bugallo, 2018). And Cornuet, Marin, Mira, et al. (2012) asserted that the initial distribution of their algorithm has a major impact on the accuracy of adaptive algorithms. They talked about the "what-you-get-is-what-you-see" nature of such algorithms: these methods have to guess which part of the space is charged by the target based on points of this space that have been previously visited. Several schemes have been introduced to initialise the first proposal distribution. The initialization method proposed by Cornuet, Marin, Mira, et al. (2012, Section 4) requires multidimensional simplex optimization, hence requires many evaluations of $\pi(\theta)$ that are then discarded. On the other hand, Beaujean and Caldwell (2013) runs a complete Metropolis-Hastings algorithm that can miss several modes of the target since it is based on local moves.

Numerical instability may come from the fact that the adaptive algorithm can be trapped around a point of the space that better fits the target than previously visited points. When such phenomenon occurs, the algorithm misses important parts of the core of the target: the learnt proposal distribution becomes concentrated around this point, and the rest of the space to sample is eliminated forever. When the space to sample is of moderate or large dimension, numerical instability becomes a major problem. Many ideas were proposed to tackle the issue including tempering and clipping (M. F. Bugallo, V. Elvira, Martino, et al., 2017). Tempering can be implemented as replacing the target $\pi(x)$ by $\pi(x)^\beta$, with $\beta < 1$. It eases the discovery of the core of the target since it extends the part of the space that is charged by the target. Thus, tempering can smooth the bridge from the first proposal $q_1(x)$ to the target $\pi(x)$. Clipping (Ionides, 2008; Koblents and Miguez, 2015; Vehtari, Simpson, Andrew Gelman, et al., 2021) of the importance weights is a non linear transformation of the weights that decreases the importance of points with high $w(x) = \pi(x)/q_t(x)$. The most common way to implement clipping as a variance reduction method (which introduce a bias) is the truncation that deals with the degeneracy as follows. If $w(x) > S$ where S is a threshold that needs to be calibrated, the weights $w(x)$ are replaced by some value (e.g., by S). Otherwise, they are left unchanged. As noted by Koblents and Miguez (2015) and M. F. Bugallo, V. Elvira, Martino, et al. (2017), this transformation of the weights flattens the target distribution. Therefore, truncation is redundant with tempering. Finally, in order to increase computational efficiency, schemes have been introduced to recycle the successive samples generated at every iterations. In this vein, Cornuet, Marin, Mira, et al. (2012) and Marin, Pudlo, and Sedki (2019) considered the whole set of

draws from the different proposals calibrated at each stage of the algorithm as drawn from a mixture of these distributions to significantly increase their efficiency.

In this paper, we propose an adaptive importance sampling whose sensitivity to the first proposal, and numerical instability are highly reduced. We have tried to design our algorithm to keep control on the number of evaluations of the (unnormalized) target density. In many situations where we are interested by sampling the posterior distribution, the target density is indeed a complex function of the parameters x and the data. For instance, an extreme case is a Gaussian model whose average $\mu(x)$ is a blackbox function which carries a physical model of the reality given the value of the parameters x. Thus, the time complexity of our algorithm should be assessed in number of evaluations of the proposal density. We relied on a simple form of tempering to adapt the proposal distribution. Nevertheless tempering was not enough to stabilize the algorithm on spaces of large dimension. In our algorithm, at each stage after initialization, our new proposal distribution is calibrated to approximate a tempered version of the target contaminated with draws from the previous proposal. Both tricks (tempering and contamination with previous proposal) avoid focusing too quickly on the few points with high $w(x) = \pi(x)/q_t(x)$. At least, they keep the variance of the proposal large enough to take time to explore the space to sample before exploiting the points with high importance weights.

3.2. Calibration of importance sampling

We propose here a new strategy to walk on the bridge from the first proposal $q_1(x)$ to a proposal $q_T(x)$ well adapted to the target $\pi(x)$ in terms of effective sample size. In order to adapt the proposal gradually, we introduce a sequence of temporary targets:

$$\widehat{\pi}_1(x), \ldots, \widehat{\pi}_T(x)$$

which are intermediaries between the first proposal $q_1(x)$ and the target $\pi(x)$. The precise definition of these temporary targets, given in Section 3.2.2, is paramount to the succes of the algorithm. They are based on a tempering w^β of the importance weights w. As described in Section 3.2.1, the tempering

(i) eases the discovery of the area charged by the real target $\pi(x)$,
(ii) temporarily removes the problems due to large queues of the target,
(iii) allows us to design a diagnostic based on the final of β.

To this non-linear transformation of the weights, we add an anti-truncation, defined as $\widehat{w}^\beta = w^\beta \vee s$, that pulls up all tempered weights w^β less than a threshold s to this single value, see Section 3.2.2. This anti-truncation

(iv) performs a contamination of the temporary target by the last proposal,
(v) helps to stabilize numerically the algorithm and
(vi) allows us to explore new directions in large-dimension spaces.

Both β and s are automatically calibrated at the end of stage t of the algorithm, as explained in Section 3.3.1. The new proposal $q_{t+1}(x)$ is tuned to fit the temporary

$\widehat{\pi}_t(x)$ with the EM algorithm as given in Section 3.2.3. The whole algorithm is given in Figure 3.1.

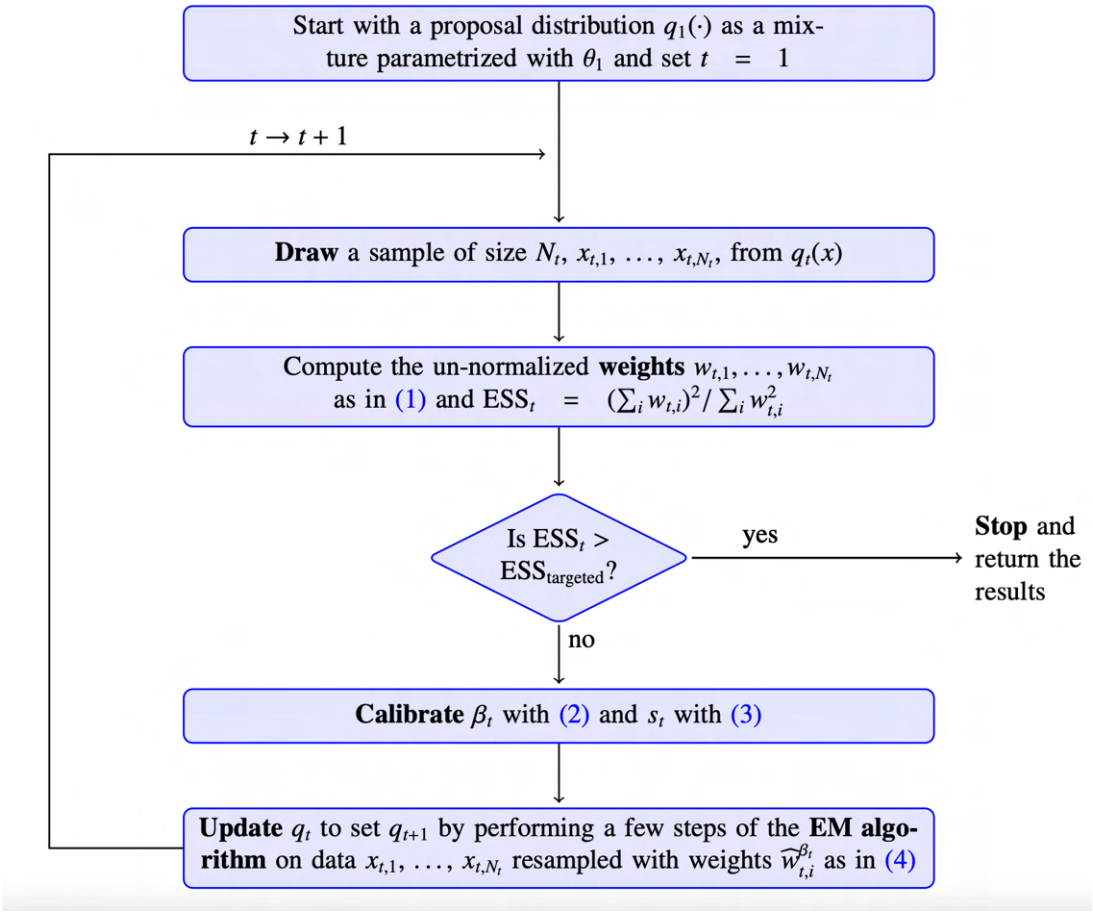

Figure 3.1. – **The tempered, anti-truncated multiple importance sampling (TAMIS) algorithm**

3.2.1. The tempering

Let us assume that, given all past draws $\mathscr{F}_{t-1}$, a set $x_{t,1}, \ldots, x_{t,N_t}$ of size N_t has been drawn independently from a distribution $q_t(x) = q(x|\theta_t)$ picked among a parametric family $\mathscr{Q}$ of laws. The importance weights at this stage are

$$w_{t,i} = \frac{\pi(x_{t,i})}{q_t(x_{t,i})}. \tag{3.1}$$

We can replace the target $\pi(x)$ by the distribution of density

$$\pi_{\beta,t}(x) \propto \pi(x)^{\beta} q_t(x)^{1-\beta} \tag{3.2}$$

with inverse temperature $\beta \in (0,1)$ as proposed by R. M. Neal (2001) in his Annealed

importance sampling. When $\beta = 0$, (3.2) is the proposal distribution that served to draw the $x_{t,i}$'s: $\pi_{\beta=0,t}(x) = q_t(x)$. When $\beta = 1$, (3.2) is the target distribution: $\pi_{\beta=1,t}(x) = \pi(x)$. Moreover, $\beta \mapsto \mathrm{KL}(\pi \| \pi_{\beta,t})$ decreases from $\mathrm{KL}(\pi|q_t)$ to 0, see Proposition 3 in Appendix C. If we use the $x_{t,i}$'s to target $\pi_{\beta,t}(x)$, the unnormalized importance weights become

$$\frac{\pi_{\beta,t}(x_{t,i})}{q_t(x_{t,i})} \propto \frac{\pi(x)^\beta q_t(x)^{1-\beta}}{q_t(x_{t,i})} = \left(\frac{\pi(x)}{q_t(x_{t,i})}\right)^\beta = w_{t,i}^\beta. \tag{3.3}$$

Such weights had been use in the past, for instance by Koblents and Miguez (2015) who relied on the $x_{t,i}$'s weighted with the $w_{t,i}^\beta$'s to get a sample from $\pi_{\beta,t}(x)$ and to tune a $q_{t+1}(x) = q(x|\theta_{t+1})$ that approximates $\pi_{\beta,t}(x)$. It is also explored by Korba and Portier (2022) as a regularization strategy.

3.2.2. Anti-trunctation and temporary targets

There are many ways to contaminate this weighted sample with draws from $q_t(x)$. The first idea is to add N_t' new draws $x_{t,N_t+1}, x_{t,N_t+2} \dots$ with all weights equal to s to the above weighted sample. This idea may add a non negligeable amount of computational time when the dimension of x is large. Another idea to contaminate this weighted sample with $q_t(x)$, is to change the weights. We introduce a deterministic contamination based on the value of $w_{t,i}^\beta$. Indeed, the $x_{t,i}$'s weighted with

$$\widehat{w}_{t,i}^\beta = s \vee w_{t,i}^\beta \tag{3.4}$$

form an approximation of the distribution with density

$$\widehat{\pi}_{\beta,t}(x) \propto s q_t(x) \mathbf{1}\{x \in E\} + \pi^\beta(x) q_t^{1-\beta}(x) \mathbf{1}\{x \notin E\}, \quad \text{where } E = \{x : \pi^\beta(x)/q_t^\beta(x) \le s\}. \tag{3.5}$$

An easy computation gives us the weights of the mixture as follows.

Lemma 1. *Let $q_t^E(x)$ denotes the normalized probability density of $q_t(x)$ knowing $x \in E$ and $\pi_{\beta,t}^{\bar{E}}(x)$ the normalized probability density of $\pi_{\beta,t}(x) = \pi^\beta(x) q_t^{1-\beta}(x)$ knowing $x \notin E$.*

We have

$$\widehat{\pi}_{\beta,t}(x) = \lambda q_t^E(x) + (1-\lambda)\pi_{\beta,t}^{\bar{E}}(x)$$

where $\lambda = s\int_E q_t(x)\mathrm{d}x = 1 - \int_{\bar{E}} \pi^\beta(x) q_t^{1-\beta}(x)\mathrm{d}x$.

Note that the scheme is different from the Safe Importance Sampling one (Delyon and Portier (2021), Owen and Zhou (2000)) as the anti-truncation contaminates the target with the current proposal q_t instead of q_0, and specifically in E. We apply in (3.4) a non-linear transformation of the weights. Yet it is the inverse of truncating the importance weights and we refer to these transformed weights as anti-truncated weights. Unlike the common truncation of the weights that replaces all weights larger

than S by S, the anti-truncation we propose in (3.4) replaces all weights smaller than s by s. Actually, we do not need to truncate large values since we relied on tempering to remove the degeneracy of the weights. However the sample drawn from $q_t(x)$ with weights $w_{t,i}^{\beta}$ may not be of sufficient size to approximate (3.2) correctly, even if β is well calibrated. If we trust that $q_t(x)$ is a decent sampling distribution, the anti-truncated, tempered weights fight against the degeneracy of the weights in importance sampling (tempering) and keep part of the old proposal (q_t) to keep exploring the space from it (anti-tempering). At the end of each stage t (except the final one), the future proposal distribution $q_{t+1}(x) = q(x|\theta_{t+1})$ is calibrated on the temporary target given by (3.5). The anti-truncated, tempered $\widehat{\pi}_{\beta,t}(x)$ defined in (3.5), is a continuous bridge from

— the real target $\pi(x)$ to

— the freshly used proposal $q_t(x) = q(x|\theta_t)$.

The tempered target $\pi_{\beta,t}(x) = \pi^{\beta}(x)q_t^{1-\beta}(x)$ is already such a continuous bridge. But, when β is fixed, the anti-truncated, tempered $\widehat{\pi}_{\beta,t}$ is in-between the tempered $\pi_{\beta,t}$ and the freshly used proposal $q_t(x)$ in terms of Kullback divergence as given by Proposition 2. Let us recall first that, if both f and g are probability densities, then the Kullback divergence is defined as

$$\mathrm{KL}(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x.$$

If f and g are unnormalized probability densities, we will still denote by $\mathrm{KL}(f||g)$ the Kullback divergence between their normalized versions.

The following proposition is proved in Appendix D.

Proposition 2. *When $s \leq 1$, we have*

$$0 = \mathrm{KL}\left(\pi_{\beta,t} \middle\| \pi_{\beta,t}\right) \leq \mathrm{KL}\left(\pi_{\beta,t} \middle\| \widehat{\pi}_{\beta,t}\right) \leq \mathrm{KL}\left(\pi_{\beta,t} \middle\| q_t\right)$$

3.2.3. Updating the proposal

The family of proposals we recommend for TAMIS is composed of Gaussian mixture models, with diagonal covariance matrix for each component. The density of a distribution $q(x|\theta) \in \mathcal{Q}$ is defined as

$$q(x|\theta) = \sum_{k=1}^{K} \mathfrak{p}_k \varphi(x|\mu_k, \Sigma_k)$$

where $\varphi(x|\mu, \Sigma)$ is the multivariate Gaussian density with mean μ and covariance matrix Σ. This family is parametrized by $\theta = (\mathfrak{p}_1, \ldots, \mathfrak{p}_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$.

The future proposal distribution $q_{t+1}(x) \in \mathcal{Q}$ is set by using the EM algorithm. Let us assume that $q_t(x) = q(x|\theta_t) \in \mathcal{Q}$ is the Gaussian mixture with parameter θ_t. We tune $q_{t+1}(x) = q(x|\theta_{t+1}) \in \mathcal{Q}$, that is to say, we pick θ_{t+1} with the help of the $x_{t,i}$'s weighted with $\widehat{w}_{t,i}^{\beta}$ as given in (3.4). After resampling this sample according to their weights

$\widehat{w}_{t,i}^{\beta}$, we resort to iterations of the EM algorithm, starting from θ_t, to get θ_{t+1}. Because of well known properties of the EM algorithm (see, e.g., Fruhwirth-Schnatter, Celeux, and Christian P Robert, 2019), we have that

$$\mathrm{KL}\big(\widehat{\pi}_{\beta,t} \big\| q_{t+1}\big) < \mathrm{KL}\big(\widehat{\pi}_{\beta,t} \big\| q_t\big).$$

3.3. Practical aspects of the TAMIS algorithm

We can now discuss pratical aspects of the proposed algorithm, based on numerical results that demonstrate the typical behavior of the method.

3.3.1. Choosing the inverse temperature β and the anti-truncation s

The inverse temperature β has to be chosen at each stage of the algorithm (except the last one). We follow the path open by by Beskos, Jasra, Kantas, et al. (2016) to chose β. To ensure that the $x_{t,i}$'s weighted with $\widehat{w}_{t,i}^{\beta}$ is a sample that can approximate $\widehat{\pi}_{\beta,t}$, we set β automatically at each stage with

$$\beta_t = \sup\big\{\beta \in (0,1) : \mathrm{ESS}(\beta) > \mathrm{ESS}_{\min}\big\}, \quad \text{where } \mathrm{ESS}(\beta) = \left(\sum_{i=1}^{N_t} w_{t,i}^{\beta}\right)^2 \Big/ \sum_{i=1}^{N_t} w_{t,i}^{2\beta}. \tag{3.6}$$

The function $\beta \mapsto \mathrm{ESS}(\beta)$ is continuously decreasing (see Proposition 4 of the Appendix). Hence the optimization problem stated in (3.6) can be solved easily by a simple one-dimensional bisection method and do not require a new sampling step, contrary to Korba and Portier (2022)'s adaptive regularization scheme. Note that the weights $\widehat{w}^{\beta}$ related to the temporary target (3.5) are used only to calibrate the next proposal $q_{t+1}(x)$ — this is an important difference with the algorithm proposed by Koblents and Miguez (2015). Hence the value of $\mathrm{ESS}_{\min}$ should be fixed such that the fit of $q_{t+1}(x)$ with the EM algorithm provides stable estimates with an iid sample of size $\mathrm{ESS}_{\min}$.

A good choice of $\mathrm{ESS}_{\min}$ is paramount to get numerical stability in our algorithm. If $\mathrm{ESS}_{\min}$ is much larger than really needed, the algorithm will remain stable numerically. But convergence to the target will be slow down: as the tempering will be more aggressive at each stage, more iterations will be needed to move from the first proposal $q_1(x)$ to the target $\pi(x)$. The typical effect of changing the value of $\mathrm{ESS}_{\min}$ is studied in Figure 3.2. For example, if $\mathscr{Q}$ is the set of mixtures of K Gaussian densities with diagonal covariances, the update of the proposal with EM steps require to calibrate Kd mean paramaters and Kd variance parameters. Thus, we should have $2Kd \ll \mathrm{ESS}_{\min} \leq N_t$.

The value of s that set the amount of anti-truncation is more easy to tune. We chose

s to be the quantile of order τ of the tempered weights:

$$s_t = \text{quantile}_{\text{order}=\tau}\left(w_{t,1}^{\beta}, \dots, w_{t,N_t}^{\beta}\right). \tag{3.7}$$

Although the required number of iterations may be suboptimal, the value $\tau = 0.4$ appears to be a universal compromise, working flawlessly in every numerical example considered in this paper. Lower values of τ picked in $(0, 0.1)$ can speed up the algorithm in low dimensional problems, but can induce instability. Hence, we strongly advocate for the almost universal $\tau = 0.4$, see Figure 3.3.

3.3.2. Numerical diagnostics

In order to assess the convergence of the algorithm we monitor the inverse temperature and the estimated Kullback-Leibler divergence along iterations. Following Cappé, Douc, Guillin, et al., 2008, we estimate the Kullback-Leibler divergence between the target density and the mixture proposal using the Shannon entropy of the normalised IS weights. Indeed since the normalised perplexity $\exp\left(H^{t,N}\right)/N$ is a consistent estimator of $exp(-KL(\pi||q_t)$,where $H^{t,N} = -\sum_{i=1}^{N_t} \omega_{i,t} log\omega_{i,t}$ (Cappé, Douc, Guillin, et al., 2008), we simply estimate $KL(\pi||q_t) \approx \sum_{i=1}^{N_t} \omega_{i,t} log\omega_{i,t} + logN_t$. Note that this estimate is upper bounded by $logN_t$, leading to an obvious bias when $KL(\pi||q_t)$ is large or N_t small. However this bias does not practically prevent the use of this estimate as a monitoring tool.

We show in Figure 3.4 the typical evolution of both the inverse temperature β and the estimated KL divergence along iteration. The inverse temperature starts increasing slowly during the first iterations, followed by a strong acceleration until it stabilises. The estimated KL divergence on the other hand starts with a plateau at its upper bound ($\log N_t$), then drops to a much small value as β reaches it maximum.

In some cases, β does not reach 1, nor does the estimated KL divergence reach 0. Indeed if the target density can't be well approximated by any proposal in $\mathscr{Q}$, $KL(\pi||q_t)$ never reaches 0. This behaviour is also observed on targets of very high dimension regardless of the proposal distribution family (see Section 3.4.2). Even in those pathological cases, the convergence of TAMIS can be simply assessed by the sharp increase of β followed by its stabilization (or the sharp decrease of the estimated KL).

3.3.3. Stopping criterion and recycling

When the iterative algorithm is stopped at time T, we end with a set of weighted simulations:

$$x_{t,i} \sim q_t(\cdot) = q(\cdot|\theta_t), \quad \text{with weight } w_{t,i} = \frac{\pi(x_{t,i})}{q_t(x_{t,i})}.$$

Table 3.1. – Parameter tuning and monitoring experiments

Experiment	E3.1	E3.2	E3.3
Dimension	$d = 50$	$d = 50$	$d = 1,000$
Target	$\mathcal{N}(50,5)^{\otimes d}$	$\mathcal{N}(50,5)^{\otimes d}$	$\mathcal{N}(10,5)^{\otimes d}$
Proposals	Gaussian mixture with 5 components		Gaussian
Draws	$N_t = 2,000$	$N_t = 2,000$	$N_t = 2,000$
ESS$_{\text{min}}$	$\in \{100, 200, 1400\}$	300	1,000
$\boldsymbol{\tau}$	0	$\in \{0, 0.1, \ldots, 0.9, 0.95\}$	0.4
Stop	$\sum_t \text{ESS}_t > 10,000$		$t = 500$

As in many iterative importance sampling algorithms such as AMIS Cornuet, Marin, Mira, et al., 2012, we recycle all these draws and change their weights to

$$w_{t,i} = \frac{\pi(x_{t,i})}{Q(x_{t,i})}, \quad \text{where } Q(x) = \frac{1}{N_1 + \cdots + N_T} \sum_{t=1}^{T} N_t q_t(x).$$

We use the usual effective sample size estimate to assess the quality of the IS sample given by TAMIS. Thus we suggest stopping the algorithm when the predifined ESS or the maximal number of iterations is reached. As usual in such adaptive algorithms, we recycle all particles with their weights after stopping the iterations. This recycling improve the efficiency of the algorithm. Thus, the ESS of the final sample returned by the algorithm is underestimated by the sum of the effective sample sizes at each iteration. Hence, to monitor that we have reached the predefined level, we stop at the first time where

$$\text{ESS}_1 + \cdots + \text{ESS}_t > \text{ESS}_{\text{predefined}}$$

or when we reach the maximal number of iterations.

3.3.4. Parameter tuning and monitoring

We start by illustrating the effect of parameter tuning on TAMIS with the experiments targeting various multivariate Gaussian distribution as given in Table 3.1. The proposal at first iteration was a Gaussian mixture with 5 components: each component is centered around a μ_k drawn at random from $\mathcal{U}([-4,4])^{\otimes d}$ and has covariance matrix $\Sigma_k = 200 \times \mathbb{I}_{50}$ with large eigenvalues. To approximate de MSE, we ran 20 replicates of the experiences for each set of parameters.

Figure 3.2 shows Experiment E3.1 described in Figure 3.1 and Figure 3.3 shows Experiment E3.2. The conclusion is that we should set ESS_{min} so that the calibration of the new proposal (i.e., of θ_{t+1}) is stable and that $\tau = 0.4$ is a decent value.

To illustrate monitoring in Figure 3.4, we first plot a typical tempering path (obtained on Experiment E3.1 with $\text{ESS}_{\text{min}} = 100$ and $\tau = 0$) along with the estimated KL divergence. As mentioned in section 3.3.1, the auto calibrated tempering path has

a rather sigmoid-like shape with a clear transition and stabilization to $\beta = 1$, while the KL-divergence decreases (despite the estimator bias at the beginning) until both quantities stabilizes together around 1 and 0 respectively.

