Inferring the photometric and size evolution of galaxies from image simulations. I. Method

Sébastien Carassou, Valérie de Lapparent, Emmanuel Bertin, Damien Le Borgne

▶ **To cite this version:**

Sébastien Carassou, Valérie de Lapparent, Emmanuel Bertin, Damien Le Borgne. Inferring the photometric and size evolution of galaxies from image simulations. I. Method. Astronomy and Astrophysics - A&A, 2017, 605, 10.1051/0004-6361/201730587 . insu-03747445

HAL Id: insu-03747445
https://insu.hal.science/insu-03747445

Submitted on 8 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A&A 605, A9 (2017)
DOI: 10.1051/0004-6361/201730587
© ESO 2017

Astronomy & Astrophysics

Inferring the photometric and size evolution of galaxies from image simulations

I. Method

Sébastien Carassou, Valérie de Lapparent, Emmanuel Bertin, and Damien Le Borgne

Institut d'Astrophysique de Paris, CNRS, UMR 7095 et Sorbonne Universités, UPMC Univ. Paris 6, 98bis bd Arago, 75014 Paris, France
e-mail: lapparent@iap.fr

Received 9 February 2017 / Accepted 15 April 2017

ABSTRACT

Context. Current constraints on models of galaxy evolution rely on morphometric catalogs extracted from multi-band photometric surveys. However, these catalogs are altered by selection effects that are difficult to model, that correlate in non trivial ways, and that can lead to contradictory predictions if not taken into account carefully.
Aims. To address this issue, we have developed a new approach combining parametric Bayesian indirect likelihood (pBIL) techniques and empirical modeling with realistic image simulations that reproduce a large fraction of these selection effects. This allows us to perform a direct comparison between observed and simulated images and to infer robust constraints on model parameters.
Methods. We use a semi-empirical forward model to generate a distribution of mock galaxies from a set of physical parameters. These galaxies are passed through an image simulator reproducing the instrumental characteristics of any survey and are then extracted in the same way as the observed data. The discrepancy between the simulated and observed data is quantified, and minimized with a custom sampling process based on adaptive Markov chain Monte Carlo methods.
Results. Using synthetic data matching most of the properties of a Canada-France-Hawaii Telescope Legacy Survey Deep field, we demonstrate the robustness and internal consistency of our approach by inferring the parameters governing the size and luminosity functions and their evolutions for different realistic populations of galaxies. We also compare the results of our approach with those obtained from the classical spectral energy distribution fitting and photometric redshift approach.
Conclusions. Our pipeline infers efficiently the luminosity and size distribution and evolution parameters with a very limited number of observables (three photometric bands). When compared to SED fitting based on the same set of observables, our method yields results that are more accurate and free from systematic biases.

Key words. galaxies: evolution – galaxies: bulges – galaxies: spiral – galaxies: luminosity function, mass function – galaxies: statistics – methods: numerical

1. Introduction

During the last decades our understanding of galaxy formation and evolution has been largely shaped by the results of deep multicolor photometric surveys. We can now extract the spectro-photometric properties of millions of galaxies, over large volumes that cover more than ten billion years of cosmic history. Despite this wealth of data, we are still incapable of deriving strong constraints on the free parameters of current semi-analytic models that describe quantitatively how galaxies evolve in color, size, and shape from their high redshifts counterparts. The main reason is that, missing physical ingredients in our models aside, the galaxy catalogs derived from surveys are often incomplete.

First of all, surveys are limited in flux. Consequently, intrinsically faint sources tend to be under-represented because they are above the limiting magnitude only at small distances. This effect, called Malmquist bias (Malmquist 1920), introduces correlations between probably non-correlated variables, mainly distance and other parameters such as luminosity (e.g., Singal & Rajpurohit 2014). Additionally, some galaxies overlap and may be blended into single objects. Source confusion (Condon 1974), caused by unresolved faint sources blended by the point spread function, can act as a signal at the detection limit and also affects number counts in a non-trivial way. Moreover, source confusion affects background estimation by adding a non-uniform component to the background noise, which is correlated with the spatial distribution of unresolved sources (Helou & Beichman 1990). Statistical fluctuations in flux measurements give rise to the Eddington bias (Eddington 1913). As galaxy number counts increase as a power of the flux, there are more overestimated fluxes for faint sources than underestimated fluxes for bright sources. This results in a general increase in the number of sources detected at a given flux (Hasinger & Zamorani 2000; Loaring et al. 2005). Because of the cosmological dimming, the bolometric surface brightness of galaxies gets dimmer with increasing redshift proportionally to $(1+z)^{-4}$ (Tolman & Richard 1934), which makes many faint extended sources undetectable. Finally, stellar contamination affects the bright end of the source counts (e.g., Pearson et al. 2014).

Apparent magnitudes in catalogs also have to be corrected for Galactic extinction (e.g., Schlegel et al. 1998), and to account for redshift effects, K-corrections (Hogg et al. 2002) that are sensitive to galaxy spectral type must be applied on the magnitudes of high-redshift galaxies (e.g., Ramos et al. 2011). Both corrections, however, are applied only after the sample is

truncated at its flux limit, which causes biases at the survey limit. Inclination-dependent internal absorption from dust lanes in the disk of galaxies also tends to draw a fraction of edge-on spirals below the survey flux limit (e.g., Kautsch et al. 2006). Because of these various selection effects, that correlate in ways that are poorly understood, and that may be spatially variable over the field of view of the survey, observations undergo complex selection functions that are difficult to treat analytically, and the resulting catalogs tend to be biased towards intrinsically brighter, compact, and low dust content sources.

The determination of the luminosity function (LF) of galaxies, a fundamental tool for characterizing galaxy populations that is often used for constraining models of galactic evolution, is particularly sensitive to these biases. As input data, analyses use catalogs containing the photometric properties, such as apparent magnitudes, of a selected galaxy sample. LF estimation requires the knowledge of the absolute magnitude of the sources, which itself depends upon the determination of their redshift. The number density per luminosity bin can be determined by a variety of methods, parametric or non-parametric, described in detail in Binggeli et al. (1988), Willmer (1997), and Sheth (2007). The resulting distribution is usually fitted by a Schechter function (Schechter 1976), but other functions are sometimes required (e.g., Driver & Phillipps 1996; Blanton et al. 2005). The Schechter function is characterized by three parameters: ϕ^* the normalization density, α the faint end slope, and M^* a characteristic absolute magnitude. The LF at $z \sim 0$ is presently well constrained thanks to the analysis of high-resolution spectroscopic surveys, such as the 2dF Galaxy Redshift Survey (2dFGRS, Norberg et al. 2002) or the Sloan