Finally we illustrate the typical behavior of the monitoring on targets of very high dimension with Experiment E3.3. The first proposal distribution to initialize TAMIS is a Gaussian distribution centered at μ drawn from $\mathcal{U}([-4,4])^{\otimes d}$ and with covariance matrix $\Sigma = 100 \times \mathbb{I}_d$. Figure 3.5 shows that TAMIS provide more than decent results in high dimension.

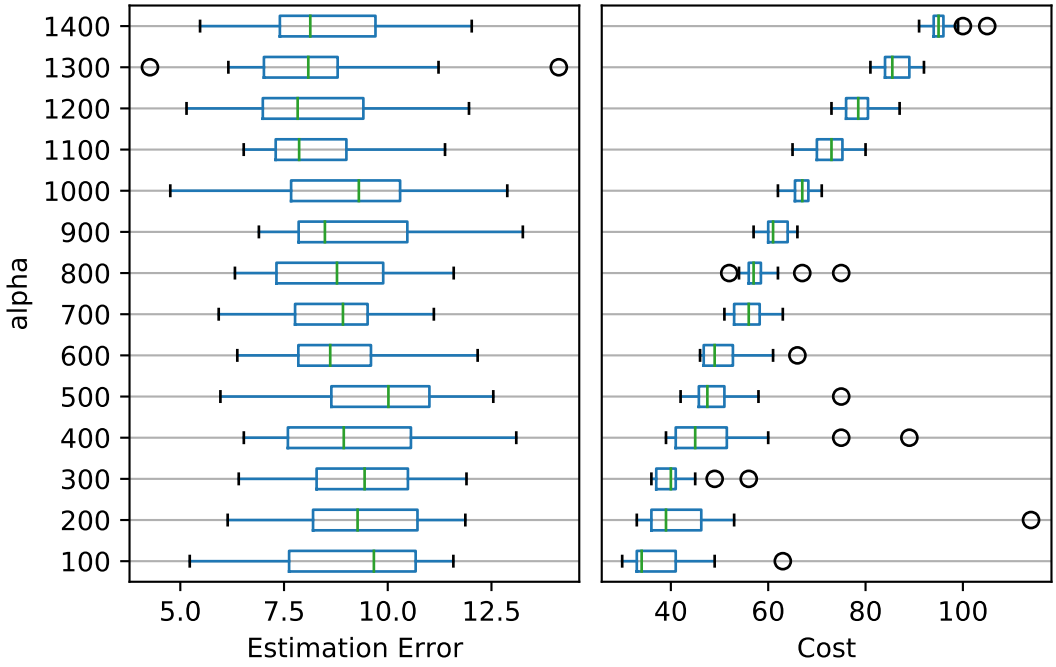

Figure 3.2. – Effect of varying the $\mathrm{ESS}_{\min}$ parameter (y-axis) defining the minimum ESS to be reached for calibration of the interse temperature. As stopping depends on the total estimated ESS, the MSE of the variance (x-axis on the left) estimation doesn't depend on $\mathrm{ESS}_{\min}$, but the number of required iterations (x-axis on the right) before convergence of the sequence of proposal distributions increases. Increasing $\mathrm{ESS}_{\min}$ further than the minimum required to stabilize the calibration of the new proposal (i.e., of θ_{t+1}) with the EM step results in an increased computational cost.

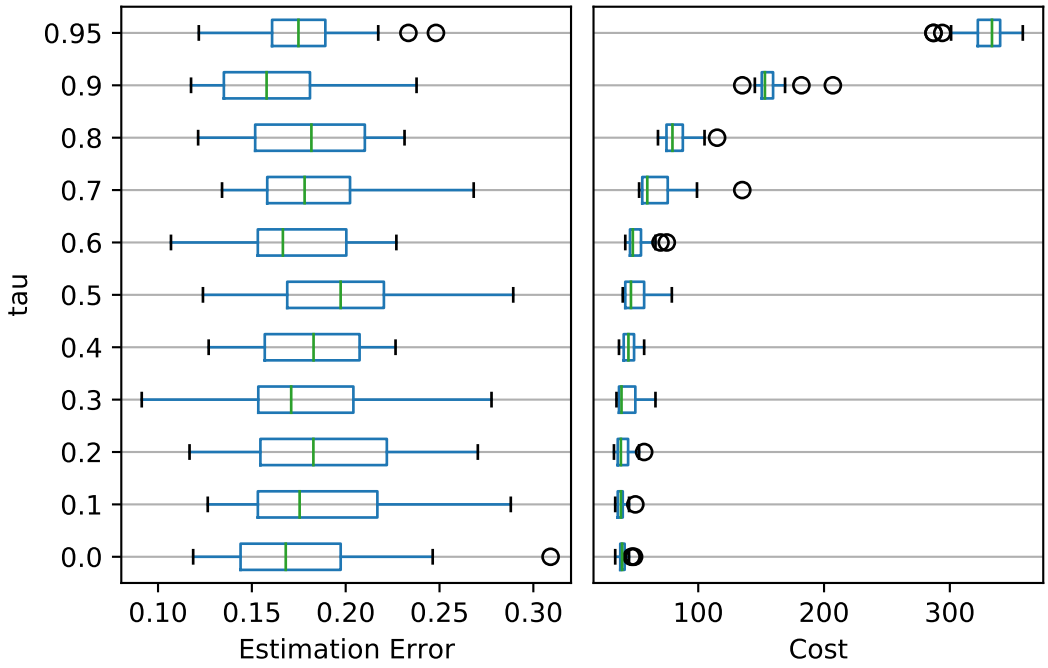

Figure 3.3. – Effect of varying the τ parameter (y-axis) defining the antitruncation threshold. Except for very high values, the truncation has no detrimental effect on either the MSE (x-axis on the left) of the estimated variance or the required number of iterations (x-axis on the right) before convergence of the sequence of proposal distributions increases. As for $\mathrm{ESS}_{\mathrm{min}}$, once the calibration of of the new proposal (i.e., of θ_{t+1}) with the EM step is stable, increasing τ further only increases the computational cost.

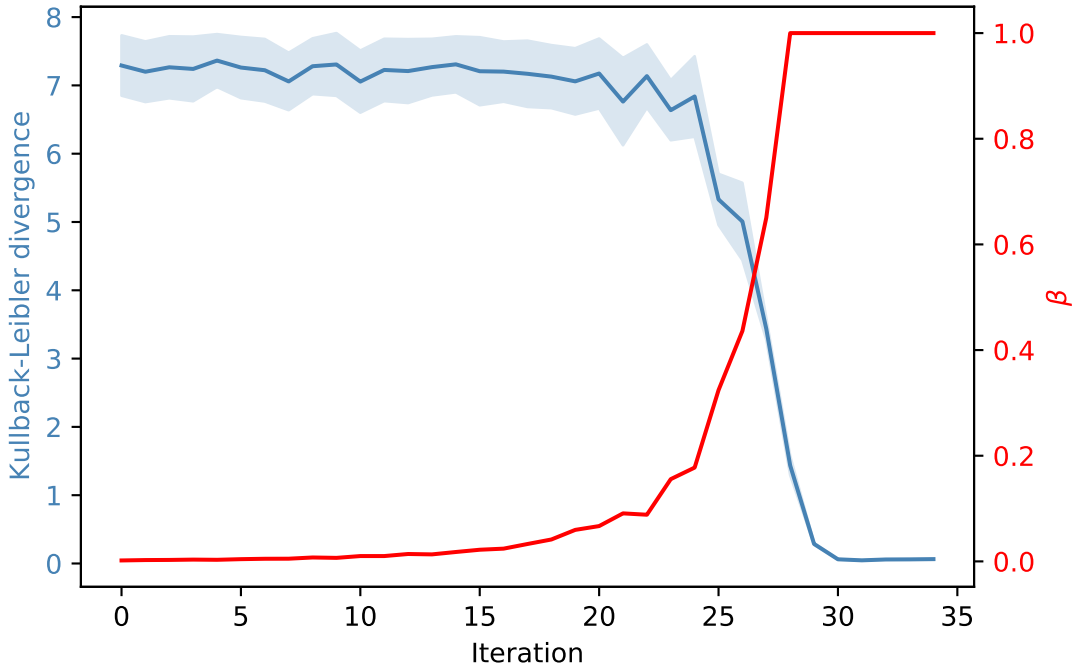

Figure 3.4. – Typical evolution of the inverse temperature β (y-axis in red) and estimated Kullback-Leibler divergence (y-axis in blue) along iterations (x-axis). The automatically calibrated β starts by increasing slowly until a sharp acceleration, followed by stabilization clearly indicating convergence of sequence of proposal distributions. The estimated KL divergence shows the upper bound biais until iteration 20, as detailed in 3.3.2. Yet its sharp decrease and stabilization mirrors β's path.

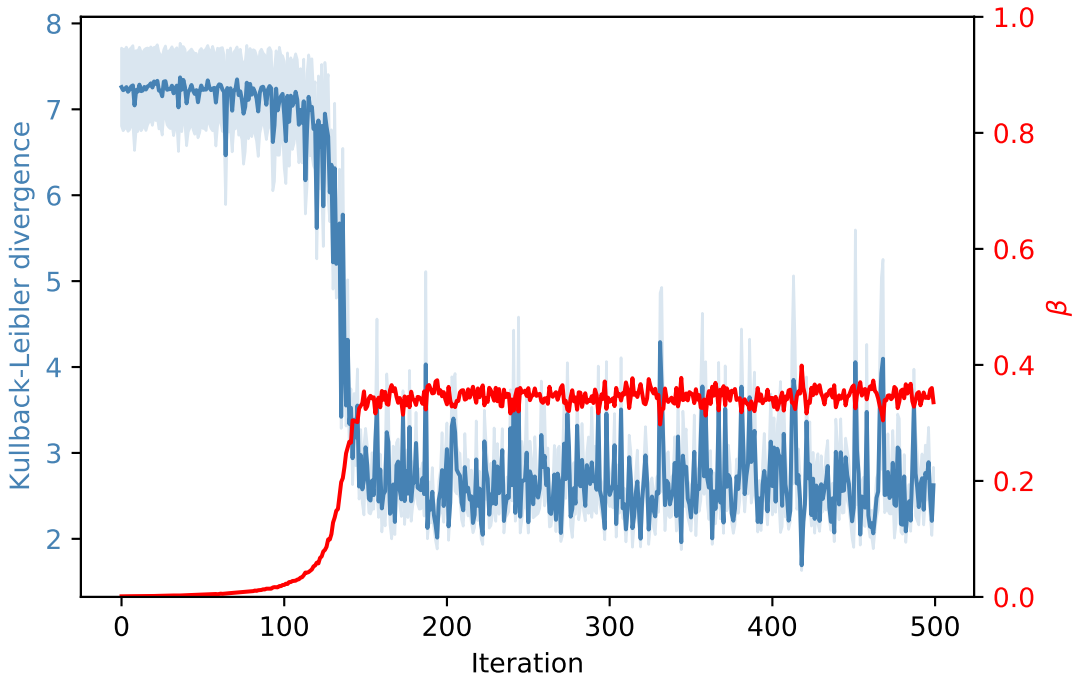

Figure 3.5. – A very high-dimensional problem : The target is a 1000-dimensional gaussian distribution, the proposals are gaussian distributions with diagonal covariance. (left) Evolution of the inverse temperature β (in red) and estimated Kullback-Leibler divergence (blue) along iterations. *(right)* the L2 distance between the moments of the target and proposal distribution at each iteration. The temperature doesn't go to 1 despite the target distribution belonging to the family of proposal distributions and the covariance of the proposal doesn't converge to the real covariance.

3.4. Numerical Experiments

We finally illustrate the good numerical properties of TAMIS relatively to its initialization and to the dimensionality of the problem.

3.4.1. On the effect of initialization

We now compare the effect of a bad initialization on TAMIS, AMIS and N-PMC with Experiment E4.1 given in Table 3.2. The example considered is the banana shape target density of Haario et al., also known as the Rosenbrock distribution. Let $\sigma^2 = 100, \Sigma = \text{diag}(\sigma^2, 1, \dots, 1), b = 0.03$ and $\Psi(x) = \left(x_1, x_2 + b(x_1^2 - \sigma^2), x_3, \dots, x_d\right)$. The target is the Rosebrock distribution with density

$$\pi(x) = \varphi(\Psi(x)|0, \Sigma).$$

Table 3.2. – Initialization and dimensionality

Experiment	E4.1	E4.2	E4.3
Dimension	$d \in \{20, 50\}$	$d \in \{5, 10, 20, 50, 100\}$	$d \in \{300, 500\}$
Target	Rosenbrock distr.	$\mathcal{N}(50, 5)^{\otimes d}$	
Proposal	Gaussian mixture with 5 components		
Draws	$N_t = 2,000$	$N_t = 1,000$	$N_t = 2,000$
ESS$_{\min}$	100	300	1,000
$\boldsymbol{\tau}$	0.4		
Stop	$t = 20$	$\sum_t \mathrm{ESS}_t > 1,000$	

For N-PMC, the inverse temperature sequence is chosen as in Koblents and Miguez, 2015, i.e., $\beta_t = 1/(1 + e^{-(t-\ell)})$ where ℓ is a tuning parameter we have set to 5.

The first proposal at initialization is a Gaussian mixture model with 5 components with covariance matrix all equal to Σ, and centered at random μ_k drawn from $\mathcal{N}(0, \Sigma_{0,k}/5)$. We used various covariance matrices Σ, starting from the diagonal matrix $\mathrm{diag}(200, 50, 4, \ldots, 4)$ used in Wraith, Kilbinger, Benabed, et al., 2009 and Koblents and Miguez, 2015. This initial covariance matrix is already adapted to the target and can be considered as an a priori informed proposal. Then, we used less informed covariance matrices for Σ:

— $\mathrm{diag}(200, 50, 10, \ldots, 10)$,

— $\mathrm{diag}(200, 50, 20, \ldots, 20)$, $\mathrm{diag}(200, 50, 50, \ldots, 50)$,

— $\mathrm{diag}(200, 100, 100, \ldots, 100)$ and

— finally $200 \times \mathrm{I}_d$ which is blind regarding the shape of the target.

Each experiment was repeated 500 times.

Figure 3.6 shows the final ESS. As expected the final ESS after a fixed number of iterations decreases as the initialization gets worse. Since the dimension is already high, AMIS fails very frequently even with the first initialization. The tempering scheme of N-PMC is effective only with a well calibrated initialization, while TAMIS remains effective and allows the algorithm to converge in every case without any additional parameter tuning.

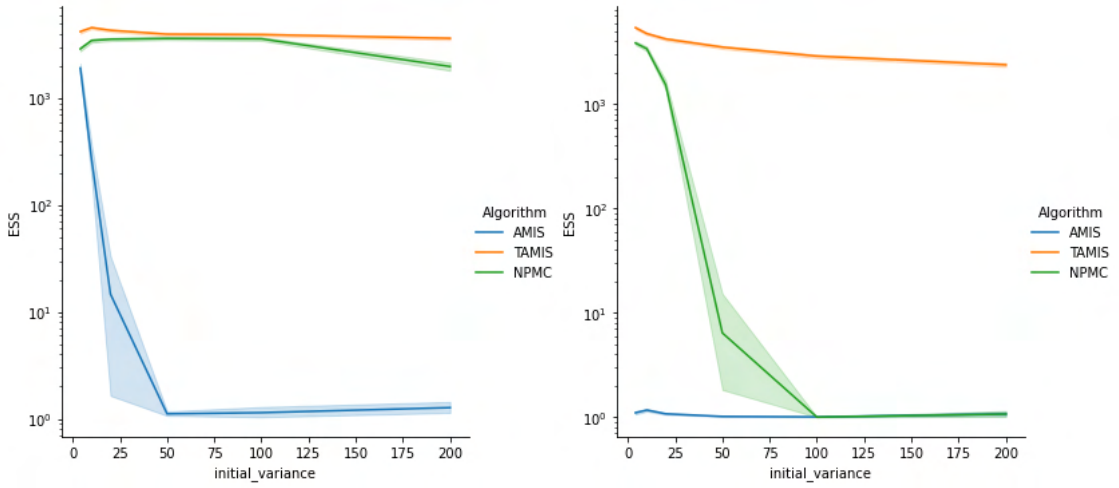

Figure 3.6. – Effective Sample Size (y-axis) of AMIS, N-PMC and TAMIS after 40,000 draws along 20 iterations, with increasingly wide covariance matrix at initialization (x-axis) in dimension 20 (left) and 50 (right). As expected from the litterature, AMIS is only performing well with a good initialization and if the dimension is relatively low. N-PMC is able to correct for bad initialization with a well chosen tempering path if the dimension is low enough, while TAMIS performs well in every case.

3.4.2. On the effect of dimensionality

We now consider a simple Gaussian target

$$\mathcal{N}(50,5)^{\otimes d}$$

of Experiment E4.3 of Table 3.2 in high dimension. We only consider TAMIS only, as both AMIS and N-PMC fail in every case. The initialization of the proposal distribution is poor for both location and for scale. The proposal distributions are Gaussian mixture models with 5 components. At initialization, they are centered at random $\mu_k \sim \text{Unif}([-4,4])^{\otimes d}$ and have covariance matrices $\Sigma_k = 200 \times \text{I}_d$. The target is therefore very concentrated and centered very far in the tail of the initial proposal. The other tuning details are given in Table 3.2.

We plot the MSE when estimating the trace of the covariance matrix along iterations. We also plot the number of likelihood evaluations required before convergence of the proposal (assessed by the number of iterations such that $\hat{\text{KL}}(\pi||q_t) > 1$ in Figure 3.7.

The number of simulations required before convergence increases as expected with the dimension. But we note that not only is TAMIS able to accurately estimate scale and location of a very high dimensional target, it does so with the same bad initialization as previously, with very little tuning required.

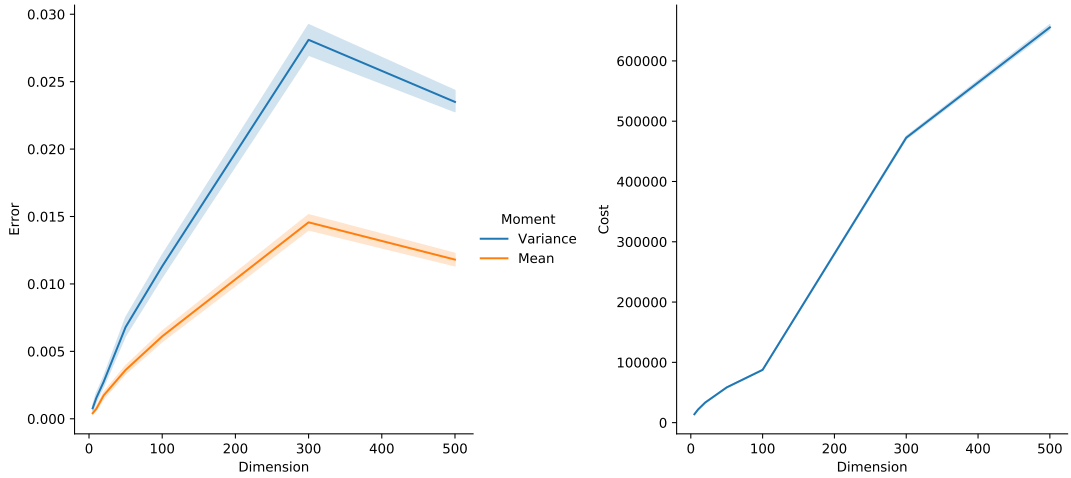

Figure 3.7. – Mean square error (y-axis on the left) of the estimates of the mean and covariance for increasing dimension (x-axis) and the required number of iterations (y-axis on the right) before convergence of the proposal to the target distribution (right).

3.5. Conclusion

We have designed an adaptive importance sampling that is

— robust to poor initialization of proposal and

— robust to high dimension of the space to sample

— efficient in the number of evaluations of the target density and

— does not rely on any gradient computation.

Very few importance sampling algorithm are stable in dimension higher than 100, and TAMIS is one of them. Therefore, TAMIS can be used to initialize other Monte Carlo algorithm such as MCMC methods that can lead to more precise estimates when correctly initialized. The phase transition observed in the decrease of the Kullbuck-Leibler divergence we monitor remains to be explained theoretically.

4. SED modeling and Neural Approximations

Sommaire

4.1. CIGALE physical modeling

To compute the spectral models, CIGALE constructs composite stellar populations from simple stellar populations combined with star formation histories.It is followed by the computation of the emission of ionized gas from massive stars, and attenuates both the stars and the ionized gas with a specified attenuation curve. Based on an energy balance principle, the absorbed energy is then re-emitted by the dust in the mid and far infrared ranges while thermal and non-thermal components are also included, extending the spectrum far into the radio wavelengths. A large grid of models is finally fitted to the data and the physical properties are estimated by analyzing the likelihood distribution.

4.1.1. A modular approach

CIGALE is split into different blocks that are as independent as can be from one another. Each physical component (stellar populations, nebular emission, attenuation by dust, dust emission, etc.) are handled separately in individual modules, and

each module is able to be substituted as transparently as possible with an other one handling the same physical component. For instance it is possible to change the attenuation law without affecting the rest of the code in any way. The next section describes those different steps.

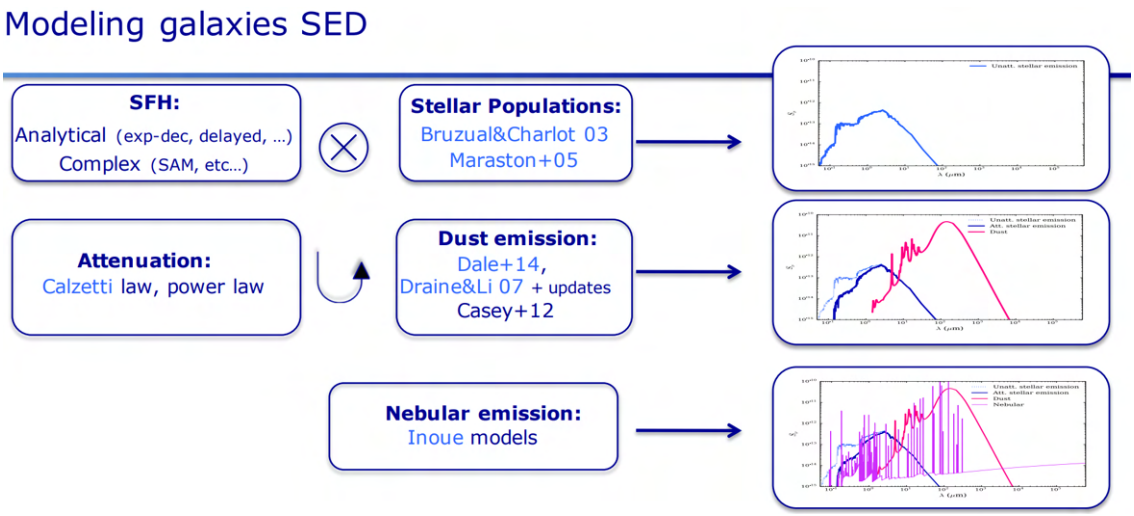

Figure 4.1. – The modular approach of CIGALE. The contributions of each physical component of galaxy emissions are computed sequentially by different modules, offering different models for each process. It starts by the combination of a chosen SFH with a SSP to obtain a first spectrum (the stellar emissions). The nebular emissions are then computed taking into account the Lyman photons from the stellar emissions. Both those contributions are then attenuated and re-emitted by dust following an energy balance principle.

The physical processes at play in galaxies provide us with a natural path to build models and compute their physical properties.In CIGALE, the models are progressively computed by a series of independent modules called successively, each corresponding to a physical component. The typical sequence to build each model is the following :

— Computation of the SFH of the galaxy.

— Computation of the stellar spectrum from the SFH and single stellar population models.

— Computation of the nebular emission (lines and continuum) from the Lyman continuum photons.

— Computation of the attenuation of the stellar and nebular emission assuming an attenuation law.

— Computation of dust emission in the mid-infrared (mid-IR)and far-IR.

— Redshifting of the model and computation of the absorption by the intergalactic medium (IGM).

In practice, the models are progressively computed by successively applying these different modules, each adding a different physical component (spectrum and associated physical parameters). For each model these individual spectral components and the combined spectrum are stored individually to ease the subsequent computation (e.g. to account for the differential reddening between younger and older stellar populations, we need to store these populations separately) and allow the user to easily retrieve the contribution from each physical component. For quantities that are more conveniently computed from the full restframe spectrum, in particular those that are directly measured observationally from the spectrum (e.g. line equivalent widths, UVslopeβ, colors, etc.), a special module can be added prior to redshifting to calculate them on the rest-frame spectrum. We describe here how each of these different physical components are modeled and parametrized

4.1.2. The different steps

4.1.2.1. Star Formation History

The first step is to compute the SFH. For each time step (1Myr) between the age of the galaxy t_0 and the current time t, CIGALE computes the corresponding SFR described by the model. The code then divides each time step in 10 bins of 0.1Myr and spreads the star formation uniformly in those bins. This spread is necessary to account for stellar evolutionary events too short to be properly modeled by greater time steps used in the SSP models. Finally, each SFH is automatically normalized so that the total mass of stars formed during the life of the galaxy is 1 $M_\odot$, and is rescaled to fit the observation after the complete modeling of the SED.

4.1.2.2. Stellar populations

The next step is to compute the intrinsic stellar spectrum, i.e the spectrum corresponding to the stellar emissions only. Now that we have the mass of stars at each time step, we can choose a Single Stellar Population (SSP) model. CIGALE relies on two pre-computed libraries of SSPs, the one from Bruzual & Charlot (2003) (Bruzual and S. Charlot, 2003 ; module bc03) and the one from Maraston (2005) (Claudia Maraston, 2005 ; module m2005). Each SSP library is available for a few values of metallicities (0.0001, 0.0004, 0.004, 0.008, 0.02, and 0.05 for ; and 0.001, 0.01, 0.02, and 0.04 for respectively) as well as two initial mass functions (IMFs) (Salpeter(1955) - Salpeter, 1955 and Chabrier (2003) for Bruzual & Charlot (2003) ; and salpeter(1955) and Kroupa (2001) (Kroupa, 2001) for Maraston (2005)). The spectrum of the composite stellar populations,is then computed by combining the SFH with the grid containing the evolution of the spectrum of an SSP with steps of 1Myr.
Since we need to take into account the difference in reddening between young populations and old populations S. Charlot and Fall, 2000 during the dust attenuation step of the modeling process, the spectra of old and young stars are computed separately.

4.1.2.3. Nebular Emissions

After the first spectrum has been computed by combining the SFH and the SSP, the contributions due to the nebular emissions can be added. Once again CIGALE relies on a pre-computed database. This database has been generating using CLOUDY 13.01(Ferland, Korista, Verner, et al., 1998, Ferland, Porter, van Hoof, et al., 2013) to compute nebular templates based on Inoue, 2011 contains relative intensities of 124 lines from Hii regions from Heii at 30.38 nm to [Nii] at 205.4μm These templates are parametrized according to a ionisation parameter U, and the metallicity Z (assumed to be the same as the stellar metallicity) . After having selected a given template (based on U, and Z), which gives line luminosities normalized to the ionizing photon luminosity, the spectrum of emission lines is computed. Each line has a Gaussian shape with a given width. Those lines are then multiplied by the ionizing photon luminosity which was computed with the intrinsic stellar spectrum. However two main processes affect the ionisation rate of the surrounding gas :

— A fraction of the Lyman continuum can simply escape from the galaxy. This fraction (f_{esc}) varies from one galaxy to the other (Inoue, Iwata, and Deharveng, 2006 ; Hayes, Schaerer, Östlin, et al., 2011).

— Another fraction (f_{dust}) can be absorbed by dust (Inoue, 2001), which results in some dust heating handled by the dust emission models.

Those two processes result in a downscaling factor (Inoue, 2011):

$$k = \frac{1 - f_{dust} - f_{esc}}{1 + \alpha_1(T_e)/\alpha_B(T_e) \times (f_{dust} + f_{esc})}$$

with $\alpha_1 = 1.54 \times 10^{-19} m^3 s^{-1}$ and $\alpha_B = 2.58 \times 10^{-19} m^3 s^{-1}$(Ferland, 1980)
The hydrogen nebular continuum is computed in the same way as the emission lines ; and the other elements continua are considered negligible.

4.1.2.4. Attenuation laws

Galaxies contain dust, and this dust is very efficient at absorbing short-wavelength radiation. The energy absorbed from the UV to the NIR is then re–emitted in the mid– and far-IR. CIGALE is based on this energy balance The vast literature focussed studying attenuation laws in galaxies (e.g. Wild, Stéphane Charlot, Brinchmann, et al., 2011;Steidel, Strom, Pettini, et al., 2016; Lo Faro, Buat, Roehlly, et al., 2017; Buat, Boquien, Małek, et al., 2018), shows the need for the attenuation laws to be highly flexible in order to properly model the diversity of observed curves. CIGALE implements two ways of modelling attenuation curves: the one of S. Charlot and Fall, 2000 , and flexible laws inspired from the starburst curve (Calzetti et al. 2000).

Charlot & Fall (2000) The idea behind this first model is to assume two different attenuation curves to take into account both the birth dust cloud of young stars and the Interstellar medium. Both attenuation curves have the same general parametrization :

$$A(\lambda) \propto \lambda^{\delta}$$

with $\delta = -1.3$ for the birth cloud attenuation and $\delta = -0.7$ for the ISM attenuation. Young stars are affected by both attenuation processes, while old ones are only affected by the ISM attenuation.

Starburst curve A second approach is to resort to an empirical starburst attenuation curve (D. Calzetti, Armus, Bohlin, et al., 2000 ; Leitherer, Daniela Calzetti, and Martins, 2002) and to allow for some flexibility by modifying the slope δ and adding a UV bump(parametrized by its central wavelengthλ_0, width γ and amplitude E). The resulting attenuation curve is given by

$$A(\lambda) \propto k_\lambda \times \left(\frac{\lambda}{550\text{nm}}\right)^{\delta} + D_\lambda$$

with

$$D_\lambda = \frac{E\lambda^2\gamma^2}{(\lambda^2 - \lambda_0^2)^2 + \lambda^2\gamma^2}$$

4.1.2.5. Dust emission

The modeling of dust emission is a very active domain of research, requiring a high level of flexibility. CIGALE implements three different sets of models: the Dale et al. (2014)empirical templates (Dale, Helou, Magdis, et al., 2014), the Draine & Li (2007) and 2014 models (Draine, Dale, Bendo, et al., 2007,Draine, Aniano, O. Krause, et al., 2014), and the Casey (2012) analytic model (C. M. Casey, 2012).

dale 2014 The dust templates of Dale et al. (2014) are based on a sample of nearby star-forming galaxies originally presented in Dale and Helou, 2002. This model excels by its simplicity : a single quantity α is used to parametrize the star-forming component a power law slope of the dust mass distribution over heating intensity. However this simplicity comes at a cost : the PAH emissions show very little diversity with varying α,which can be problematic for some galaxies which are known to have only little PAH emission (e.g. Engelbracht, Gordon, Rieke, et al., 2005)

dl2007 and dl2014 The main feature of the Draine & Li (2007) templates is to divide the dust emission into two components.

— The dust illuminated by a single radiation field $Umin$, modelling the diffuse dust emission heated by the general stellar population.

— The dust illuminated with a variable radiation field going from $Umin$ to $Umax$ through a power–law parametrized by α, modelling the dust linked to star-forming regions.

Those two components represents relative fraction γ and $1-\gamma$ of the total dust mass. Finally the mass fraction of the PAH, is set as a parameter $qpah$. Those models are very flexible and can account for very different physical conditions with a variety of radiation fields and a variable PAH emission, at the cost of a broader parameter space to explore during the fit to an observation. For the purpose of this section, the Draine & Li (2014) templates mainly extends the range of possible parameter values.

casey2012 Finally CIGALE also implements the analytic model of Casey (2012). the module depends on three parameters: the temperature and the emissivity index of the dust, and a mid-IR power law index. While less physically motivated than the Draine & Li (2007) models and not based on observations as the Dale et al. (2014)templates, the Casey (2012) models are very flexible but do not include PAH emissions.

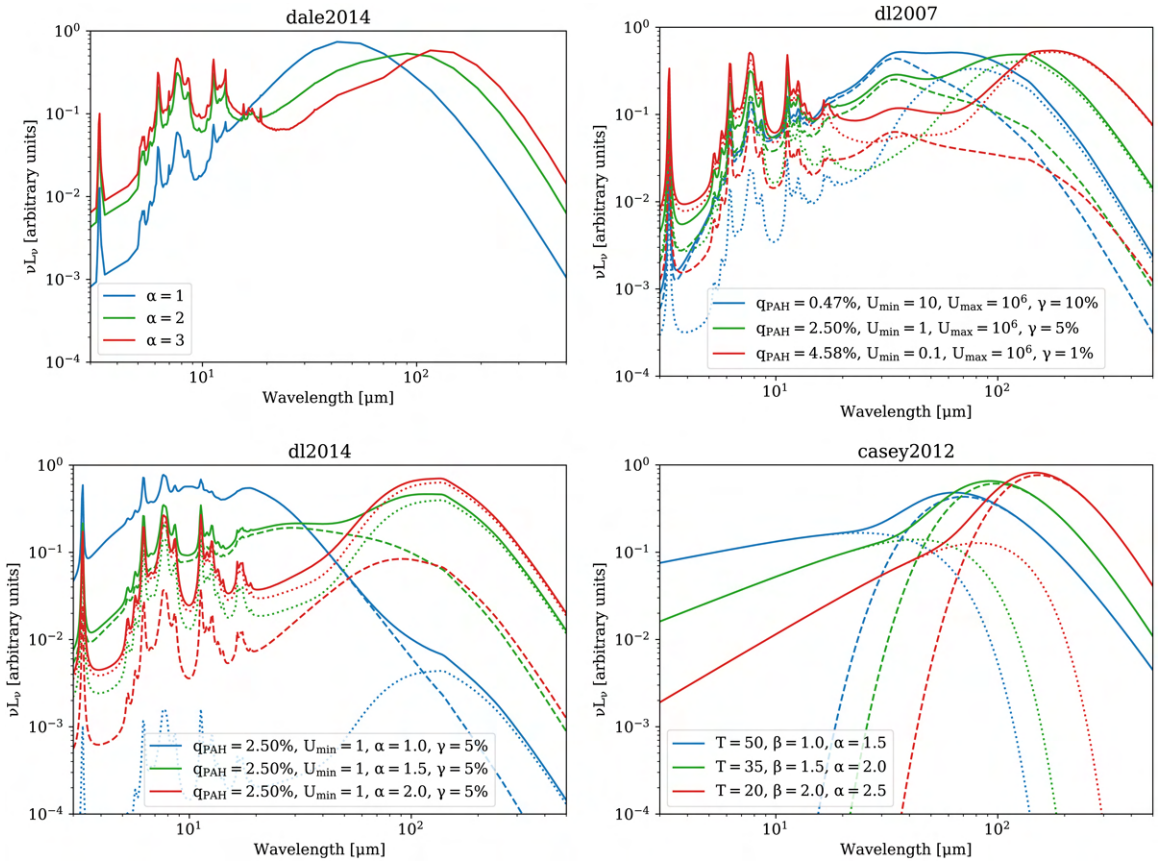

Figure 4.2. – Illustration of the difference between the different attenuation models: dale2014 (top–left), dl2007 (top–right), dl2014 (bottom–left), and casey2012 (bottom–right). Each color corresponds a different set of parameters. The solid lines represent the total SED, summing up the different components specific to each model (e.g diffuse and star-forming for the two Draine and Li models). The smoothness of the SED resulting from casey2012 is due to the absence of PAH emissions in the model.Figure extracted from Boquien, Burgarella, Roehlly, et al., 2019.

4.1.2.6. Redshifting

The final step of CIGALE's modelling pipeline is the 'redshifting' module. The spectrum is redshifted and dimmed by multiplying the wavelengths by $1+z$ and dividing the spectrum by $1+z$. Finally he model of Meiksin, 2006 is used to account for the IGM absorption.

4.1.3. Statistical Inference

CIGALE allows users to model each physical process at play in galaxies' light emissions. From this modeling we obtain spectra, which can then be compared to observations to infer the different physical properties of interest thanks to a simple statistical

model. CIGALE currently implements photometric SED fitting via a grid-based importance sampling approach :

— The user selects the modules they wish to use for each physical component.

— The user sets a list of values for each parameter of each module.

— CIGALE creates the grid of every possible combination of parameter values.

— CIGALE computes the spectra associated with each set of parameters.

— CIGALE computes the likelihood of each set of parameters by computing the χ^2 distance from the observation to each simulation.

However this approach suffers from the exponential growth of the number of spectra to compute with the number of free-parameters. This means the user has to either restrict the number of values for each parameter, use simpler physical models, or set some parameters to a fixed value altogether. When dealing with a vague prior knowledge but highly informative data, this necessary relative coarseness of the sampling grid leads to biases and poor uncertainty estimates.

4.2. Neural Network approximations

As the goal of this work is to provide an effective methodology for Bayesian inference, we need the posterior distribution sampling algorithm to efficiently explore the parameter space. In particular, in order to use the TAMIS algorithm described in chapter 3, we need to be able to sample randomly from a proposal distribution without worrying about a grid of precomputed values. However leaving the parameter grid makes it difficult to use CIGALE directly as implemented.

Indeed several modules (Nebular and SFH/BC03 in particular) exploit the efficiency of a grid. The *Nebular* module for instance relies on using a precomputed database of emission lines calculated by CLOUDY which simply cannot be supplemented on the fly. An other problem is that the computational efficiency of CIGALE in general relies on the combination of its modularity and the grid sampling. Assuming we want to use 4 modules, with 3 free-parameters in each and 5 values for each parameter : The complete grid has size 5^{3*4} and we have to compute that many ($244,140,625$) spectra. But the intrinsic computation of each module only have $5^3 = 125$ different sets of parameters. Storing those results allows CIGALE to exponentially reduce the number of actual computations, but requires the grid to do so.

We therefore propose to approach this process of physical modeling using some neural networks approximations. In order to control the interpretability of errors and the modularity of CIGALE as a whole, we only use approximations for certain modules (see Fig. 4.3). These new "deep" modules must be able to be as interchangeable as possible with the original modules, be used both for grid use and for randomized use, and keep the inputs and outputs of the original modules. This allows them to fit naturally into the CIGALE simulation process and not to require new learning of a "global" approximation neural network in the event of customization of one of the modules by a user. The learning itself is also simplified since the parameter space from which we construct the learning set remains small and the approximation errors easily identified at the level of each module.
We now describe the chosen approximation method which is similar to the one presented in Alsing, Peiris, Leja, et al., 2020

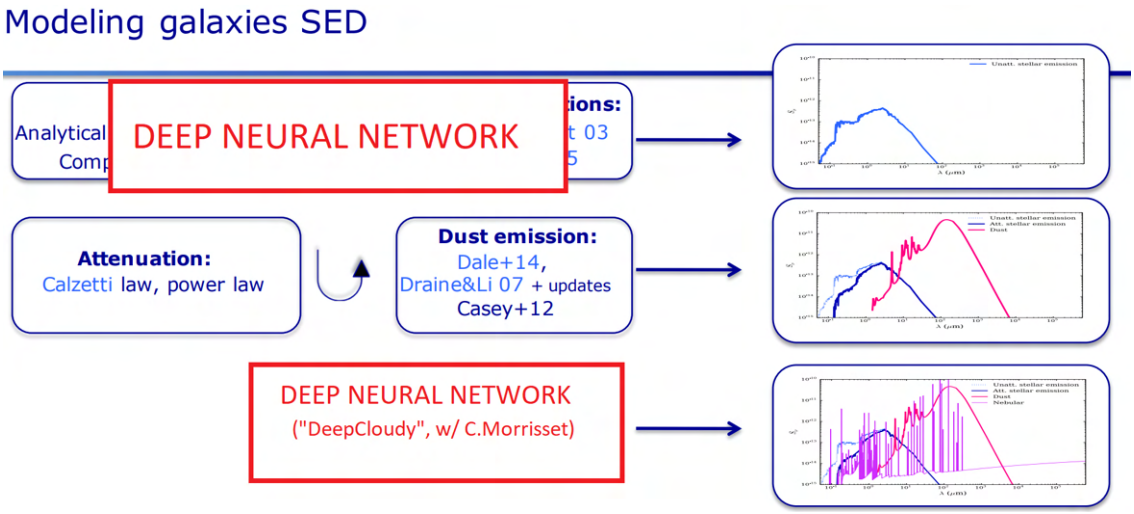

Figure 4.3. – Our proposed approach to extend CIGALE's framework : combining Neural Network approximations replacing the expensive or unwieldy computation steps while keeping the exact physical modules as much as possible. This reduces the computational cost, and allows for interpolation of precomputed values while retaining CIGALE's modularity and explainability, confining the approximation and black-box aspects to specific part of the model.

4.2.1. Methodology

Whether to approximate full spectra (e.g the young and old stellar emissions) or only a few hundred emission lines (Nebular), we implement a two-step approximation: We first perform a Principal Component Analysis (below PCA) on the target in order to reduce the dimension of the problem - therefore the number of values to approximate - and we train a neural network to estimate the PCA coefficients using the physical properties as input.

This two-step process allows for the use of a very simple neural network (fully connected with few neurons) which avoids the infamous difficulty of tuning a deep Neural Network with complex architecture despite advances in automatic hyper-parameter tuning (e.g Bergstra and Bengio, 2012 ; Bengio, 2012 ; Lam, Ling, Leung, et al., 2001 ; Golovin, Solnik, Moitra, et al., 2017)

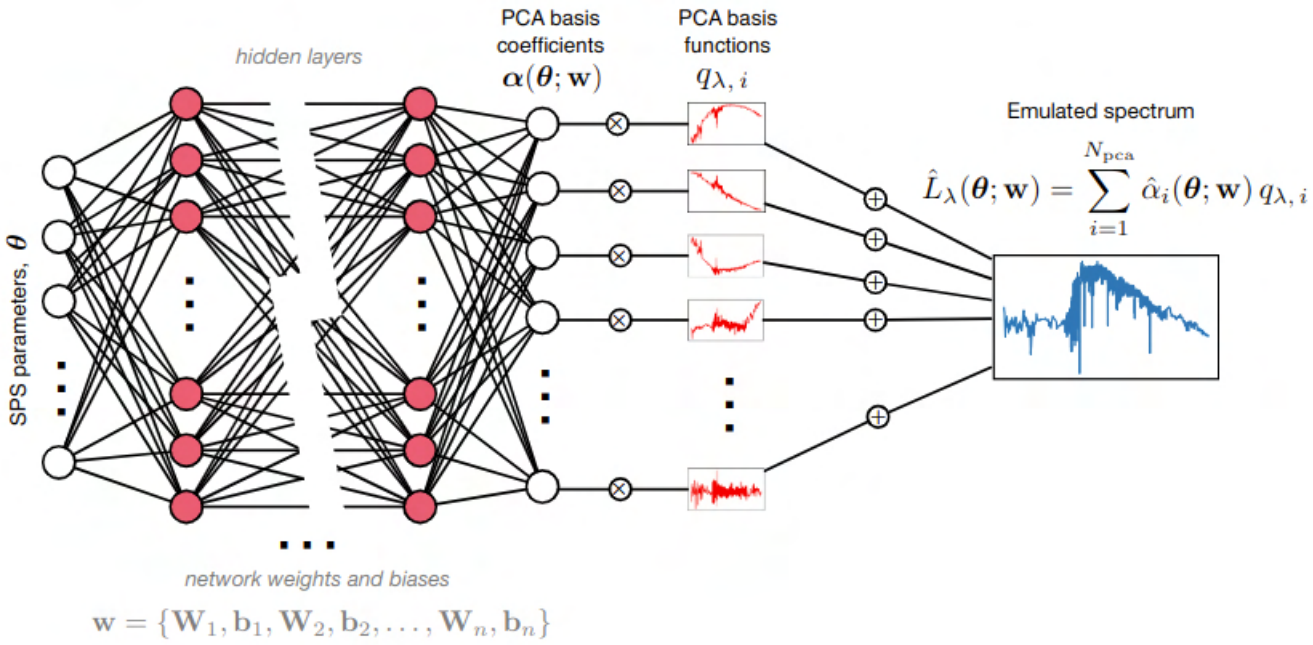

Figure 4.4. – The two-step Neural network approximations we use to replace specific modules of CIGALE. A PCA reduction is first performed on the spectra composing the training set. A Neural network is then trained to approximate the PCA coefficients corresponding to each spectrum using the physical parameters as input. The approximated spectrum is then computing inverting the PCA reduction. Figure adapted from Alsing, Peiris, Leja, et al., 2020

4.2.2. Star population contributions

We start by focusing on the biggest computational burden regarding our work with the current CIGALE implementation : The computation of the stellar contributions from the combination of SFH and the SSP. It is also one of the easiest step to replace as none of its inputs depend on the result of a previous computation. The use of the modules modeling the SFH and the SSP being done jointly for the simulation of the stellar contribution, we choose to create a single joint approximate model taking both of SFH and SSP parameters as input. This single model approximate all the quantities necessary for the downstream modules of CIGALE:

— the spectrum of the young stellar population

Table 4.1. – CIGALE parameters used to generate the first train set for our Stellar emissions neural approximator. All parameters are sample uniformly in the reported interval, except for τ_{main} which is sample in log scale to account for the non linearity of its effect on the SED.

Parameter	Values
Delayed SFH	
age_{main} (Myr)	$[500; 10000]$
τ_{main} (Myr)	$[10^2; 10^5]$
τ_{burst} (Myr)	$[100; 10000]$
f_{burst}	$[0; 0.0.5]$
age_{burst} (Myr)	$[1; 500]$
'bc03'	
IMF	$Chabrier$
$Metallicity$	$0.0001, 0.0004, 0.004, 0.008, 0.02, 0.05$

— the spectrum of the old stellar population

— The number of ionizing photons n_{ly} used to rescale emission lines.

A first training set is created by simulating 400,000 spectra with CIGALE from the following distribution distribution described in table 4.1.

After checking for potential systematic errors (specific parts of the parameter space where the error approximation is greater), we produce a second 400,000 spectra sample with CIGALE, oversampling the small values of age_{burst} . This new training sample (described in is table 4.2) is combined to the first one, and our neural PCA approximation is learned on the joint training set. This modification in the parameter distribution of the training set might introduces biases to be studied in a future work but participate to the increased accuracy of the estimator (Fig 4.5)

Table 4.2. – CIGALE parameters used to generate the second train set for our Stellar emissions neural approximator. It is coupled with the set presented in 4.1 to oversample regions in the parameter space where the neural network errors are too important.

Parameter	Value
Delayed SFH	
age_{main} (Myr)	$[500; 10000]$
τ_{main} (Myr)	$[10^2; 10^5]$
τ_{burst} (Myr)	$[100; 10000]$
f_{burst}	$[0; 0.0.5]$
age_{burst} (Myr)	$[1; 10]$
'bc03'	
IMF	$Chabrier$
$Metallicity$	$0.0001, 0.0004, 0.004, 0.008, 0.02, 0.05$

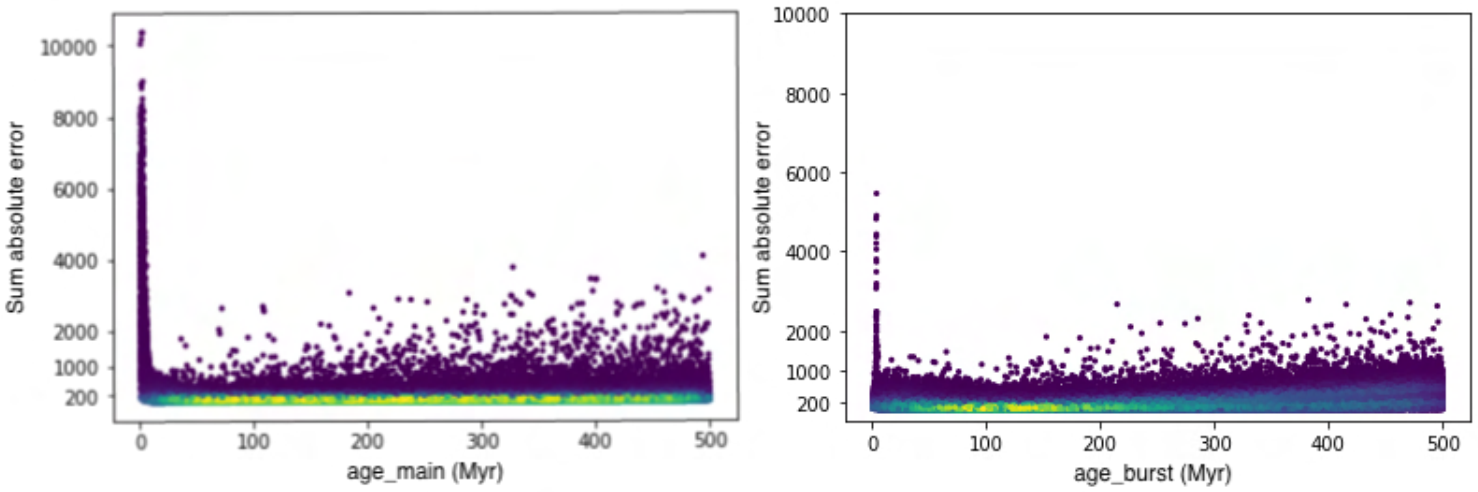

Figure 4.5. – Heatmaps of the spectra approximation errors (y-axis) as a function of age_{burst} (x-axis) on the test set. Left : after training on the first training set only. There is a clear explosion of the errors for very low values of age_{burst}. Right : After training on the completed training set. The catastrophic errors for low age_{burst} have been greatly reduced.

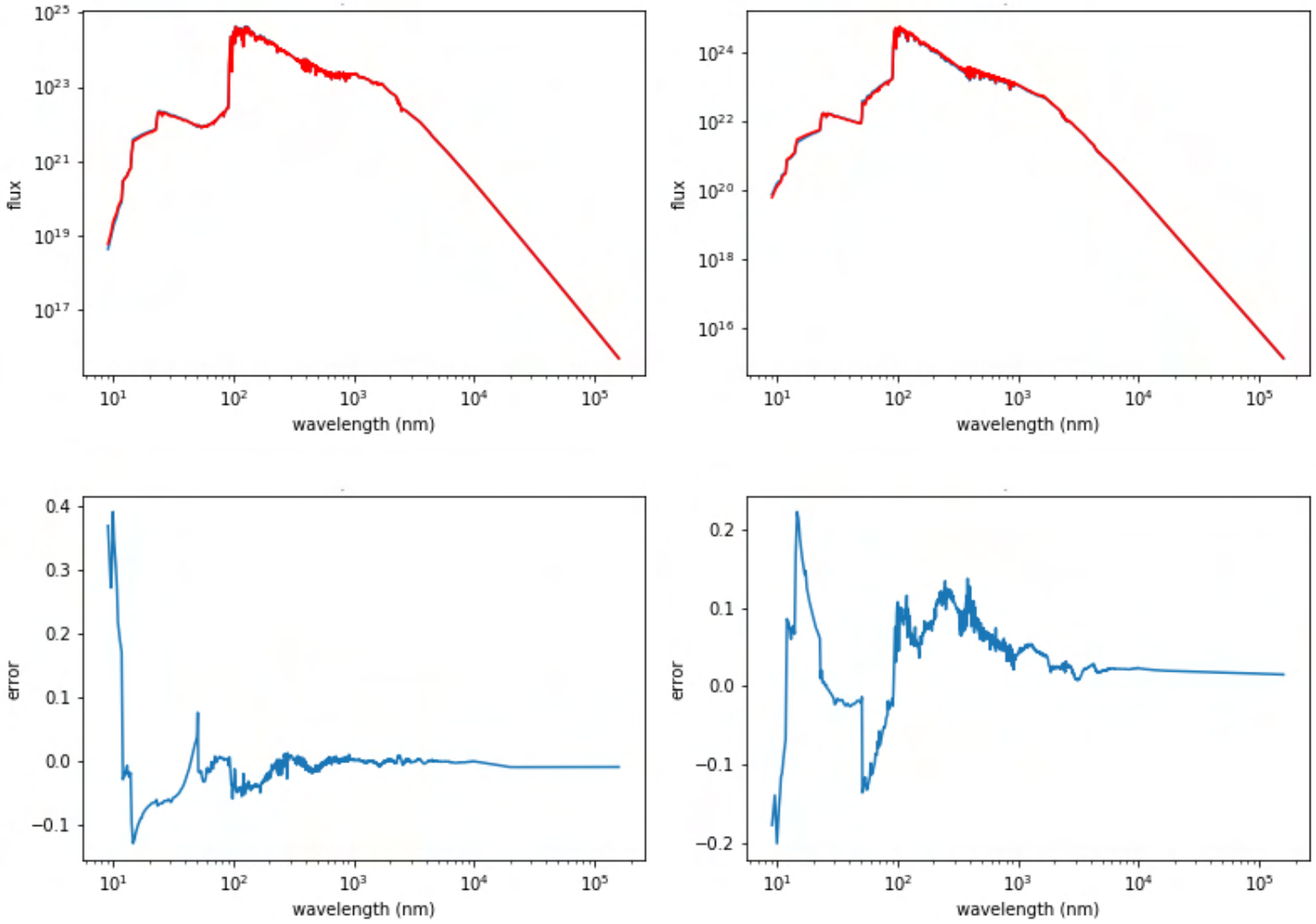

Figure 4.6. – Error of the Stellar emissions approximation. Top : the true and approximated spectrum for the 50th percentile and 99th percentile of errors. Bottom : The relative error of both approximations. The clear error increase at low wavelengths is due to both a large variability in the training set and a flux magnitudes lower than the ones at higher wavelength.

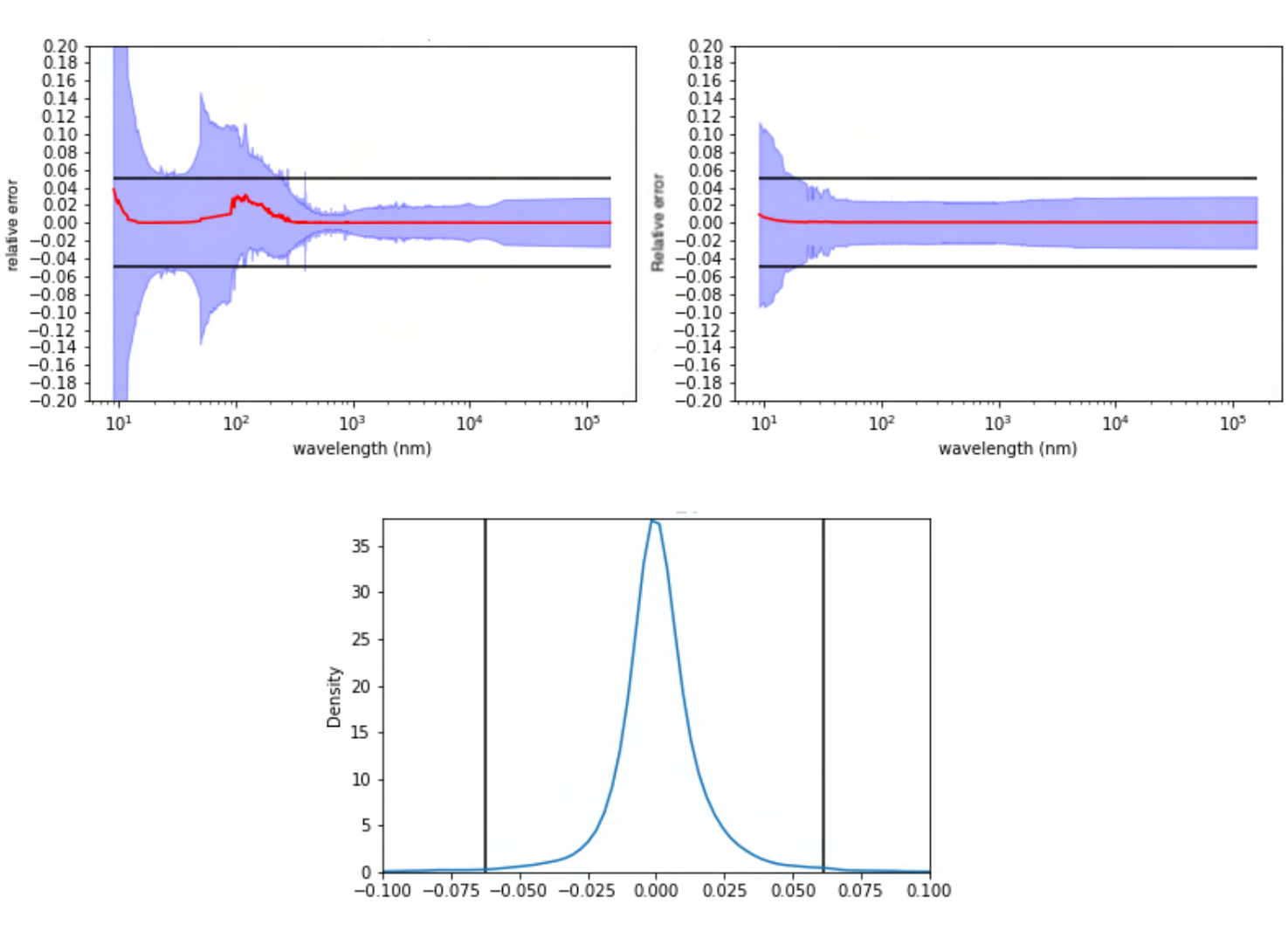

Figure 4.7. – Error of the Stellar emissions approximation. Top : The relative flux error (y-axis) along wavelength (x-axis) across the entire test set The red line is the mean of the distribution and the 5th and 95th percentiles are in blue.Bottom : The relative error of the number of ionizing photons estimations. In both cases the black lines represents a 5% relative error. As seen in Fig. 4.6, the error greatly increase at very low wavelength. However we expect to mitigate this error as the number of ionizing photos in directly estimated instead of being derived from the spectrum

The module implementing this approximation is able to process 10,000 about 500 times faster than the original, with the actual spectra computation being about 10^3 times faster as presented in table 4.3.

Table 4.3. – Wall-clock time (in seconds) for the original modules(*sfhdelayed* and *bc03*) and their 'deep' counterpart to process 10,000 SEDs using CIGALE, with random input parameters, on a single CPU core.

Step	Original	Deep Approximation
Computation time		
Spectra computation	~ 4300	~ 4
Other	5	5
Total	~ 4300	~ 9

4.2.3. Nebular emissions

We replace the CIGALE database pre-computed by CLOUDY by an approximation. For this we use the 3MDB database developed by Morisset, Delgado-Inglada, and Flores-Fajardo, 2015 [1].
We start by removing the lines with negligible flux (for which the median ratio with $H\beta$ is less than $1e-4$) . The variable "age" is first normalized, then all the input parameters are normalized while the fluxes of the lines are normalized by the flux values in $H\beta$. The database is then split 80-20 into training and testing, then the training set is again split 80-20 into training and validation.
A dimension reduction by PCA is then performed on the values of the lines (still normalized by $H\beta$) in the training and validation sets (the decomposition being learned only on the training set), keeping a number of components such that the percentage of explained variance reaches 99.95%.
Finally, we use scikit-optimize (Head, Kumar, Nahrstaedt, et al., 2021) combined with scikit-learn (Pedregosa, Varoquaux, Gramfort, et al., 2011)to learn a neural network taking the 5 CLOUDY parameters as input and estimating the values of the associated lines transformed by the PCA. The neural network is a multilayer perceptron composed of 4 layers of 8 to 256 neurons. We use a hyperbolic tangent as the activation function and L-BFGS as the optimizer. The setting of the number of neurons per layer is performed Bayesian optimization of the MSE error on the validation set (specifically by Gaussian process Bandit optimization, Srinivas, A. Krause, Kakade, et al., 2010 ;Golovin, Solnik, Moitra, et al., 2017).

When calling the Deep-nebular module, a set of parameter sets corresponding to the CLOUDY parameters is read, the PCA components are estimated by the network. These estimated values are then retransformed by inversion of the PCA into the values of the flux of the lines normalized by the value of the flux of $H\beta$ and $logQ$,the number of ionizing photons emitted per second. The flux in $H\beta$ and $logQ$ are estimated separately by another neural network trained in a similar way (without PCA).
Once these approximate values have been estimated, they are managed exactly as

1. https://sites.google.com/site/mexicanmillionmodels

for the standard Nebular module (fraction escaped or absorbed, width of the lines, multiplication by the number of photons of the Lyman continuum).
The approximation errors for each supported line in the dataset is presented in Fig. 4.8

The nebular continuum is considered negligible compared to the stellar continuum and is not implemented for the moment.

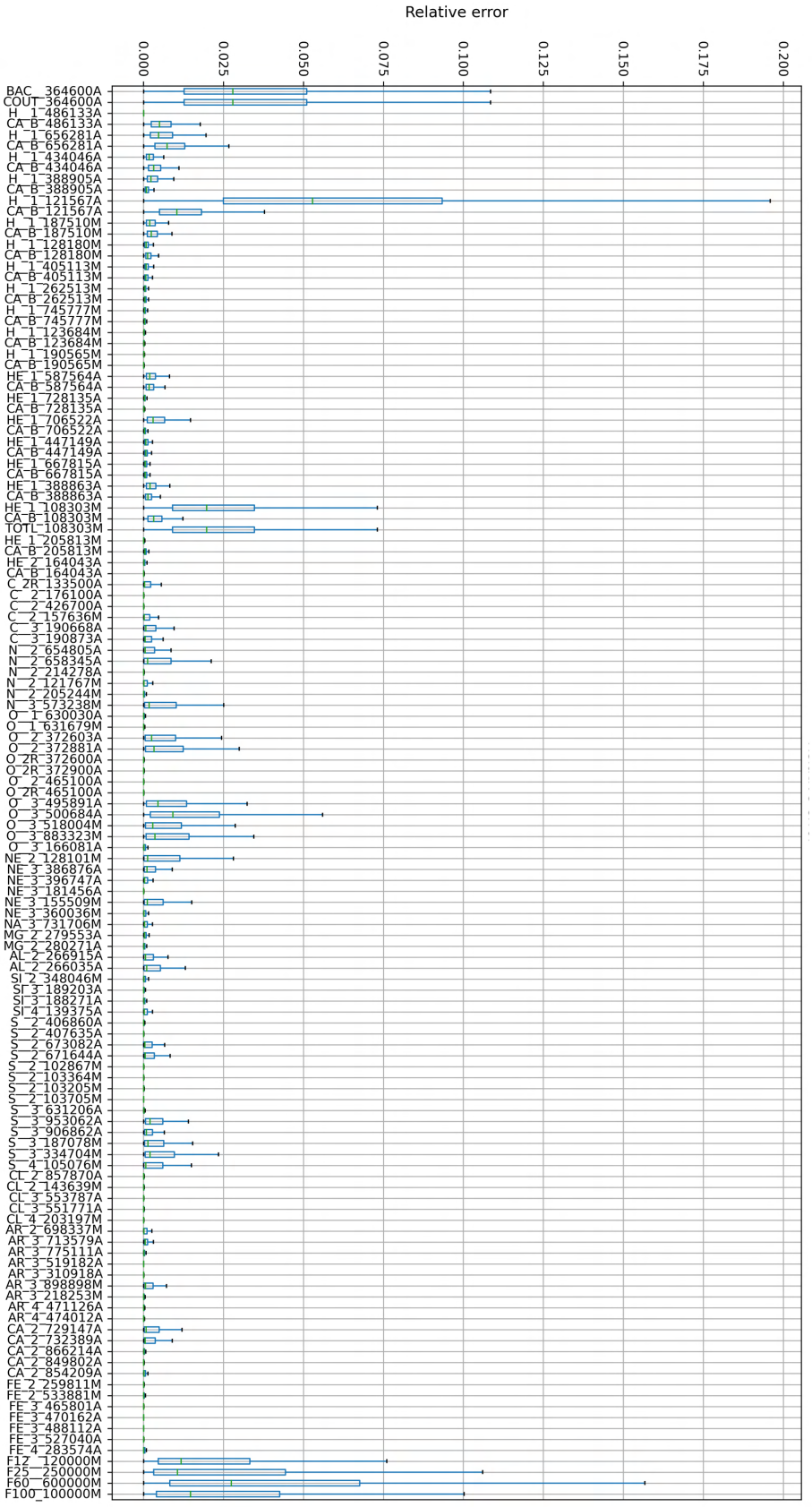

Figure 4.8. – For each line supported by our CLOUDY approximation (x-axis), boxplots of the absolute relative error (normalized by Hβ)

5. Bayesian spectro-photometric SED Fitting

Sommaire

5.1. Introduction

Our goal is to propose an inference model to fit physical parameters to each observed galaxy spectrum. For each galaxy, we have a spectrum x represented by a vector of N positive real numbers $x_1 \dots x_N$. Those numbers are measures of the flux density at wavelengths $\lambda_i, i = 1, \dots N$. The x_i can come either from spectroscopy, or from broadband photometry (in which case we consider the flux density integrated over the filter's width).

Since the resolution of spectroscopic measurement is far greater than the number of broadband filter, our observed x is composed of many more points from spectroscopy (several thousands) than from photometry (about 5 to 30). Those spectroscopic datapoints are probing a small part of the spectrum wavelength with a high-resolution, whereas the broadband filters can be collectively covering the spectrum from UV to FIR , but with a very low resolution.

We assume that the physical model is implemented as a blackbox. This blackbox takes as input a parameter of physical properties θ and outputs the corresponding theoretical spectrum $S(\theta)$. From this physical model and an observed spectrum x_{obs}, we wish to perform Bayesian inference and estimate the posterior distribution $p(\theta|x_{obs})$ and its marginal likelihood $p(x_{obs})$, given a prior distribution $p(\theta)$.

The main modeling questions are the way to handle

— nuisance parameters,
— the difference in the spectrum sampling resolution due to the different types of measurements,
— the plausible differences and imprecisions regarding the estimated redshift when dealing with high-resolution measurements,
— the specific treatment of emission lines,
— the missing (censored) values expressed as upper or lower limits on an observed flux.

The next section explores each of those problems.

5.2. A general SED fitting Bayesian Model

As is usual in the SED fitting literature (Walcher, Groves, Budavári, et al., 2011 ; Roehlly, Burgarella, Buat, Giovannoli, et al., 2011 ; Boquien, Burgarella, Roehlly, et al., 2019 ; G. Kauffmann, T. M. Heckman, S. D. M. White, et al., 2003), we assume the likelihood to be Gaussian with variance Σ and mean $\alpha \times S(\theta)$:

— Σ is most often diagonal;

— $S(\theta)$ is the spectrum corresponding to physical parameters θ. S is a black-box, highly non-linear function encoding the physical modeling;

— α is a scaling factor linked to the mass of the galaxy and considered here as a nuisance parameter.

Therefore the likelihood is

$$p(x|\theta,\alpha) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\alpha\times S(\theta))^{\top}\Sigma^{-1}(x-\alpha\times S(\theta))\right] \tag{5.1}$$

We propose to integrate it out over the nuisance parameter as is common in Bayesian inference, hence to consider

$$p(x|\theta) = \int p(x|\theta,\alpha)p(\alpha)d\alpha$$

However since computing exactly this quantity would require numerical integration over the α space, spanning several orders of magnitudes, for each evaluation of the likelihood, we rely on a Laplace approximation (Tierney and Kadane, 1986). Indeed as the profile likelihood is extremely concentrated around its unique maximum, the prior distribution $p(\alpha)$ to cover galaxies of greatly varying sizes, and the number of observed datapoints is large (especially due to the spectroscopy). If we set

$$\begin{aligned}\hat{\alpha}_\theta &= \underset{\alpha}{\operatorname{argmax}}\, p(x|\theta,\alpha)\\ &= \frac{x^T\Sigma^{-1}S(\theta)}{S(\theta)^T\Sigma^{-1}S(\theta)}\end{aligned}$$

the Laplace approximation leads to

$$\begin{aligned}p(x|\theta) &= \int p(x|\theta,\alpha)p(\alpha)d\alpha\\ &\approx p(x|\theta,\hat{\alpha}_\theta)N^{-\frac{1}{2}}.\end{aligned}$$

This derivation also justifies the use of the maximum profile likelihood in CIGALE, when the number of observed datapoints along the spectrum is large enough.

5.3. Redshift, covariance and binning

As the redshift z shifts the wavelengths λ_i, the direct comparison of theoretical and observed spectrum wavelength by wavelength would require perfect calibration of the estimated z as well and the exact same sampling resolution between the observation and the simulation, especially around the emission and absorption lines. To avoid all those complications with a minimum of approximation and computation cost, we propose to integrate the spectroscopic density fluxes over wavelength bins. The number and width of the bins would become tuning parameters of the model, reducing furthermore the computational cost of each likelihood evaluation.
Let us assume that we have n_{bins}, and that, for bin number j we have observed N_j fluxes $x_1^j, \dots, x_{N_j}^j$ over wavelength bandwidth of size $\Delta\lambda_1^j, \dots, \Delta\lambda_{N_j}^j$. We consider the average

$$B_j = \frac{1}{\sum_{i=1}^{N_j} \Delta\lambda_i^j} \sum_{i=1}^{Nj} \Delta\lambda_i^j x_i^j. \tag{5.2}$$

Likewise on the expected fluxes $\hat{\alpha}_\theta S(\theta)$, we can compute an expected average $B_j(\theta)$.

Since the vector of all observed fluxes is a Gaussian vector, the random variable B_j is Gaussian and

$$B_j \sim \mathcal{N}\left(B_j(\theta), \Sigma_{B_j}\right)$$

where

$$\Sigma_j = \sum_{i=1}^{N_j} \sum_{l=1}^{N_j} \Delta\lambda_i^j \Delta\lambda_l^j \operatorname{Cov}\left(x_i^j, x_l^j\right).$$

Moreover

$$\begin{aligned}
\operatorname{Cov}(B_j, B_k) &= \operatorname{Cov}\left(\frac{1}{\sum_{i=1}^{N_{B_j}} \Delta\lambda_i^j} \sum_{i=1}^{N_{B_j}} \Delta\lambda_i^j x_i^j, \frac{1}{\sum_{i=1}^{N_{B_k}} \Delta\lambda_i^k} \sum_{i=1}^{N_k} \Delta\lambda_i^k x_i^k\right) \\
&= \frac{1}{(\sum_{i=1}^{N_{B_j}} \Delta\lambda_i^j)(\sum_{i=1}^{N_{B_k}} \Delta\lambda_i^k)} \operatorname{Cov}\left(\sum_{i=1}^{N_{B_j}} \Delta\lambda_i^j x_i^j, \sum_{i=1}^{N_{B_k}} \Delta\lambda_i^k x_i^k\right) \\
&= \frac{1}{(\sum_{i=1}^{N_{B_j}} \Delta\lambda_i^j)(\sum_{i=1}^{N_{B_k}} \Delta\lambda_i^k)} \sum_{i=1}^{N_{B_j}} \sum_{l=1}^{N_{B_k}} \Delta\lambda_i^j \Delta\lambda_l^k \operatorname{Cov}\left(x_i^j, x_l^k\right)
\end{aligned}$$

5.4. Combining spectroscopy and photometry

The number of datapoints is far greater in spectroscopy than in photometry. Not taking this disparity into account would result in a fitting procedure completely neglecting the general shape of the spectra to concentrate only in the small scale variation in the spectroscopy. It is therefore necessary to consider both contributions to the likelihood separately and to weight them adequately. This weighting could take several forms, we propose to consider a model of the form

$$p(x|\theta) = p(x_{\text{spectro}}|\theta)^{P_{\text{spectro}}}\, p(x_{\text{photo}}|\theta)^{P_{\text{photo}}} \tag{5.3}$$

where the weights P_{spectro} and P_{photo} are new hyperparameters of the model. Intuitive choices for those hyper-pameters could be $P_{\text{spectro}} = P_{\text{photo}} = 1$ (not accounting for the nature of measurements), or follow the relative wavelength coverage of each data type:

$$P_{\text{photo}} = \frac{\Delta\lambda_{\text{photo}}}{\Delta\lambda_{\text{total}}}, \quad P_{\text{spectro}} = \frac{\Delta\lambda_{\text{spectro}}}{\Delta\lambda_{\text{total}}} \tag{5.4}$$

or

$$P_{\text{photo}} = 1, \quad P_{\text{spectro}} = \frac{\Delta\lambda_{\text{spectro}}}{\Delta\lambda_{\text{photo}}}.$$

As we separated the likelihood into two weighted terms, we need to reformulate the computation method for α. An easy computation leads to

$$\begin{aligned}
\hat{\alpha}_\theta &= \underset{\alpha}{\text{argmax}}\; p(x|\theta,\alpha) \\
&= \underset{\alpha}{\text{argmax}}\; p(x_{\text{spectro}}|\theta,\alpha)^{P_{\text{spectro}}}\, p(x_{\text{photo}}|\theta,\alpha)^{P_{\text{photo}}} \\
&= \frac{P_{\text{photo}} x_{\text{photo}}^T \Sigma_{\text{photo}}^{-1} S_{\text{photo}}(\theta) + P_{\text{spectro}} x_{\text{spectro}}^T \Sigma_{\text{spectro}}^{-1} S_{\text{spectro}}(\theta)}{P_{\text{photo}} S_{\text{photo}}(\theta)^T \Sigma_{\text{photo}}^{-1} S_{\text{photo}}(\theta) + P_{\text{spectro}} S_{\text{spectro}}(\theta)^T \Sigma_{\text{spectro}}^{-1} S_{\text{spectro}}(\theta)}
\end{aligned}$$

The latter can again be plugged into the Laplace Approximation

$$\begin{aligned}
p(x|\theta) &= \int p(x|\theta,\alpha)\, p(\alpha)\, d\alpha \\
&\approx p(x|\theta,\hat{\alpha}_\theta) N^{-\frac{1}{2}}.
\end{aligned}$$

5.5. Emission lines

We expect emission lines bring a crucial amount of information about certain parameters (especially about the metallicity and nebular properties), but they are neglected by the previously detailed likelihood model. Indeed as they span relatively narrow wavelengths, their contribution to the likelihood is significantly affected by the number of spectroscopic measurements probing the continuum. Two possible solutions would be to either identify the lines in the spectroscopy and consider them

as separate inputs (e.g Boquien, Burgarella, Roehlly, et al., 2019, Franzetti, Scodeggio, Garilli, et al., 2008, Bowman, Zeimann, Nagaraj, et al., 2020) or simply let them be binned with the rest of the spectroscopy.

5.6. Censored photometric values

In some cases, the photometric fluxes are only partially known. Depending on the observation instruments characteristics, we may have lower or upper bounds for a given flux : the galaxy has been observed, but the measurement is outside the measurement range of the instrument (either too low, resulting in an upper bound, or too high, resulting in a lower bound).
To account for those censored values, we split the fluxes in three categories, namely:

— not censored when $x_j \in [L_j, U_j]$ the measurement range,

— below the lower bound when $x_j < L_j$

— avove the upper bound when $x_j > U_j$

In the two last cases, the contribution of the j^{th} measurement to the likelihood is $p(x_j < L_j|\theta)$ when below the lower bound and $p(x_i > U_i|\theta)$ when above the upper bound.

5.7. Sampling algorithm

Once the statistical model is chosen, we need to be able to compute the posterior distribution. This is necessary to perform either model choice through the computation of the marginal likelihoods and/or the associated Bayes factor, or to perform parameter estimation. Since our likelihood model is based on the numerical computation of spectrum for each likelihood evaluation, we have a few constraints regarding the sampling algorithm to use.

— Most of the computational cost in the model evaluation is due to CIGALE, that is to say the numerical computation of $\hat{\alpha}_\theta S(\theta)$ for a new value of θ

— We do not have access to the gradient of $\hat{\alpha}_\theta S(\theta)$, so we cannot compute the gradient of the likelihood either.

— If most of the physical parameters of interest can be continuously sampled, some of them are discrete in nature.

— As the intent is to fit a large sample of observed SEDs, we cannot rely on tuning the initialization manually for each one.

— The computation of $\hat{\alpha}_\theta S(\theta)$ is time consuming but highly parallelizable over different values of θ.

Those constraints make impractical the use of most Monte Carlo algorithms commonly used in Bayesian inference such as MCMC. The unavailability of the gradient prevents the use of Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, et al., 1987 ; R. Neal, 2011) and its variants (Homan and Andrew Gelman, 2014; Carpenter, Andrew Gelman, Hoffman, et al., 2017) ; the lack of initialization tuning would make most Importance Sampling schemes unreasonably expensive (Liu, 2001; Beaujean and Caldwell, 2013 ; Cornuet, Marin, Mira, et al., 2012) ; more sequential algorithms - like Metropolis-Hastings Metropolis and Ulam, 1949; or the Gibbs sampler S. Geman and D. Geman, 1984 ; Casella and George, 1992 - would not benefit from the parallelization ; and finally Sequential Monte Carlo methods would require tuning a Markov kernel for both continuous and discrete parameters (Doucet, Smith, Freitas, et al., 2001)
The TAMIS algorithm presented in chapter 3 is perfectly suited to accommodate all those constraints. This section presents the sampling methodology for both continuous and discrete parameter spaces using TAMIS.

5.7.1. Sampling the continuous parameters

The proposal distributions used for Adaptive Importance Sampling need to have larger support than the target distribution and be easily updated. A common choice

(Cappé, Douc, Guillin, et al., 2008) is to restrict the family of proposal density to mixtures of Gaussian or Student distributions. As described in chapter 4, nearly all parameters of interest for Astronomers resorting to SED fitting are physical quantities. As such, the prior distributions for each parameter has a compact support, and the value ranges involved varies a lot depending on the specific parameter and its unit (see e.g Table 5.3). Since it is easier to tune the proposal distribution from an unconstrained space, we standardize the intervals and apply a probit transform (see e.g Carpenter, Andrew Gelman, Hoffman, et al., 2017). The parameter sampling is therefore done in the unconstrained space. The transform is then inverted to compute the likelihood, the posterior distribution and the IS weights. Note that if the prior distribution is uniform over the intervals, the transformed distribution is a Gaussian distribution centered on the origin. This allows a systematic initialization of TAMIS regardless of the original parameter spaces.

5.7.2. Sampling the discrete parameters

Since the previously developed TAMIS approach relies on using the EM algorithm to fit a Gaussian Mixture q_t at each iteration, we handle the discrete parameters separately. This approach prevents the adaptive procedure to exploit correlations between the different kinds of parameters (as it could through the estimation of the covariance matrix during the EM step) but remains relevant for obtaining an IS estimate. However in the specific case of TAMIS it leads to computational problems, which we present and solve in the following paragraph.

On one hand one could simply not adjust the proposal distribution over the discrete parameters at all. It would be computationally inefficient, but an IS estimate would remain valid. However such a scheme is completely inadequate for TAMIS. Let us consider a realistic use case where we want to fit 10 continuous parameters, and 5 discrete ones. Assume the prior probability over each discrete parameter is uniform with 6 values each. Let's assume further that among those 6 values, 3 have a near 0 weight in the posterior distribution. Then in the best case (where the continuous parameters are sampled directly from the posterior) $1 - 3^5/6^5 = 97\%$ of any drawn θ would still have negligible weight. This is not only an enormous computational waste. Since TAMIS uses the current estimated ESS to tune automatically the tempering path, this decrease in the ESS biases the TAMIS adaptive step and make extremely difficult the convergence and its assessment.

In order to adapt the distribution over the discrete parameters for iteration $t+1$, the new proposal is the estimated posterior distribution at iteration t contaminated with a uniform distribution. This is akin to Safe Adaptive Importance Sampling (SAIS, see e.g Delyon and Portier, 2021). It ensures that the support of the proposal distribution remains large enough to avoid local optima.

5.8. Inference

Statistical inference is performed using the weighted sample generated by TAMIS using the specified statistical model. This specification involves the choice of the prior distribution and the likelihood hyper-parameters, as inference can be performed using photometry, spectroscopy or both, with a contribution weighted arbitrarily, and can include specific emission lines or not. It also involves the choice of a covariance structure and binning if spectroscopy is used.
From the TAMIS output we use the self-normalized IS estimator of the posterior mean

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \omega_i \theta_i}{\sum_{i=1}^{N} \omega_i}$$

and covariance

$$\hat{\Sigma} = \frac{1}{\sum_{i=1}^{N} \omega_i} \sum_{i=1}^{N} \omega_i (\theta_i - \hat{\mu})^T (\theta_i - \hat{\mu})$$

as well as the 1D credibility intervals.

The Maximum a Posteriori (MAP) is estimated from the empirical maximum in TAMIS

$$\underset{\theta_i}{\text{argmax}} p(\theta_i | x).$$

The Maximum Likelihood estimator is approximated in the same way by

$$\underset{\theta_i}{\text{argmax}} p(x | \theta_i).$$

We also compute the IS estimator of the Marginal Likelihood

$$\frac{\sum_{i=1}^{N} \omega_i}{N}.$$

Finally, for visualization, we perform 2D kernel density estimation of the posterior density over the continuous parameter space to identify possible complex structures in the joint distributions (degeneracies, multimodalities). We also draw simple weighted histograms to visualize the posterior over the discrete parameters.

5.9. Numerical Results

In order to evaluate the performances of our SED fitting tool, we use the dataset from Villa-Vellez (Villa-Vélez, 2021). This galaxy sample was constructed by combining photometry from the well-studied COSMOS2015 (C. Laigle, McCracken, Ilbert, Hsieh, I. Davidzon, P. Capak, Hasinger, Silverman, Pichon, Coupon, Aussel, Le Borgne, Caputi, Cassata, Chang, Civano, Dunlop, Fynbo, Kartaltepe, Koekemoer, Le Fèvre, Le Floc'h, Leauthaud, Lilly, Lin, Marchesi, Milvang-Jensen, et al., 2016b) catalog and the FMOS-

Table 5.1. – Broad bands used in this work

Instrument	**Band**	$\boldsymbol{\lambda}$ ($\boldsymbol{\mu}$**m**)
GALEX	NUV	0.229
CFHT	u'	0.355
SUBARU	B	0.443
SUBARU	V	0.544
SUBARU	r	0.622
Suprime Cam	i'	0.767
Suprime Cam	z'	0.902
HSC	Y	1.019
WFCAM	J	1.250
WIRcam	H	1.639
WFcam	K	2.142
Spitzer	IRAC1	3.6
Spitzer	IRAC2	4.5
Spitzer	IRAC3	4.5
Spitzer	IRAC4	4.5

COSMOS (Kashino, Silverman, Rodighiero, et al., 2013 ; Silverman, Kashino, Sanders, et al., 2015) high-resolution spectroscopy. The sample was carefully selected as to have good SNR for all fluxes and avoid discrepancies between measurements.
As we need a controlled environment to assess our fit quality, we use those sources as the base for a catalog of synthetic galaxies. First, the real sources are fit using the grid implementation of CIGALE. For each galaxy, the best fitting set of parameters is recovered, and then used to simulate a new synthetic SED. This process ensures that our test catalog is realistic in terms of the physical parameters used (as estimated from real observed sources), and the errors of our estimates are easily available (as we know both the simulator and the true values used for the simulation).
From those synthetic SEDs, we extract both photometric (5.1)and spectroscopic data (all the fluxes between 600 et 1800 nm from CIGALE High Resolution spectra). We denote the 15 photometric measurement x^{sim}_{photo} and the spectroscopic ones $x^{sim}_{spectro}$. Finally we add a gaussian noise to both data types by setting the Signal-to-Noise Ratio at 5 for the photometry and 2 for the spectroscopy, and obtain :

$$x_{photo} = x^{sim}_{photo} \times (1+\delta_1), \qquad \delta_1 \sim \mathcal{N}(0, 0.2^2)$$
$$x_{spectro} = x^{sim}_{spectro} \times (1+\delta_2), \qquad \delta_2 \sim \mathcal{N}(0, 0.5^2)$$

We set the data weights $P_{photo} = P_{spectro} = 1$, and the number of bins $n_{bins} = 20$. TAMIS with tuning parameters described in table 5.2. The different steps of our pipeline are illustrated in figure 5.1 : After binning the spectroscopy (including the

Table 5.2. – TAMIS parameters used for the SED fitting

Proposal	Gaussian mixture with 4 components
Draws	$N_t = 500$
ESS$_{\min}$	100
$\boldsymbol{\tau}$	0.4
Stop	$t = 40$ or $\sum_t \mathrm{ESS}_t > 600$

emission lines), the TAMIS is run as described in sections 5.2 and 5.7 , resulting in the predictive MAP and complete posterior predictive distribution.
Finally we compare the performances of our fitting procedure using either photometric datapoints alone, spectroscopy alone, or both combined. The three resulting distributions are plotted in figures 5.3 and 5.2. As expected some parameters are well constrained using on or the either types, but combining both types yields better estimates for all parameters.

Table 5.3. – Prior range of the parameters for the fits

Parameter	Value
Delayed SFH	
age_{main} (Myr)	$[1000;10000]$
τ_{main} (Myr)	$[1500;3000]$
τ_{burst} (Myr)	$[100;10000]$
f_{burst}	$[0;0.2]$
age_{burst} (Myr)	$[10;100]$
'bc03'	
IMF	$Chabrier$
$Metallicity$	$0.0004, 0.008, 0.05, 0.02, 0.004$
'Nebular'	
$logU$	$-4.0, -3.5, -3.0, -2.5, -2.0, -1.5, -1.0$
$zgas$	$0.004, 0.008, 0.011, 0.022, 0.007, 0.014$
f_{esc}	0
f_{dust}	0
Dust attenuation	
E_BV_lines	$[0;2]$
rv	1
'dl2014'	
$qpah$	0.47,1.12,1.77,2.5
$umin$	5.0,10.0,25.0
α	2
γ	0.02

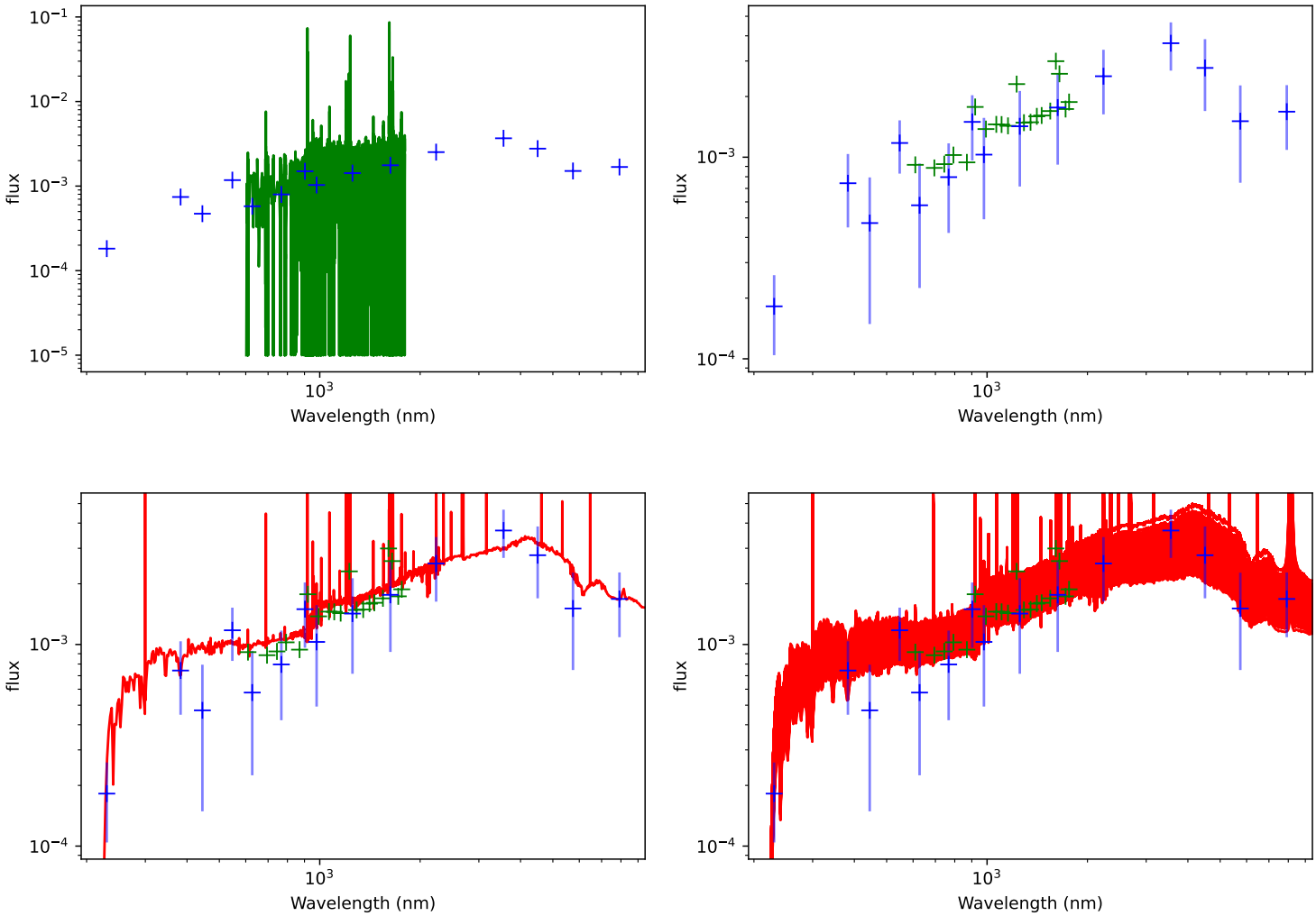

Figure 5.1. – Our proposed SED fitting pipeline. Top Left : The observed spectroscopy (green) and photometry (blue). Top Right : The spectroscopy is binned in 20 values (including the emission lines).The errors bars are represented to account for the noise (vertical lines). This is the data used to compute the likelihood. Bottom Left : the SED corresponding to the MAP is computed (red). Bottom Right : SEDs are sampled from the posterior predictive distribution to visualize the prediction uncertainties and assess proper coverage.

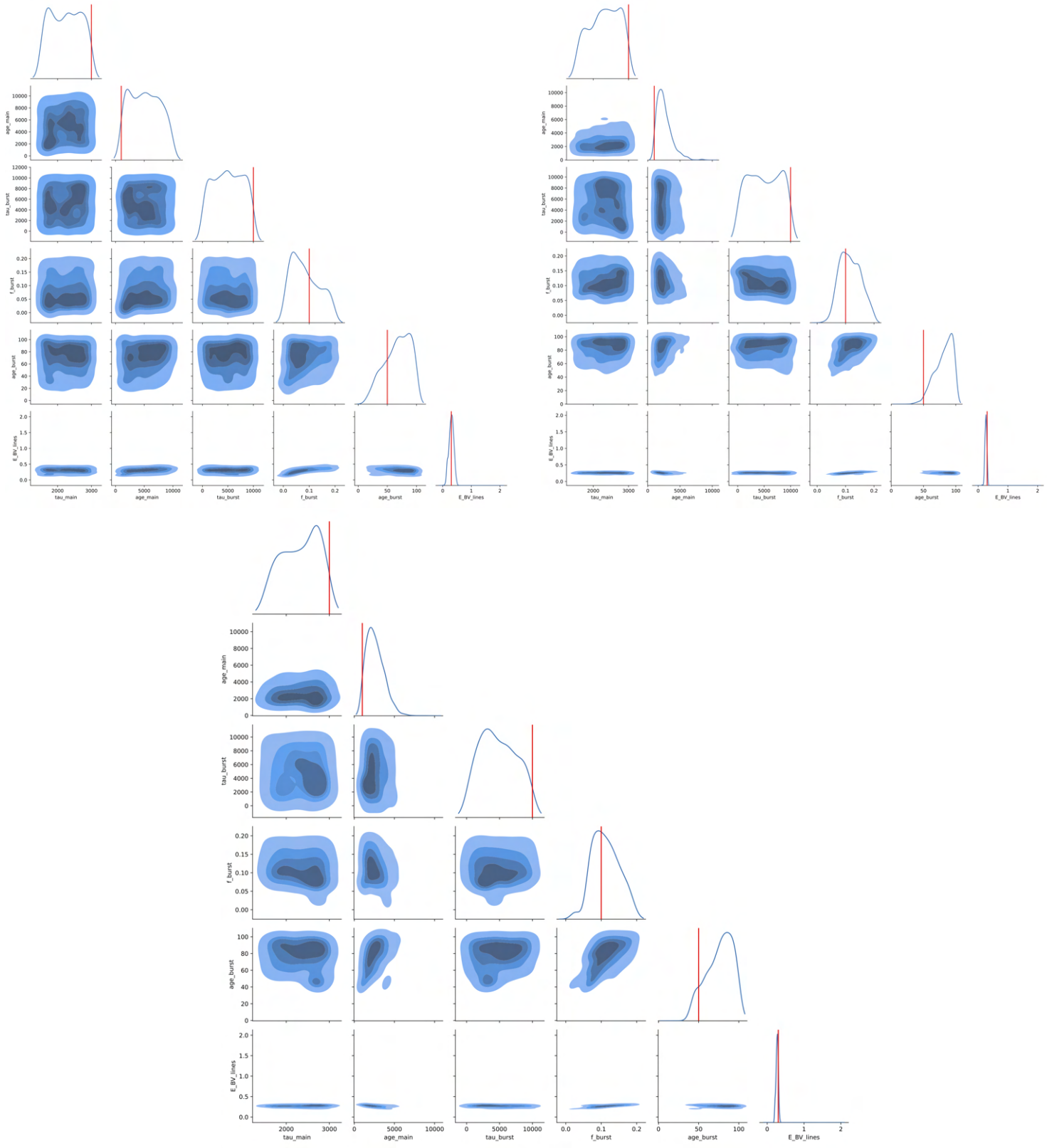

Figure 5.2. – Comparison of the estimated posterior distributions over the continuous parameter space using each datatype. Top Left : using only photometry. Top Right : Using only spectroscopy. Bottom : Using both. The red line represents the true simulating value.

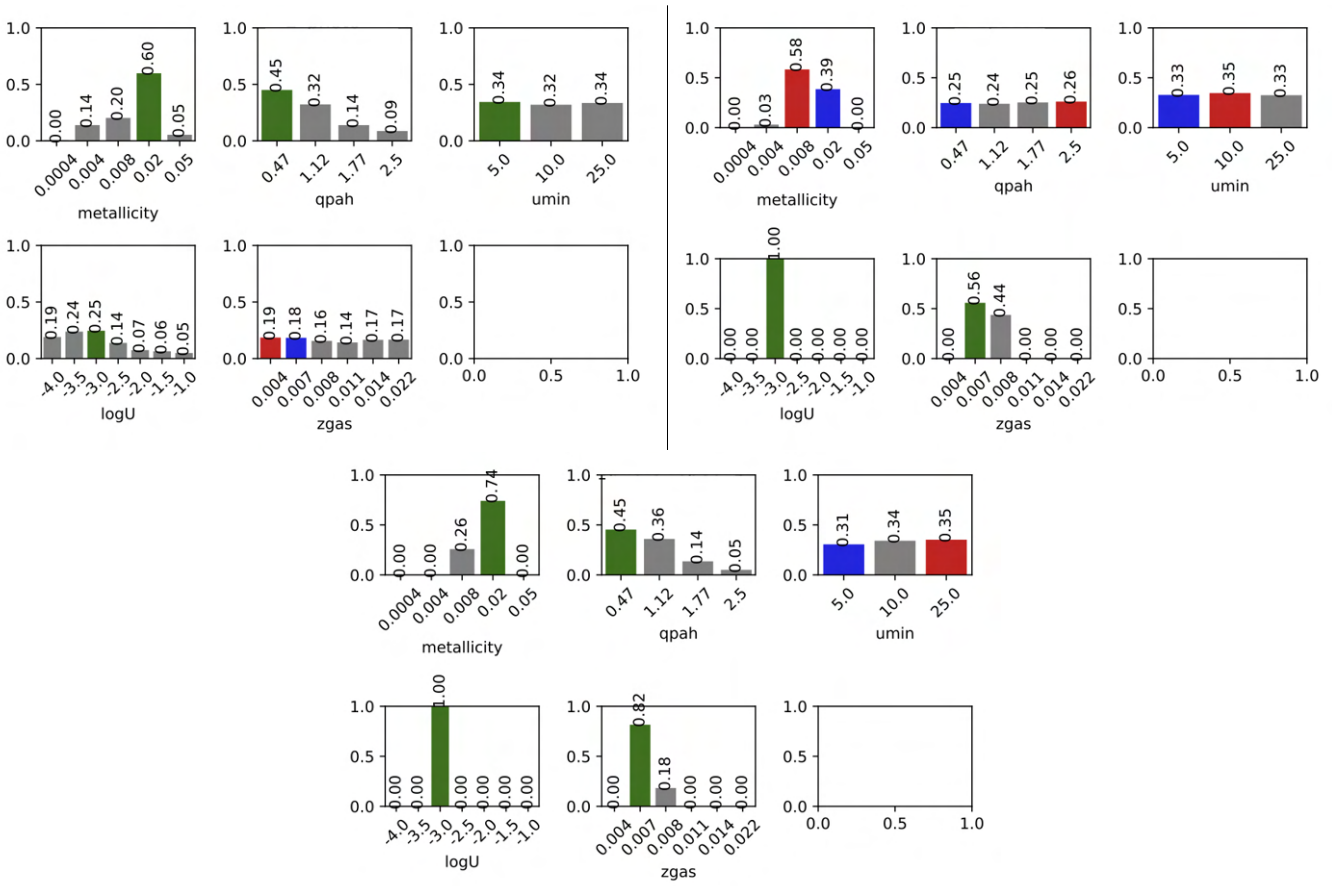

Figure 5.3. – Comparison of the estimated discrete distributions over the continuous parameter space using each datatype. The bars are colored in green if the MAP estimate is the true simulating value. If the MAP is not the simulating value, it is colored in red and the true value in blue. Top Left : using only photometry. Top Right : Using only spectroscopy. Bottom : Using both. Some parameters are well estimated using only one type of data or the other, but combining spectroscopy and photometry successfully exploits the strong suits of both.

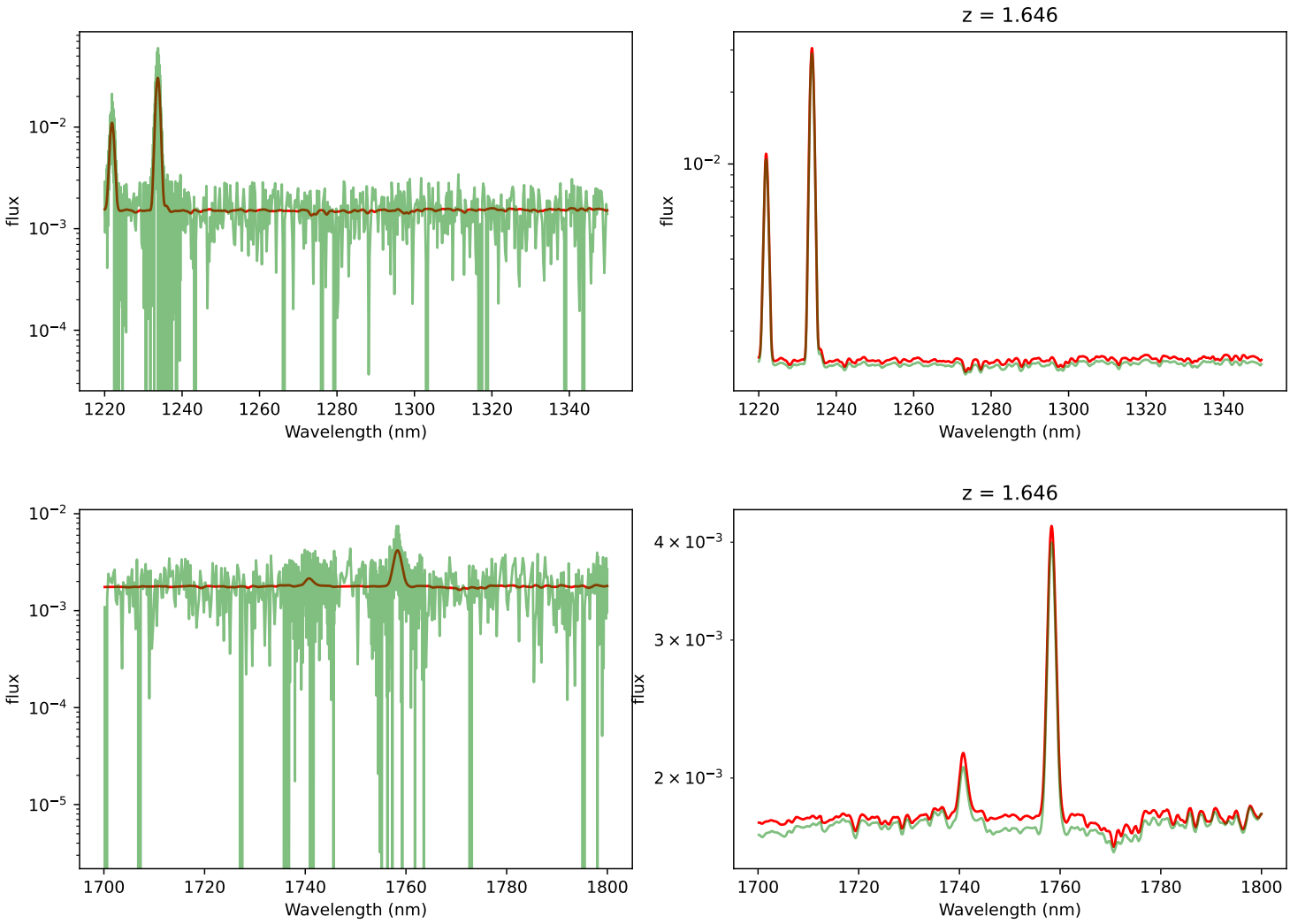

Figure 5.4. – Zoom on the high resolution spectra reconstruction. On the left the observed noisy spectrum (green) with the MAP estimate (red). On the right the original spectrum (before adding noise, green) and the MAP estimate (red). The excellent reconstruction of detailed features despite the noise level is likely a bias due to a lack of model complexity and both original and reconstructed spectrum being generated by the same model.

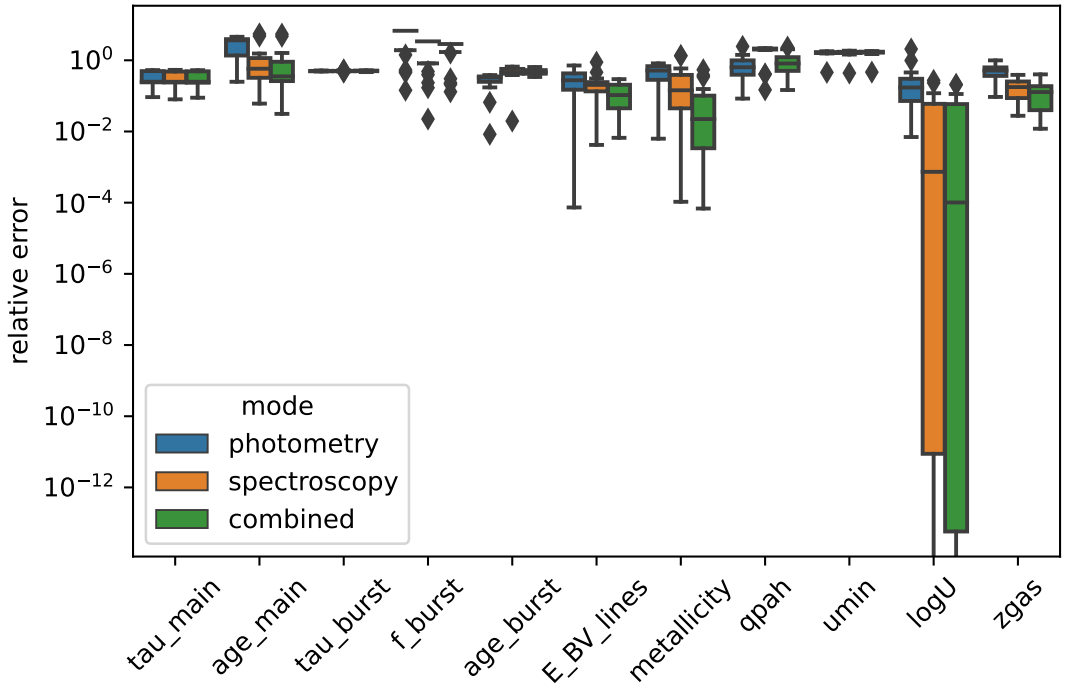

Figure 5.5. – Comparison of the Mean errors between the simulating value and the posterior mean estimate obtained using the 3 fitting methods (photometry, spectroscopy, or both) for each parameter of interest. As expected, spectroscopy is able to better constrain the nebular parameters (metallicity, logU, zgas), and combining both spectroscopy and photometry almost always yields lower error.

6. Conclusion

Sommaire

6.1. Thesis summary

This thesis develops new statistical tools to perform Bayesian inference in Spectral Energy Distribution analysis for different use-cases. The first chapter presented an Approximate Bayesian Computation scheme based on state-of-the art machine learning algorithms to perform Bayesian Star Formation History model choice on a large number of similar datasets. The goal was to approximate the posterior probability of each model given a photometric observation for thousands of galaxies efficiently. The computation time constraint excluded the use of Monte Carlo integration for each galaxy separately. We proposed a way to pool all the required simulations under the prior distribution in a single database and to train a crossentropy minimizing classifier on this training set. The resulting classifier yields a well calibrated approximation of the posterior probabilities of each model, with a computational cost being practically independent of the number of galaxies to study. A practical application on a sample of galaxies from the COSMOS survey shows strong evidence for the need of a more flexible SFH model with short-term fluctuation of the star formation rate, especially for galaxies with lower stellar mass.

The second chapter introduced a new Adaptive Importance Sampling algorithm. At each step, two simple modifications of the importance weights a new family of auxiliary targets. A simple criterion based on effective sample size leads to the automatic-tuning of the sequence of targets allowing for numerical stability regardless of the current proposal. This property is of crucial importance when dealing with Bayesian inference in presence of a vague prior, where the initial proposal distribution cannot be hand-tuned to the target distribution. This is especially the case when analyzing a large variety of datasets for which we seek to develop an automatic fitting procedure. Despite not relying on any gradient computation, the algorithm scales very well with dimension, with numerical examples considered up to the dimension 1000. Very few IS based algorithms are able to sample spaces of such high dimension. It is also extremely efficient in terms of the number of likelihood evaluations required at each

iteration, making it extremely competitive.

The third chapter presented the details of CIGALE physical modeling and a modular neural network approximation to reduce the computation time and extend the range of customizable values. CIGALE is based on a modular downstream pipeline. Each physical process contributing to the light spectrum of a galaxy is modeled using customizable modules. This allows users to choose between different models for each process, or to develop their own and have them interact with the other physical processes. The pipeline starts by computing the stellar emissions of the galaxy using a Star Formation History and a Single Stellar Population library. Nebular emissions are then computed by combining the ionizing photons of the stellar emissions and pre-computed values for the emission lines and continuum of the different elements. The effect of dust is then handled by two modules implementing the attenuation and re-emission of light following an energy-balance principle. Finally the redshift and IGM are taken into account.
We introduced two neural networks following the methodology of Alsing, Peiris, Leja, et al., 2020 to reduce the computation time due to the stellar emission model, and extend the parameter space of the Nebular module to be able to sample continuously instead of relying on a pre-computed grid. We show a significant increase of the computation speed of the stellar emission and satisfying accuracy on the interpolation of the Nebular emissions.

The final chapter presented a complete and flexible framework for Bayesian SED fitting. We proposed a statistical model accounting for the different types of data available, and a data preprocessing pipeline. Using this statistical model, we combined CIGALE simulations and the TAMIS algorithm to provide a principled approach to spectro-photometric Bayesian SED fitting. This new tool enjoys the modularity and customization of CIGALE with state-of-the-art Monte Carlo integration. It supports both continuous and discrete parameter spaces (or a mix), and the use of both spectroscopic and photometric measurements. Its use of ease includes flexible hyperparameters for the statistical model (including flexible prior specification contrary to the previous grid sampling), simple stopping criterion, automatic-tuning of the Monte Carlo scheme and efficient sampling of the parameter space.
We show on simulated spectra that combining spectroscopy with photometry benefits the parameter estimation, especially regarding metallicity.

6.2. Perspective and Future work

Every contribution of this thesis work can be extended or applied in new ways. This section proposes a few such avenues.

Bridging the gap between ABC and SOM for SED fitting Self-Organizing Maps (Kohonen, 1990) are a popular tool in the Astrophysics community (Davidzon, Laigle,

P. L. Capak, et al., 2019,Masters, Peter Capak, D. Stern, et al., 2015 ; Hemmati, Peter Capak, Pourrahmani, et al., 2019 ; Geach, 2012). However once the SOM is learned, several schemes have been proposed to obtain a prediction and potentially derive a probability distribution to quantify the prediction uncertainty. As the interpretation of those prediction uncertainty depends on the way they are derived, an idea would be to reframe this SOM methodology in the ABC setting. This would allow us to obtain a proper approximation of the posterior distribution, derive a Bayesian analysis pipeline and formalize the underlying prior elicitations.

TAMIS and the normalizing constant The computation of the posterior normalizing constant (i.e the evidence, or marginal likelihood) is an important problem in a number of tasks. In particular there is significant statistical literature on the link between the estimation of the normalizing constant in a Bayesian setting and its role in generative models (Kingma and Welling, 2019 ; Thin, Janati El Idrissi, Le Corff, et al., 2021 ;Geach, 2012). As we showed TAMIS to be both cost effective and scalable to high-dimensional settings, an interesting avenue would be to study TAMIS' normalizing constant estimate and bias, and its possible application within probabilistic deep neural networks.

On the biases introduced by the neural network sampling. It is well known that the construction of the training sample of a Neural Network induces biases (S. Geman, Bienenstock, and Doursat, 1992, B. Kim, H. Kim, K. Kim, et al., 2018) in its predictions. As we rely on such a network for the physical simulation used in the likelihood computation, we expect those biases to influence the shape of the likelihood, and thus the resulting posterior approximation. A quantification of this influence is necessary to assess the pertinence of the method proposed in chapter 4. Assuming this influence is not negligible, deriving a correction of this bias directly on the posterior approximation would significantly improve the method.

A completely NN likelihood. The approach proposed in 4 has the advantage of keeping CIGALE's flexibility, modularity and interpretability, but introduces a lot of computational overhead. A more efficient method would be to replace the entire CIGALE pipeline with a single neural network (as proposed in Alsing, Peiris, Leja, et al., 2020) and use it to compute the likelihood. Another possibility would be to directly train a Neural density estimator (Papamakarios, Pavlakou, and Murray, 2017 ; Papamakarios, 2019) to approximate the posterior directly. This solution would be less flexible in terms of the input, would accentuate the Black-box aspect of the process and require retraining the neural network for each modification of a single module in the CIGALE pipeline, but would be magnitudes faster and convenient enough when the problem is to fit millions of observations from the same survey using the same modeling assumptions. TAMIS would still be useful to explore the parameter space with few evaluations and the study of a big survey could be done on a personal computer in a reasonable amount of time.

Transferring the importance samples. The main disadvantage of our TAMIS based SED fitting approach when compared to the current grid-based Importance Sampling is the dependency between the number of spectra simulations required and the number of galaxies to study. When working with the current version of CIGALE, the number of simulations only depends on the chosen parameter grid. Once every simulation is computed and stored, they can be compared with an arbitrary number of observed SEDs. That is to say, from a statistical point of view, we perform importance sampling from a common proposal that is the prior and recompute the weight for each galaxy to analyze. Fitting 1 or 10 observations takes practically the same time, as the likelihood computation time is negligible once the simulation is done. This is not the case with TAMIS, as the parameter sampling directly depends on the observation. This leads to a way more efficient sampling (thousands of simulations instead of hundreds of millions) for a single observation, but scales linearly with the number of observations to fit. Since the current extragalactic surveys include millions of spectra, this linear scaling is problematic.
Nevertheless we could explore the idea to transfer the importance samples from one fit to the next: If two spectra are close in the observation space, their posterior distributions might be close in the parameter space. Extending this idea, denoting x_1 and x_2 two observations, $\theta_1 \sim p(\theta|x_1)$ and $\theta_2 \sim p(\theta|x_2)$ random variables from each posterior distribution, we could learn a transformation $f_{x_1,x_2}(\theta_1)$ such that $p(f_{x_1,x_2}(\theta_1)|x_2)) \approx p(\theta_2|x_2)$. This idea was explored in Paananen, Piironen, Bürkner, et al., 2019 for the problem of leave-one-out cross-validation. This transformation would enable a very good initialization of TAMIS requiring less adaptation steps.

Application to new surveys. As the TAMIS-CIGALE combination allows to fit complex emission models using both spectroscopy and photometry, we will be able to leverage the data from upcoming observation systems. Such systems include ESO's Multi-Object Optical and Near-IR Spectrograph (MOONS) to be installed in the Very Large Telescope in Chile, as we already collaborate with the MOONS Science Team to prepare for the data analysis aspect of the program. It is also of interest for the data analysis of the Subaru Prime Focus Spectrograph (PFS) and the EUCLID space telescope in which the Laboratoire d'Astrophysique de Marseille is involved. Developing specific analysis pipelines and models for each program, or combinations of program, is of crucial importance to successfully learn as much as possible from those future surveys.

Bibliography

[Abr+16] L. E. Abramson, M. D. Gladders, A. Dressler, et al. “Return to [Log-]Normalcy: Rethinking Quenching, The Star Formation Main Sequence, and Perhaps Much More”. In: *ApJ* 832, 7 (Nov. 2016), p. 7. DOI: 10.3847/0004-637X/832/1/7. arXiv: 1604.00016.

[Aka73] Hirotugu Akaike. “Information Theory and an Extension of the Maximum Likelihood Principle”. In: 1973.

[Als+20] Justin Alsing, Hiranya Peiris, Joel Leja, et al. “SPECULATOR: Emulating Stellar Population Synthesis for Fast and Accurate Galaxy Spectra and Photometry”. In: *The Astrophysical Journal Supplement Series* 249.1 (June 2020), p. 5. DOI: 10.3847/1538-4365/ab917f. URL: https://doi.org/10.3847/1538-4365/ab917f.

[Arn+13] S. Arnouts, E. Le Floc’h, J. Chevallard, et al. “Encoding of the infrared excess in the NUVrK color diagram for star-forming galaxies”. In: *A&A* 558, A67 (Oct. 2013), A67. DOI: 10.1051/0004-6361/201321768. arXiv: 1309.0008.

[BC13] Frederik Beaujean and Allen Caldwell. *Initializing adaptive importance sampling with Markov chains*. 2013. arXiv: 1304.7808 [stat.CO].

[BWC13] P. S. Behroozi, R. H. Wechsler, and C. Conroy. “The Average Star Formation Histories of Galaxies in Dark Matter Halos from z = 0-8”. In: *ApJ* 770, 57 (June 2013), p. 57. DOI: 10.1088/0004-637X/770/1/57. arXiv: 1207.6105 [astro-ph.CO].

[Ben12] Yoshua Bengio. “Practical recommendations for gradient-based training of deep architectures”. In: *arXiv e-prints*, arXiv:1206.5533 (June 2012), arXiv:1206.5533. arXiv: 1206.5533 [cs.LG].

[BB12] James Bergstra and Yoshua Bengio. “Random Search for Hyper-Parameter Optimization”. In: *J. Mach. Learn. Res.* 13.null (Feb. 2012), pp. 281–305. ISSN: 1532-4435.

[Bes+16] Alexandros Beskos, Ajay Jasra, Nikolas Kantas, et al. “On the convergence of adaptive sequential Monte Carlo methods”. In: *The Annals of Applied Probability* 26.2 (2016), pp. 1111–1146. DOI: 10.1214/15-AAP1113. URL: https://doi.org/10.1214/15-AAP1113.

[BBP14] M. Boquien, V. Buat, and V. Perret. “Impact of star formation history on the measurement of star formation rates”. In: *A&A* 571, A72 (Nov. 2014), A72. DOI: 10.1051/0004-6361/201424441. arXiv: 1409.5792.

[Boq+19] M. Boquien, D. Burgarella, Y. Roehlly, et al. "CIGALE: a python Code Investigating GALaxy Emission". In: *A&A* 622, A103 (Feb. 2019), A103. DOI: 10.1051/0004-6361/201834156. arXiv: 1811.03094 [astro-ph.GA].

[Bos+16] A. Boselli, Y. Roehlly, M. Fossati, et al. "Quenching of the star formation activity in cluster galaxies". In: *A&A* 596, A11 (Nov. 2016), A11. DOI: 10.1051/0004-6361/201629221. arXiv: 1609.00545 [astro-ph.GA].

[Bou+11] R. J. Bouwens, G. D. Illingworth, I. Labbe, et al. "A candidate redshift z~10 galaxy and rapid changes in that population at an age of 500Myr". In: *Nature* 469.7331 (Jan. 2011), pp. 504–507. DOI: 10.1038/nature09717. arXiv: 0912.4263 [astro-ph.CO].

[Bow+20] William P. Bowman, Gregory R. Zeimann, Gautam Nagaraj, et al. "MCSED: A Flexible Spectral Energy Distribution Fitting Code and Its Application to z ~ 2 Emission-line Galaxies". In: *ApJ* 899.1, 7 (Aug. 2020), p. 7. DOI: 10.3847/1538-4357/ab9f3c. arXiv: 2006.13245 [astro-ph.GA].

[Bre01] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[BC03] G. Bruzual and S. Charlot. "Stellar population synthesis at the resolution of 2003". In: *MNRAS* 344 (Oct. 2003), pp. 1000–1028. DOI: 10.1046/j.1365-8711.2003.06897.x. eprint: arXiv:astro-ph/0309134.

[Bua+18] V. Buat, M. Boquien, K. Małek, et al. "Dust attenuation and Hα emission in a sample of galaxies observed with Herschel at 0.6 < z < 1.6". In: *A&A* 619, A135 (Nov. 2018), A135. DOI: 10.1051/0004-6361/201833841. arXiv: 1809.00161 [astro-ph.GA].

[Bua+14] V. Buat, S. Heinis, M. Boquien, et al. "Ultraviolet to infrared emission of z 1 galaxies: Can we derive reliable star formation rates and stellar masses?" In: *A&A* 561, A39 (Jan. 2014), A39. DOI: 10.1051/0004-6361/201322081. arXiv: 1310.7712 [astro-ph.CO].

[Bug+17] M. F. Bugallo, V. Elvira, L. Martino, et al. "Adaptive Importance Sampling: The past, the present, and the future". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 60–79. DOI: 10.1109/MSP.2017.2699226.

[Cal+00] D. Calzetti, L. Armus, R. C. Bohlin, et al. "The Dust Content and Opacity of Actively Star-forming Galaxies". In: *ApJ* 533 (Apr. 2000), pp. 682–695. DOI: 10.1086/308692. eprint: arXiv:astro-ph/9911459.

[Cap+08] Olivier Cappé, Randal Douc, Arnaud Guillin, et al. "Adaptive importance sampling in general mixture classes". In: *Statistics and Computing* 18.4 (Apr. 2008), pp. 447–459. ISSN: 1573-1375. DOI: 10.1007/s11222-008-9059-x. URL: http://dx.doi.org/10.1007/s11222-008-9059-x.

[Car+19] Adam C. Carnall, Joel Leja, Benjamin D. Johnson, et al. "How to Measure Galaxy Star Formation Histories. I. Parametric Models". In: *ApJ* 873.1, 44 (Mar. 2019), p. 44. DOI: 10.3847/1538-4357/ab04a2. arXiv: 1811.03635 [astro-ph.GA].

[Car+17] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, et al. "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software* 76.1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v076i01.

[CG92] George Casella and Edward I. George. "Explaining the Gibbs Sampler". In: *The American Statistician* 46.3 (1992), pp. 167–174. ISSN: 00031305. URL: http://www.jstor.org/stable/2685208 (visited on 07/08/2022).

[Cas12] C. M. Casey. "Far-infrared spectral energy distribution fitting for galaxies near and far". In: *MNRAS* 425 (Oct. 2012), pp. 3094–3103. DOI: 10.1111/j.1365-2966.2012.21455.x. arXiv: 1206.1595 [astro-ph.CO].

[Cas+15] Caitlin M. Casey et al. "Next Generation Very Large Array Memo No. 8 Science Working Group 3: Galaxy Assembly through Cosmic Time". In: (2015). arXiv: 1510.06411 [astro-ph.GA].

[CF00] S. Charlot and S. M. Fall. "A Simple Model for the Absorption of Starlight by Dust in Galaxies". In: *ApJ* 539 (Aug. 2000), pp. 718–731. DOI: 10.1086/309250. eprint: astro-ph/0003128.

[CG16] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[Cie+16] L. Ciesla, A. Boselli, D. Elbaz, et al. "The imprint of rapid star formation quenching on the spectral energy distributions of galaxies". In: *A&A* 585, A43 (Jan. 2016), A43. DOI: 10.1051/0004-6361/201527107. arXiv: 1510.07657.

[Cie+15] L. Ciesla, V. Charmandaris, A. Georgakakis, et al. "Constraining the properties of AGN host galaxies with spectral energy distribution modelling". In: *A&A* 576, A10 (Apr. 2015), A10. DOI: 10.1051/0004-6361/201425252. arXiv: 1501.03672.

[CEF17a] L. Ciesla, D. Elbaz, and J. Fensch. "The SFR-M$_?$ main sequence archetypal star-formation history and analytical models". In: *A&A* 608, A41 (Dec. 2017), A41. DOI: 10.1051/0004-6361/201731036. arXiv: 1706.08531.

[CEF17b] L. Ciesla, D. Elbaz, and J. Fensch. "The SFR-M$_*$ main sequence archetypal star-formation history and analytical models". In: *A&A* 608, A41 (Dec. 2017), A41. DOI: 10.1051/0004-6361/201731036. arXiv: 1706.08531.

[Cie+18] L. Ciesla, D. Elbaz, C. Schreiber, et al. "Identification of galaxies that experienced a recent major drop of star formation". In: *A&A* 615, A61 (July 2018), A61. DOI: 10.1051/0004-6361/201832715. arXiv: 1803.10239 [astro-ph.GA].

[CT20] Tine Colman and R. Teyssier. "On the origin of the peak of the stellar initial mass function: exploring the tidal screening theory". In: *Monthly Notices of the Royal Astronomical Society* (2020).

[Cor+12] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, et al. "Adaptive Multiple Importance Sampling". In: *Scandinavian Journal of Statistics* 39.4 (2012), pp. 798–812. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/23357226.

[DH02] D. A. Dale and G. Helou. "The Infrared Spectral Energy Distribution of Normal Star-forming Galaxies: Calibration at Far-Infrared and Submillimeter Wavelengths". In: *ApJ* 576 (Sept. 2002), pp. 159–168. DOI: 10.1086/341632. eprint: arXiv:astro-ph/0205085.

[Dal+14] D. A. Dale, G. Helou, G. E. Magdis, et al. "A Two-parameter Model for the Infrared/Submillimeter/Radio Spectral Energy Distributions of Galaxies and Active Galactic Nuclei". In: *ApJ* 784, 83 (Mar. 2014), p. 83. DOI: 10.1088/0004-637X/784/1/83. arXiv: 1402.1495 [astro-ph.CO].

[Dav+19] I Davidzon, C Laigle, P L Capak, et al. "horizon-AGN virtual observatory - 2. Template-free estimates of galaxy properties from colours". In: *Monthly Notices of the Royal Astronomical Society* 489.4 (Sept. 2019), pp. 4817–4835. ISSN: 0035-8711. DOI: 10.1093/mnras/stz2486. eprint: https://academic.oup.com/mnras/article-pdf/489/4/4817/30046320/stz2486.pdf. URL: https://doi.org/10.1093/mnras/stz2486.

[DB14] A. Dekel and A. Burkert. "Wet disc contraction to galactic blue nuggets and quenching to red nuggets". In: *MNRAS* 438 (Feb. 2014), pp. 1870–1879. DOI: 10.1093/mnras/stt2331. arXiv: 1310.1074.

[DP21] Bernard Delyon and François Portier. "Safe adaptive importance sampling: A mixture approach". In: *The Annals of Statistics* 49.2 (2021), pp. 885–917. DOI: 10.1214/20-AOS1983. URL: https://doi.org/10.1214/20-AOS1983.

[Dou+01] A. Doucet, A. Smith, N. de Freitas, et al. *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York, 2001. ISBN: 9780387951461. URL: https://books.google.fr/books?id=uxX-koqKtMMC.

[Dra+14] B. T. Draine, G. Aniano, Oliver Krause, et al. "Andromeda's Dust". In: *ApJ* 780.2, 172 (Jan. 2014), p. 172. DOI: 10.1088/0004-637X/780/2/172. arXiv: 1306.2304 [astro-ph.CO].

[Dra+07] B. T. Draine, D. A. Dale, G. Bendo, et al. "Dust Masses, PAH Abundances, and Starlight Intensities in the SINGS Galaxy Sample". In: *ApJ* 663 (July 2007), pp. 866–894. DOI: 10.1086/518306. eprint: arXiv:astro-ph/0703213.

[Dua+87] S. Duane, A. D. Kennedy, B. J. Pendleton, et al. "Hybrid Monte Carlo". In: *Phys. Lett. B* 195 (1987), pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

[Elb+07] D. Elbaz, E. Daddi, D. Le Borgne, et al. "The reversal of the star formation-density relation in the distant universe". In: *A&A* 468 (June 2007), pp. 33–48. DOI: 10.1051/0004-6361:20077525. eprint: astro-ph/0703653.

[Eng+05] C. W. Engelbracht, K. D. Gordon, G. H. Rieke, et al. "Metallicity Effects on Mid-Infrared Colors and the 8 μm PAH Emission in Galaxies". In: *ApJ* 628 (July 2005), pp. L29–L32. DOI: 10.1086/432613. eprint: arXiv:astro-ph/0506214.

[Fer80] G. J. Ferland. "Hydrogenic emission and recombination coefficients for a wide range of temperature and wavelength." In: *PASP* 92 (Oct. 1980), pp. 596–602. DOI: 10.1086/130718.

[Fer+98] G. J. Ferland, K. T. Korista, D. A. Verner, et al. "CLOUDY 90: Numerical Simulation of Plasmas and Their Spectra". In: *PASP* 110.749 (July 1998), pp. 761–778. DOI: 10.1086/316190.

[Fer+13] G. J. Ferland, R. L. Porter, P. A. M. van Hoof, et al. "The 2013 Release of Cloudy". In: *Rev. Mexicana Astron. Astrofis.* 49 (Apr. 2013), pp. 137–163. arXiv: 1302.4485 [astro-ph.GA].

[For+13] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, et al. "emcee: The MCMC Hammer". In: *PASP* 125.925 (Mar. 2013), p. 306. DOI: 10.1086/670067. arXiv: 1202.3665 [astro-ph.IM].

[Fra+08] P. Franzetti, M. Scodeggio, B. Garilli, et al. "GOSSIP, a New VO Compliant Tool for SED Fitting". In: *Astronomical Data Analysis Software and Systems XVII*. Ed. by R. W. Argyle, P. S. Bunclark, and J. R. Lewis. Vol. 394. Astronomical Society of the Pacific Conference Series. Aug. 2008, p. 642. arXiv: 0801.2518 [astro-ph].

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[FCR19] Sylvia Fruhwirth-Schnatter, Gilles Celeux, and Christian P Robert. *Handbook of mixture analysis*. CRC press, 2019.

[GPB96] G. Gavazzi, D. Pierini, and A. Boselli. "The phenomenology of disk galaxies." In: *A&A* 312 (Aug. 1996), pp. 397–408.

[Gea12] James E. Geach. "Unsupervised self-organized mapping: a versatile empirical tool for object selection, classification and redshift estimation in large surveys". In: *Monthly Notices of the Royal Astronomical Society* 419.3 (Jan. 2012), pp. 2633–2645. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2011.19913.x. eprint: https://academic.oup.com/mnras/article-pdf/419/3/2633/18719180/mnras0419-2633.pdf. URL: https://doi.org/10.1111/j.1365-2966.2011.19913.x.

[Gel+13] A. Gelman, J.B. Carlin, H.S. Stern, et al. *Bayesian Data Analysis*. Ed. by Chapman and Hall/CRC. 3rd. 2013.

[GBD92] Stuart Geman, Elie Bienenstock, and René Doursat. "Neural Networks and the Bias/Variance Dilemma". In: *Neural Computation* 4.1 (1992), pp. 1–58. DOI: 10.1162/neco.1992.4.1.1.

[GG84] Stuart Geman and Donald Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pp. 721–741. DOI: 10.1109/TPAMI.1984.4767596.

[Gla+13] M. D. Gladders, A. Oemler, A. Dressler, et al. "The IMACS Cluster Building Survey. IV. The Log-normal Star Formation History of Galaxies". In: *ApJ* 770, 64 (June 2013), p. 64. DOI: 10.1088/0004-637X/770/1/64. arXiv: 1303.3917 [astro-ph.CO].

[Gol+17] *Google Vizier: A Service for Black-Box Optimization*. Proceedings of the 23rd International Conference on Knowledge Discovery and Data Mining. 2017, pp. 1487–1495.

[GZF13] K. Guo, X. Z. Zheng, and H. Fu. "The Intrinsic Scatter along the Main Sequence of Star-forming Galaxies at z ~ 0.7". In: *ApJ* 778, 23 (Nov. 2013), p. 23. DOI: 10.1088/0004-637X/778/1/23. arXiv: 1309.4093 [astro-ph.CO].

[Hat+16] N. A. Hatch, S. I. Muldrew, E. A. Cooke, et al. "The structure and evolution of a forming galaxy cluster at z = 1.62". In: *Mon. Not. Roy. Astron. Soc.* 459.1 (2016), pp. 387–401. DOI: 10.1093/mnras/stw602. arXiv: 1603.03774 [astro-ph.GA].

[Hay+11] Matthew Hayes, Daniel Schaerer, Göran Östlin, et al. "ON THE REDSHIFT EVOLUTION OF THE Lyα ESCAPE FRACTION AND THE DUST CONTENT OF GALAXIES". In: *The Astrophysical Journal* 730.1 (Feb. 2011), p. 8. DOI: 10.1088/0004-637x/730/1/8. URL: https://doi.org/10.1088/0004-637x/730/1/8.

[Hea+21] Tim Head, Manoj Kumar, Holger Nahrstaedt, et al. "scikit-optimize/scikit-optimize". In: *Zenodo* (Oct. 2021).

[Hem+19] Shoubaneh Hemmati, Peter Capak, Milad Pourrahmani, et al. "Bringing Manifold Learning and Dimensionality Reduction to SED Fitters". In: *ApJ* 881.1, L14 (Aug. 2019), p. L14. DOI: 10.3847/2041-8213/ab3418. arXiv: 1905.10379 [astro-ph.GA].

[HG14] Matthew D. Homan and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *J. Mach. Learn. Res.* 15.1 (Jan. 2014), pp. 1593–1623. ISSN: 1532-4435.

[Ilb+15] O. Ilbert, S. Arnouts, E. Le Floc'h, et al. "Evolution of the specific star formation rate function at z 1.4 Dissecting the mass-SFR plane in COSMOS and GOODS". In: *A&A* 579, A2 (July 2015), A2. DOI: 10.1051/0004-6361/201425176. arXiv: 1410.4875.

[Ino01] Akio K. Inoue. "Lyman Continuum Extinction by Dust in H [CSC]ii[/CSC] Regions of Galaxies". In: *The Astronomical Journal* 122.4 (Oct. 2001), pp. 1788–1795. DOI: 10.1086/323095. URL: https://doi.org/10.1086/323095.

[Ino11] Akio K. Inoue. "Rest-frame ultraviolet-to-optical spectral characteristics of extremely metal-poor and metal-free galaxies". In: *Monthly Notices of the Royal Astronomical Society* 415.3 (Aug. 2011), pp. 2920–2931. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2011.18906.x. eprint: https://academic.oup.com/mnras/article-pdf/415/3/2920/5982955/mnras0415-2920.pdf. URL: https://doi.org/10.1111/j.1365-2966.2011.18906.x.

[IID06] Akio K. Inoue, Ikuru Iwata, and Jean-Michel Deharveng. "The escape fraction of ionizing photons from galaxies at z = 0-6". In: *MNRAS* 371.1 (Sept. 2006), pp. L1–L5. DOI: 10.1111/j.1745-3933.2006.00195.x. arXiv: astro-ph/0605526 [astro-ph].

[Ion08] Edward L Ionides. "Truncated Importance Sampling". In: *Journal of Computational and Graphical Statistics* 17.2 (2008), pp. 295–311. DOI: 10.1198/106186008X320456. eprint: https://doi.org/10.1198/106186008X320456. URL: https://doi.org/10.1198/106186008X320456.

[Jef98] H. Jeffreys. *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. OUP Oxford, 1998. ISBN: 9780191589676. URL: https://books.google.fr/books?id=vh9Act9rtzQC.

[Kas+13] D. Kashino, J. D. Silverman, G. Rodighiero, et al. "The FMOS-COSMOS Survey of Star-forming Galaxies at z ∼1.6. I. Hα-based Star Formation Rates and Dust Extinction". In: *ApJ* 777.1, L8 (Nov. 2013), p. L8. DOI: 10.1088/2041-8205/777/1/L8. arXiv: 1309.4774 [astro-ph.CO].

[KR95] Robert E. Kass and Adrian E. Raftery. "Bayes Factors". In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. ISSN: 01621459. URL: http://www.jstor.org/stable/2291091 (visited on 07/14/2022).

[Kau+03a] G. Kauffmann, T. M. Heckman, S. D. M. White, et al. "Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey". In: *MNRAS* 341 (May 2003), pp. 33–53. DOI: 10.1046/j.1365-8711.2003.06291.x. eprint: astro-ph/0204055.

[Kau+03b] Guinevere Kauffmann, Timothy M. Heckman, Simon D. M. White, et al. "Stellar masses and star formation histories for 105 galaxies from the Sloan Digital Sky Survey". In: *Monthly Notices of the Royal Astronomical Society* 341.1 (May 2003), pp. 33–53. ISSN: 0035-8711. DOI: 10.1046/j.1365-8711.2003.06291.x. eprint: https://academic.oup.com/mnras/article-pdf/341/1/33/18650863/341-1-33.pdf. URL: https://doi.org/10.1046/j.1365-8711.2003.06291.x.

[Kim+18] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, et al. “Learning Not to Learn: Training Deep Neural Networks with Biased Data”. In: *arXiv e-prints*, arXiv:1812.10352 (Dec. 2018), arXiv:1812.10352. arXiv: 1812.10352 [cs.CV].

[KW19] Diederik P. Kingma and Max Welling. “An Introduction to Variational Autoencoders”. In: *arXiv e-prints*, arXiv:1906.02691 (June 2019), arXiv:1906.02691. arXiv: 1906.02691 [cs.LG].

[KM15] Eugenia Koblents and Joaquin Miguez. “A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models”. In: *Statistics and Computing* 25.2 (2015), pp. 407–425.

[Koh90] T. Kohonen. “The self-organizing map”. In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480. DOI: 10.1109/5.58325.

[KP22] Anna Korba and François Portier. “Adaptive Importance Sampling meets Mirror Descent : a Bias-variance Tradeoff”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 28–30 Mar 2022, pp. 11503–11527. URL: https://proceedings.mlr.press/v151/korba22a.html.

[Kro01] P. Kroupa. “On the variation of the initial mass function”. In: *MNRAS* 322 (Apr. 2001), pp. 231–246. DOI: 10.1046/j.1365-8711.2001.04022.x. eprint: arXiv:astro-ph/0009005.

[EEB18] Yousef El-Laham, Victor Elvira, and Monica Bugallo. “Robust Covariance Adaptation in Adaptive Importance Sampling”. In: *IEEE Signal Processing Letters* 25.7 (July 2018), pp. 1049–1053. URL: https://hal.archives-ouvertes.fr/hal-02019041.

[Lai+16a] C. Laigle, H. J. McCracken, O. Ilbert, B. C. Hsieh, I. Davidzon, P. Capak, G. Hasinger, J. D. Silverman, C. Pichon, J. Coupon, H. Aussel, D. Le Borgne, K. Caputi, P. Cassata, Y. -Y. Chang, F. Civano, J. Dunlop, J. Fynbo, J. S. Kartaltepe, A. Koekemoer, O. Le Fèvre, E. Le Floc’h, A. Leauthaud, S. Lilly, L. Lin, S. Marchesi, B. Milvang- Jensen, et al. “The COSMOS2015 Catalog: Exploring the 1 < z < 6 Universe with Half a Million Galaxies”. In: *The Astrophysical Journal Supplement Series* 224, 24 (June 2016), p. 24. DOI: 10.3847/0067-0049/224/2/24.

[Lai+16b] C. Laigle, H. J. McCracken, O. Ilbert, B. C. Hsieh, I. Davidzon, P. Capak, G. Hasinger, J. D. Silverman, C. Pichon, J. Coupon, H. Aussel, D. Le Borgne, K. Caputi, P. Cassata, Y. -Y. Chang, F. Civano, J. Dunlop, J. Fynbo, J. S. Kartaltepe, A. Koekemoer, O. Le Fèvre, E. Le Floc’h, A. Leauthaud, S. Lilly, L. Lin, S. Marchesi, B. Milvang-Jensen, et al. “The COSMOS2015 Catalog: Exploring the 1 < z < 6 Universe with Half a Million Galaxies”. In: *ApJS* 224.2, 24 (June 2016), p. 24. DOI: 10.3847/0067-0049/224/2/24. arXiv: 1604.02350 [astro-ph.GA].

[Lam+01] Hak-Keung Lam, S. H. Ling, F.H.F. Leung, et al. "Tuning of the structure and parameters of neural network using an improved genetic algorithm". In: vol. 14. Feb. 2001, 25–30 vol.1. ISBN: 0-7803-7108-9. DOI: 10.1109/IECON.2001.976448.

[Lee+10] S.-K. Lee, H. C. Ferguson, R. S. Somerville, et al. "The Estimation of Star Formation Rates and Stellar Population Ages of High-redshift Galaxies from Broadband Photometry". In: *ApJ* 725 (Dec. 2010), pp. 1644–1651. DOI: 10.1088/0004-637X/725/2/1644. arXiv: 1010.1966 [astro-ph.CO].

[LCM02] Claus Leitherer, Daniela Calzetti, and Lucimara P. Martins. "Ultraviolet Spectra of Star-forming Galaxies with Time-dependent Dust Obscuration". In: *The Astrophysical Journal* 574.1 (July 2002), pp. 114–125. DOI: 10.1086/340902. URL: https://doi.org/10.1086/340902.

[Lej+19] Joel Leja, Adam C. Carnall, Benjamin D. Johnson, et al. "How to Measure Galaxy Star Formation Histories. II. Nonparametric Models". In: *ApJ* 876.1, 3 (May 2019), p. 3. DOI: 10.3847/1538-4357/ab133c. arXiv: 1811.03637 [astro-ph.GA].

[Liu01] Jun S Liu. *Monte Carlo strategies in scientific computing*. Vol. 10. Springer, 2001.

[Lo +17] B. Lo Faro, V. Buat, Y. Roehlly, et al. "Characterizing the UV-to-NIR shape of the dust attenuation curve of IR luminous galaxies up to z 2". In: *Monthly Notices of the Royal Astronomical Society* 472.2 (July 2017), pp. 1372–1391. ISSN: 0035-8711. DOI: 10.1093/mnras/stx1901. eprint: https://academic.oup.com/mnras/article-pdf/472/2/1372/19886576/stx1901.pdf. URL: https://doi.org/10.1093/mnras/stx1901.

[Lor13] Thomas J. Loredo. "Bayesian Astrostatistics: A Backward Look to the Future". In: *Astrostatistical Challenges for the New Astronomy*. Ed. by Joseph M. Hilbe. New York, NY: Springer New York, 2013, pp. 15–40. ISBN: 978-1-4614-3508-2. DOI: 10.1007/978-1-4614-3508-2_2. URL: https://doi.org/10.1007/978-1-4614-3508-2_2.

[Mag+12] G. E. Magdis, E. Daddi, M. Béthermin, et al. "The Evolving Interstellar Medium of Star-forming Galaxies since z = 2 as Probed by Their Infrared Spectral Energy Distributions". In: *ApJ* 760, 6 (Nov. 2012), p. 6. DOI: 10.1088/0004-637X/760/1/6. arXiv: 1210.1035 [astro-ph.CO].

[Mar05a] C. Maraston. "Evolutionary population synthesis: models, analysis of the ingredients and application to high-z galaxies". In: *MNRAS* 362 (Sept. 2005), pp. 799–825. DOI: 10.1111/j.1365-2966.2005.09270.x. eprint: arXiv:astro-ph/0410207.

[Mar+10] C. Maraston, J. Pforr, A. Renzini, et al. “Star formation rates and masses of z ~ 2 galaxies from multicolour photometry”. In: *MNRAS* 407 (Sept. 2010), pp. 830–845. DOI: 10.1111/j.1365-2966.2010.16973.x. arXiv: 1004.4546 [astro-ph.CO].

[Mar05b] Claudia Maraston. “Evolutionary population synthesis: models, analysis of the ingredients and application to high-z galaxies”. In: *MNRAS* 362.3 (Sept. 2005), pp. 799–825. DOI: 10.1111/j.1365-2966.2005.09270.x. arXiv: astro-ph/0410207 [astro-ph].

[Mar+18] Jean-Michel Marin, Pierre Pudlo, Arnaud Estoup, et al. “Likelihood-free Model Choice”. In: *Handbook of Approximate Bayesian Computation*. Ed. by Scott A. Sisson, Yanan Fan, and Mark Beaumont. Chapman and Hall/CRC, 2018. Chap. Chapter 6.

[Mar+12] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, et al. “Approximate Bayesian computational methods”. In: *Statistics and Computing* 22.6 (2012), pp. 1167–1180.

[MPS19] Jean-Michel Marin, Pierre Pudlo, and Mohammed Sedki. “Consistency of adaptive importance sampling and recycling schemes”. In: *Bernoulli* 25.3 (2019), pp. 1977–1998.

[MR09] Jean-Michel Marin and Christian Robert. “Importance sampling methods for Bayesian discrimination between embedded models”. In: (Oct. 2009).

[Mas+15] Daniel Masters, Peter Capak, Daniel Stern, et al. “Mapping the Galaxy Color-Redshift Relation: Optimal Photometric Redshift Calibration Strategies for Cosmology Surveys”. In: *ApJ* 813.1, 53 (Nov. 2015), p. 53. DOI: 10.1088/0004-637X/813/1/53. arXiv: 1509.03318 [astro-ph.CO].

[Mei06] Avery Meiksin. “Colour corrections for high-redshift objects due to intergalactic attenuation”. In: *Monthly Notices of the Royal Astronomical Society* 365.3 (Jan. 2006), pp. 807–812. ISSN: 0035-8711. DOI: 10.1111/j.1365-2966.2005.09756.x. eprint: https://academic.oup.com/mnras/article-pdf/365/3/807/2905061/365-3-807.pdf. URL: https://doi.org/10.1111/j.1365-2966.2005.09756.x.

[MU49] Nicholas Metropolis and S. Ulam. “The Monte Carlo Method”. In: *Journal of the American Statistical Association* 44.247 (1949). PMID: 18139350, pp. 335–341. DOI: 10.1080/01621459.1949.10483310. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483310. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310.

[MDF15] C. Morisset, G. Delgado-Inglada, and N. Flores-Fajardo. “A virtual observatory for photoionized nebulae: the Mexican Million Models database (3MdB).” In: *Rev. Mexicana Astron. Astrofis.* 51 (Apr. 2015), pp. 103–120. arXiv: 1412.5349 [astro-ph.GA].

[Nea11] Radford Neal. “MCMC Using Hamiltonian Dynamics”. In: *Handbook of Markov Chain Monte Carlo*. 2011, pp. 113–162. DOI: 10.1201/b10905.

[Nea01] Radford M. Neal. “Annealed importance sampling”. In: *Statistics and Computing* 11.22 (2001), pp. 125–139. DOI: 10.1023/A:1008923215028. URL: https://doi.org/10.1023/A:1008923215028.

[NC12] Alexandru Niculescu-Mizil and Rich Caruana. “Obtaining Calibrated Probabilities from Boosting”. In: (July 2012). URL: https://www.cs.cornell.edu/~caruana/niculescu.scldbst.crc.rev4.pdf.

[Noe+07] K. G. Noeske, B. J. Weiner, S. M. Faber, et al. “Star Formation in AEGIS Field Galaxies since z=1.1: The Dominance of Gradually Declining Star Formation, and the Main Sequence of Star-forming Galaxies”. In: *ApJ* 660.1 (May 2007), pp. L43–L46. DOI: 10.1086/517926. arXiv: astro-ph/0701924 [astro-ph].

[OB92] Man-Suk Oh and James O Berger. “Adaptive importance sampling in Monte Carlo integration”. In: *Journal of Statistical Computation and Simulation* 41.3-4 (1992), pp. 143–168.

[OB93] Man-Suk Oh and James O Berger. “Integration of multimodal functions by Monte Carlo importance sampling”. In: *Journal of the American Statistical Association* 88.422 (1993), pp. 450–456.

[Orl20] Ivana Orlitova. “Starburst Galaxies”. In: *Reviews in Frontiers of Modern Astrophysics: From Space Debris to Cosmology*. Ed. by Petr Kabáth, David Jones, and Marek Skarka. Cham: Springer International Publishing, 2020, pp. 379–411. ISBN: 978-3-030-38509-5. DOI: 10.1007/978-3-030-38509-5_13. URL: https://doi.org/10.1007/978-3-030-38509-5_13.

[OZ00] Art Owen and Yi Zhou. “Safe and Effective Importance Sampling”. In: *Journal of the American Statistical Association* 95.449 (2000), pp. 135–143.

[Paa+19] Topi Paananen, Juho Piironen, Paul-Christian Bürkner, et al. “Implicitly Adaptive Importance Sampling”. In: *arXiv e-prints*, arXiv:1906.08850 (June 2019), arXiv:1906.08850. arXiv: 1906.08850 [stat.CO].

[Pac+13] C. Pacifici, S. A. Kassin, B. Weiner, et al. “The Rise and Fall of the Star Formation Histories of Blue Galaxies at Redshifts 0.2 z 1.4”. In: *ApJ* 762, L15 (Jan. 2013), p. L15. DOI: 10.1088/2041-8205/762/1/L15. arXiv: 1210.0543.

[Pac+16] C. Pacifici, S. Oh, K. Oh, et al. “Timing the Evolution of Quiescent and Star-forming Local Galaxies”. In: *ApJ* 824, 45 (June 2016), p. 45. DOI: 10.3847/0004-637X/824/1/45. arXiv: 1604.02460.

[Pap19] George Papamakarios. “Neural Density Estimation and Likelihood-free Inference”. In: *arXiv e-prints*, arXiv:1910.13233 (Oct. 2019), arXiv:1910.13233. arXiv: 1910.13233 [stat.ML].

[PPM17] George Papamakarios, Theo Pavlakou, and Iain Murray. "Masked Autoregressive Flow for Density Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf.

[PDF01] C. Papovich, M. Dickinson, and H. C. Ferguson. "The Stellar Populations and Evolution of Lyman Break Galaxies". In: *ApJ* 559 (Oct. 2001), pp. 620–653. DOI: 10.1086/322412. eprint: astro-ph/0105087.

[Ped+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[PMT12] J. Pforr, C. Maraston, and C. Tonini. "Recovering galaxy stellar population properties from broad-band spectral energy distribution fitting". In: *MNRAS* 422 (June 2012), pp. 3285–3326. DOI: 10.1111/j.1365-2966.2012.20848.x. arXiv: 1203.3548 [astro-ph.CO].

[Pud+16] Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, et al. "Reliable ABC model choice via random forests". In: *Bioinformatics* 32.6 (2016), pp. 859–866. DOI: 10.1093/bioinformatics/btv684. URL: http://dx.doi.org/10.1093/bioinformatics/btv684.

[Rob07] Christian Robert. *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation*. Springer Science & Business Media, 2007.

[RCC99] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.

[RC04] Christian P. Robert and George Casella. "Monte Carlo Statistical Methods". In: *Technometrics* 47 (2004), pp. 243–243.

[Roe+14] Y. Roehlly, D. Burgarella, V. Buat, M. Boquien, et al. "Pcigale: Porting Code Investigating Galaxy Emission to Python". In: *Astronomical Data Analysis Software and Systems XXIII*. Ed. by N. Manset and P. Forshay. Vol. 485. Astronomical Society of the Pacific Conference Series. May 2014, p. 347.

[Roe+11] Y. Roehlly, D. Burgarella, V. Buat, É. Giovannoli, et al. "CIGALE: Code Investigating GALaxy Emission". In: *ArXiv e-prints* (Nov. 2011). arXiv: 1111.1117 [astro-ph.CO].

[Sal+12] F. Salmi, E. Daddi, D. Elbaz, et al. "Dissecting the Stellar-mass-SFR Correlation in z = 1 Star-forming Disk Galaxies". In: *ApJ* 754, L14 (July 2012), p. L14. DOI: 10.1088/2041-8205/754/1/L14. arXiv: 1206.1704.

[Sal55] E. E. Salpeter. "The Luminosity Function and Stellar Evolution." In: *ApJ* 121 (Jan. 1955), p. 161. DOI: 10.1086/145971.

[Sar+14] M. T. Sargent, E. Daddi, M. Béthermin, et al. "Regularity Underlying Complexity: A Redshift-independent Description of the Continuous Variation of Galaxy-scale Molecular Gas Properties in the Mass-star Formation Rate Plane". In: *ApJ* 793, 19 (Sept. 2014), p. 19. DOI: 10.1088/0004-637X/793/1/19. arXiv: 1303.4392.

[Sch+21] Rens van de Schoot, Sarah Depaoli, Ruth King, et al. "Bayesian statistics and modelling". In: *Nature Reviews Methods Primers* 1.1 (2021), p. 1. DOI: 10.1038/s43586-020-00001-2. URL: https://doi.org/10.1038/s43586-020-00001-2.

[Sch+15] C. Schreiber, M. Pannella, D. Elbaz, et al. "The Herschel view of the dominant mode of galaxy growth from z = 4 to the present day". In: *A&A* 575, A74 (Mar. 2015), A74. DOI: 10.1051/0004-6361/201425017. arXiv: 1409.5433.

[Sch78] Gideon Schwarz. "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2 (1978), pp. 461–464. DOI: 10.1214/aos/1176344136. URL: https://doi.org/10.1214/aos/1176344136.

[Sco+16] N. Scoville, K. Sheth, H. Aussel, et al. "ISM Masses and the Star formation Law at Z = 1 to 6: ALMA Observations of Dust Continuum in 145 Galaxies in the COSMOS Survey Field". In: *ApJ* 820, 83 (Apr. 2016), p. 83. DOI: 10.3847/0004-637X/820/2/83. arXiv: 1511.05149.

[Sil+15] J. D. Silverman, D. Kashino, D. Sanders, et al. "The FMOS-COSMOS Survey of Star-forming Galaxies at z∼1.6. III. Survey Design, Performance, and Sample Characteristics". In: *ApJS* 220.1, 12 (Sept. 2015), p. 12. DOI: 10.1088/0067-0049/220/1/12. arXiv: 1409.0447 [astro-ph.GA].

[Sim+14] V. Simha, D. H. Weinberg, C. Conroy, et al. "Parametrising Star Formation Histories". In: *ArXiv e-prints* (Apr. 2014). arXiv: 1404.0402.

[SFB18] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. Chapman and Hall/CRC, 2018.

[Ski06] John Skilling. "Nested sampling for general Bayesian computation". In: *Bayesian Analysis* 1.4 (2006), pp. 833–859. DOI: 10.1214/06-BA127. URL: https://doi.org/10.1214/06-BA127.

[Sri+10] Niranjan Srinivas, Andreas Krause, Sham Kakade, et al. "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 1015–1022. ISBN: 9781605589077.

[Ste+16] Charles C. Steidel, Allison L. Strom, Max Pettini, et al. "Reconciling the Stellar and Nebular Spectra of High-redshift Galaxies". In: *ApJ* 826.2, 159 (Aug. 2016), p. 159. DOI: 10.3847/0004-637X/826/2/159. arXiv: 1605.07186 [astro-ph.GA].

[Tac+16] Sandro Tacchella, Avishai Dekel, C. Marcella Carollo, et al. "Evolution of density profiles in high-z galaxies: compaction and quenching inside-out". In: *MNRAS* 458.1 (May 2016), pp. 242–263. DOI: 10.1093/mnras/stw303. arXiv: 1509.00017 [astro-ph.GA].

[Thi+21] Achille Thin, Yazid Janati El Idrissi, Sylvain Le Corff, et al. "NEO: Non Equilibrium Sampling on the Orbits of a Deterministic Transform". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, et al. Vol. 34. Curran Associates, Inc., 2021, pp. 17060–17071. URL: https://proceedings.neurips.cc/paper/2021/file/8dd291cbea8f231982db0fb1716dfc55-Paper.pdf.

[TK86] Luke Tierney and Joseph B. Kadane. "Accurate Approximations for Posterior Moments and Marginal Densities". In: *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86. DOI: 10.1080/01621459.1986.10478240. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1986.10478240. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240.

[VO+12] Aki Vehtari, Janne Ojanen, et al. "A survey of Bayesian predictive methods for model assessment, selection and comparison". In: *Statistics Surveys* 6 (2012), pp. 142–228.

[Veh+21] Aki Vehtari, Daniel Simpson, Andrew Gelman, et al. *Pareto Smoothed Importance Sampling*. 2021. arXiv: 1507.02646 [stat.CO].

[Vil21] J. A Villa-Vélez. "Spectrophotometric analysis around cosmic noon: emission-lines, dust attenuation, and star formation". Theses. Aix-Marseille Université, Nov. 2021. URL: https://tel.archives-ouvertes.fr/tel-03617733.

[Wal+11] Jakob Walcher, Brent Groves, Tamás Budavári, et al. "Fitting the integrated spectral energy distributions of galaxies". In: *Ap&SS* 331 (Jan. 2011), pp. 1–52. DOI: 10.1007/s10509-010-0458-z. arXiv: 1008.0395 [astro-ph.CO].

[Wil+11] Vivienne Wild, Stéphane Charlot, Jarle Brinchmann, et al. "Empirical determination of the shape of dust attenuation curves in star-forming galaxies". In: *MNRAS* 417.3 (Nov. 2011), pp. 1760–1786. DOI: 10.1111/j.1365-2966.2011.19367.x. arXiv: 1106.1646 [astro-ph.CO].

[Wil+09] Rik J. Williams, Ryan F. Quadri, Marijn Franx, et al. "Detection of Quiescent Galaxies in a Bicolor Sequence from Z = 0-2". In: *ApJ* 691.2 (Feb. 2009), pp. 1879–1895. DOI: 10.1088/0004-637X/691/2/1879. arXiv: 0806.0625 [astro-ph].

[Wra+09] Darren Wraith, Martin Kilbinger, Karim Benabed, et al. "Estimation of cosmological parameters using adaptive importance sampling". In: *Phys. Rev. D* 80 (2 July 2009), p. 023507. DOI: 10.1103/PhysRevD.80.023507. URL: https://link.aps.org/doi/10.1103/PhysRevD.80.023507.

[Wri06] E. L. Wright. “A Cosmology Calculator for the World Wide Web”. In: *PASP* 118 (Dec. 2006), pp. 1711–1715. DOI: 10.1086/510102. eprint: astro-ph/0609593.

[Wuy+11] S. Wuyts, N. M. Förster Schreiber, A. van der Wel, et al. “Galaxy Structure and Mode of Star Formation in the SFR-Mass Plane from z ~ 2.5 to z ~ 0.1”. In: *ApJ* 742, 96 (Dec. 2011), p. 96. DOI: 10.1088/0004-637X/742/2/96. arXiv: 1107.0317 [astro-ph.CO].

[Wuy+07] Stijn Wuyts, Ivo Labbé, Marijn Franx, et al. “What Do We Learn from IRAC Observations of Galaxies at 2 < z < 3.5?” In: *ApJ* 655.1 (Jan. 2007), pp. 51–65. DOI: 10.1086/509708. arXiv: astro-ph/0609548 [astro-ph].

APPENDIX

A. Impact of fluxes SNR on the distribution of $p(x_{obs}|m=1)$

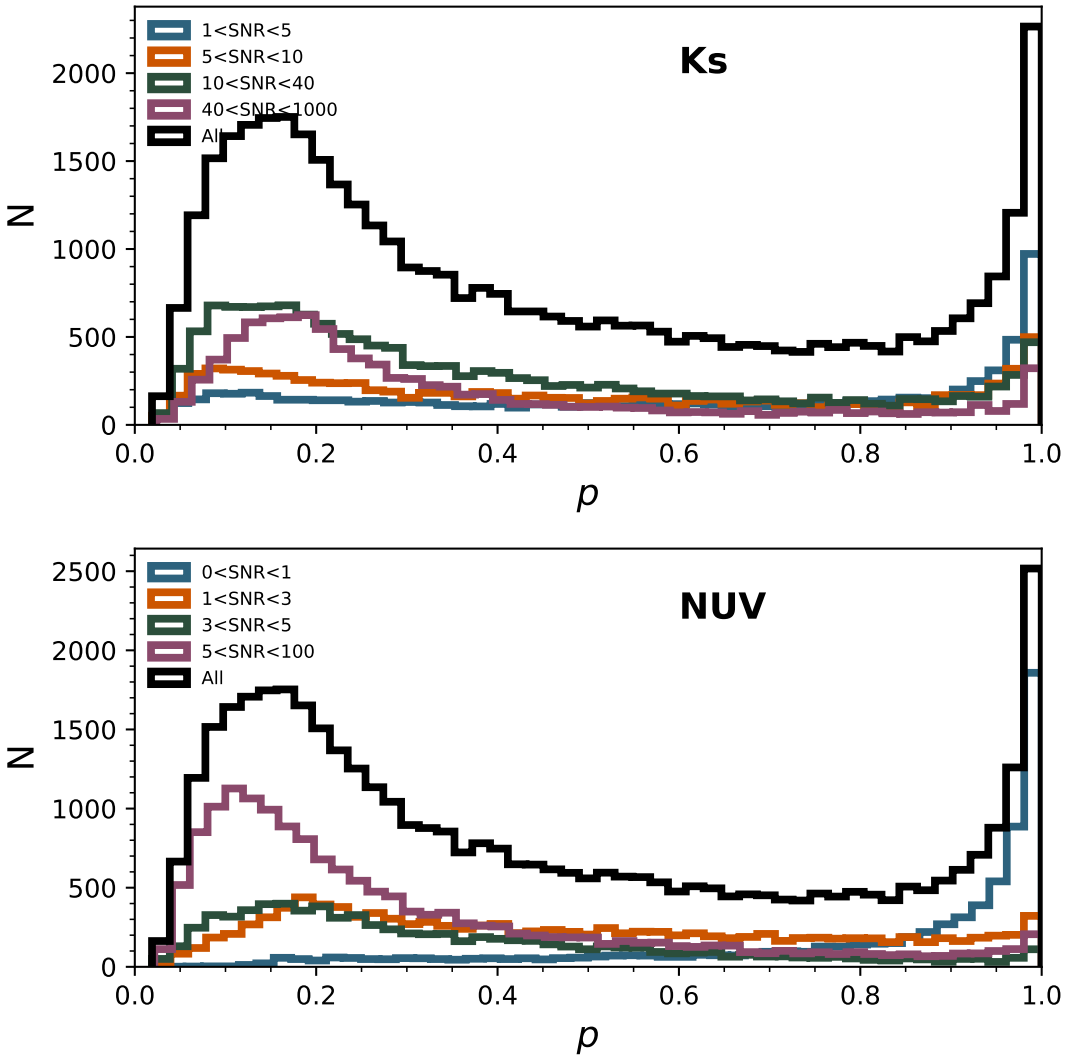

Figure 1. – Distribution of the predictions $\widehat{p}(m=1|x_{\rm obs})$ as a function of Ks band SNR (top panel) and NUV SNR (bottom panel). The different colors are for different selection in SNR in each panels.

In Fig. 1, we show the distribution of the estimated probability $\widehat{p}(m=1|x_{obs})$ for the subsample of COSMOS sources described in Sect. 2.2.2 before applying any SNR cuts. In this figure, all COSMOS sources with $M_* > 10^{8.5}\,M_\odot$ and redshift between 0.5 and 1 are used. The 0 value indicates that the delayed-τ SFH is preferred whereas $\widehat{p}=1$ indicates that the delayed-τ+flex SFH is more adapted to fit the SED of the galaxy. To understand what drives the shape of the $\widehat{p}(m=1|x_{\rm obs})$ distribution, we show in the

same figure the distributions obtained for different Ks SNR bins (top panel) and NUV SNR bins (bottom panel). Galaxies with low SNR in either NUV and Ks photometric band show flatter $\widehat{p}(m = 1|x_{\rm obs})$ distributions. This means that these low SNR sources yields to intermediate values of $\widehat{p}(m = 1|x_{\rm obs})$, translating into a difficulty to choose between the delayed-τ and the delayed-τ+flex SFHs.

B. Parameter tuning for Classification methods

The training catalog is used to optimize the value of ϕ with a specific algorithm given ψ, and the validation catalog is used to fit the tuning parameters ψ. To fit ϕ to a catalog of simulated datasets $\left(m^i, x^i\right)$, $i \in I$, the optimization algorithm specified with the machine learning model maximizes

$$\prod_{i \in I} L\Big(\widehat{p}(m = 1|x^i);\ m^i\Big)\Big(1 - \widehat{p}(m = 1|x^i)\Big)^{1-m^i}$$

given the value of ψ. Generally, this optimization algorithm is run for several values of ψ. Then, the validation catalog is used to calibrate the tuning parameters ψ based on data: the accuracy of $\widehat{p}_\psi(m = 1|x)$ for many possible values of ψ is computed on the validation catalog and we select the value $\widehat{\psi}$ that leads to the best results on this catalog. The resulting output of this two-step procedure is the approximation $\widehat{p}_{\widehat{\psi}}(m|x)$, that can be evaluated easily for new dataset x'. The accuracy of $\widehat{p}(m = 1|x)$ can be measured with various metrics. The most common metric is the classification error rate on a catalog of $\left(m^j, S(x^j)\right)$, $j \in J$, of $|J|$ simulations. We will rely on this metric. It is is defined by the frequency at which the datasets x^j are not well classified, i.e.,

$$\frac{1}{|J|} \sum_{j \in J} \mathbf{1}\Big\{\widehat{m}^j \neq m^j\Big\}, \Big(\mathbf{1}\{\widehat{p}(m = 1|x^j) \leq 1/2\}\Big)^{m^j}$$

C. Results on the tempered targets

Here, we consider that $\pi(x)$ and $q_t(x)$ are normalized densities.
For all $\beta \in [0;1]$, we introduce the normalized density

$$\pi_{\beta,t}(x) = \frac{1}{C_t(\beta)} \pi^\beta(x) q_t^{1-\beta}(x) \quad \text{where } C_t(\beta) = \int \pi^\beta(x) q_t^{1-\beta}(x) \mathrm{d}x.$$

Since the logarithm is a concave function, we have for all β and x,

$$\pi^\beta(x) q_t^{1-\beta}(x) \leq \beta\pi(x) + (1-\beta) q_t(x).$$

Thus, for all β, $C_t(\beta) \leq 1$. Moreover, $C_t(0) = C_t(1) = 1$.

Proposition 3. *The function $\beta \to \mathrm{KL}(\pi \| \pi_{\beta,t})$ is a convex, non increasing function. It decreases from $\mathrm{KL}(\pi|q_t)$ to 0.*

Proof of Proposition 3. Set for all β, $k(\beta) = \mathrm{KL}(\pi \| \pi_{\beta,t})$. We have

$$k(\beta) = \int \pi(x) \log \frac{\pi(x) C_t(\beta)}{\pi^\beta(x) q_t^{1-\beta}(x)} \mathrm{d}x = (1-\beta)\,\mathrm{KL}(\pi \| q_t) + \log C_t(\beta).$$

Hence its first and second derivatives are

$$k'(\beta) = -\mathrm{KL}(\pi \| q_t) + \frac{C_t'(\beta)}{C_t(\beta)}, \quad k''(\beta) = \frac{C_t''(\beta)}{C_t(\beta)} - \left(\frac{C_t'(\beta)}{C_t(\beta)}\right)^2. \tag{0.1}$$

On the other hand, the first and second derivative of $C_t(\beta)$ are

$$C_t'(\beta) = \int \pi^\beta(x) q_t^{1-\beta}(x) \log \frac{\pi(x)}{q_t(x)} \mathrm{d}x = C_t(\beta)\, \mathbb{E}_{\beta,t}\left(\log \frac{\pi(x)}{q_t(x)}\right),$$
$$C_t''(\beta) = \int \pi^\beta(x) q_t^{1-\beta}(x) \log^2 \frac{\pi(x)}{q_t(x)} \mathrm{d}x = C_t(\beta)\, \mathbb{E}_{\beta,t}\left(\log^2 \frac{\pi(x)}{q_t(x)}\right).$$

where $\mathbb{E}_{\beta,t}$ is the expected value when $x \sim \pi_{\beta,t}(x)$. Thus, using (0.1),

$$k''(\beta) = \mathrm{Var}_{\beta,t}\left(\log \frac{\pi(x)}{q_t(x)}\right) \geq 0$$

and $k(\beta)$ is a convex function.

Moreover, using (0.1) again, we have

$$k'(1) = -\mathrm{KL}(\pi \| q_t) + \frac{C_t'(1)}{C_t(1)} = -\mathrm{KL}(\pi \| q_t) + \int \pi(x) \log \frac{\pi(x)}{q_t(x)} \mathrm{d}x = 0.$$

Because of the convexity of k, for all $\beta \in [0,1]$, $k'(\beta) \leq k'(1) = 0$. Thus, $k(\beta)$ is decreasing and the proof is completed. □

The proposition given below is similar to the one of Beskos, Jasra, Kantas, et al., 2016, but the proof we give here deals with finite samples.

Proposition 4. *Consider a collection of positive weights w_i, $i = 1, \dots, n$. The function $\beta \to \mathrm{ESS}(\beta)$ defined by*

$$\mathrm{ESS}(\beta) = \left(\sum_{i=1}^n w_i^\beta\right)^2 \Big/ \left(\sum_{i=1}^n w_i^{2\beta}\right)$$

is decreasing.

Proof. If $x > 0$, the derivate of x^β with respect to β is $x^\beta \log x$. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\mathrm{ESS}(\beta) = \frac{2g(\beta)\sum_{i=1}^n w_i^\beta}{\left(\sum_{i=1}^n w_i^{2\beta}\right)^2} \quad \text{where}$$

$$g(\beta) = \left(\sum_{i=1}^n w_i^\beta \log w_j\right)\sum_{j=1}^n w_j^{2\beta} - \left(\sum_{j=1}^n w_j^{2\beta}\log w_i\right)\sum_{i=1}^n w_i^\beta.$$

Now,

$$\begin{aligned} g(\beta) &= \sum_{1\le i,j\le n} w_i^{2\beta} w_j^\beta \Big(\log w_j - \log w_i\Big) \\ &= \sum_{1\le i<j\le n} w_i^\beta w_j^\beta \Big(\log w_j - \log w_i\Big)\Big(w_i^\beta - w_j^\beta\Big) \\ &\le 0, \end{aligned}$$

since, for all $a, b > 0$,

$$a^{2\beta}b^\beta(\log b - \log a) + a^\beta b^{2\beta}(\log a - \log b) = a^\beta b^\beta\Big(a^\beta - b^\beta\Big)\log\frac{b}{a} \le 0. \qquad \square$$

D. Proof of Proposition 2

We start with this simple Lemma.

Lemma 5. *Let $f(x)$ and $g(x)$ be two densities on the x-space, which partitioned by $E \cup \bar{E}$. Introduce the normalized densities knowing $x \in E$ or $\bar{E}$ as*

$$f_{|E}(x) = \frac{1}{f(E)} f(x)\mathbf{1}_E(x), \quad f_{|\bar{E}}(x) = \frac{1}{f(\bar{E})} f(x)\mathbf{1}_{\bar{E}}(x)$$

and likewise for $g_{|E}$ and $g_{|\bar{E}}$. We have

$$\mathrm{KL}\big(f \big\| g\big) = f(E)\,\mathrm{KL}\big(f_{|E}\big\|g_{|E}\big) + f(\bar{E})\,\mathrm{KL}\big(f_{|\bar{E}}\big\|g_{|\bar{E}}\big) + f(E)\log\frac{f(E)}{g(E)} + f(\bar{E})\log\frac{f(\bar{E})}{g(\bar{E})}.$$

Proof. We have

$$\mathrm{KL}\left(f \middle\| g\right) = \int_E f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x + \int_{\bar{E}} f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x.$$

Moreover

$$\begin{aligned}
\int_E f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x &= \int_E f(E) f_{|E}(x) \log \frac{f(E) f_{|E}(x)}{g(E) g_{|E}(x)} \mathrm{d}x \\
&= f(E) \int_E f_{|E}(x) \log \frac{f_{|E}(x)}{g_{|E}(x)} \mathrm{d}x + f(E) \log \frac{f(E)}{g(E)} \\
&= f(E)\, \mathrm{KL}\left(f_{|E} \middle\| g_{|E}\right) + f(E) \log \frac{f(E)}{g(E)}.
\end{aligned}$$

Likewise,

$$\int_{\bar{E}} f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x = f(\bar{E})\, \mathrm{KL}\left(f_{|\bar{E}} \middle\| g_{|\bar{E}}\right) + f(\bar{E}) \log \frac{f(\bar{E})}{g(\bar{E})}. \qquad \square$$

Proof of Propostion 2. Using Lemma 5, $\mathrm{KL}\left(\pi_{\beta,t} \middle\| \widehat{\pi}_{\beta,t}\right) = \mathrm{KL}_I + \mathrm{KL}_{II} + \mathrm{KL}_{III}$ where

$$\begin{aligned}
\mathrm{KL}_I &= \pi_{\beta,t}(E)\, \mathrm{KL}\left(\pi^E_{\beta,t} \middle\| q^E_t\right) \\
\mathrm{KL}_{II} &= \pi_{\beta,t}(\bar{E})\, \mathrm{KL}\left(\pi^{\bar{E}}_{\beta,t} \middle\| \pi^{\bar{E}}_{\beta,t}\right) = 0 \\
\mathrm{KL}_{III} &= \pi_{\beta,t}(E) \log \frac{\pi_{\beta,t}(E)}{\lambda} + (1 - \pi_{\beta,t}(E)) \log \frac{1 - \pi_{\beta,t}(E)}{1 - \lambda}.
\end{aligned}$$

Likewise, $\mathrm{KL}\left(\pi_{\beta,t} \middle\| q_t\right) = \mathrm{KL}'_I + \mathrm{KL}'_{II} + \mathrm{KL}'_{III}$ where

$$\begin{aligned}
\mathrm{KL}'_I &= \pi_{\beta,t}(E)\, \mathrm{KL}\left(\pi^E_{\beta,t} \middle\| q^E_t\right) = \mathrm{KL}_I \\
\mathrm{KL}'_{II} &= \pi_{\beta,t}(\bar{E})\, \mathrm{KL}\left(\pi^{\bar{E}}_{\beta,t} \middle\| q^{\bar{E}}_t\right) \geq 0 = \mathrm{KL}_{II} \\
\mathrm{KL}'_{III} &= \pi_{\beta,t}(E) \log \frac{\pi_{\beta,t}(E)}{q_t(E)} + (1 - \pi_{\beta,t}(E)) \log \frac{1 - \pi_{\beta,t}(E)}{1 - q_t(E)}.
\end{aligned}$$

Moreover, when $s \leq 1$, $\lambda = s q_t(E) \leq q_t(E)$, thus $\mathrm{KL}'_{III} \geq \mathrm{KL}_{III}$. Finally,

$$\mathrm{KL}\left(\pi_{\beta,t} \middle\| \widehat{\pi}_{\beta,t}\right) = \mathrm{KL}_I + \mathrm{KL}_{II} + \mathrm{KL}_{III} \leq \mathrm{KL}'_I + \mathrm{KL}'_{II} + \mathrm{KL}'_{III} = \mathrm{KL}\left(\pi_{\beta,t} \middle\| q_t\right). \qquad \square$$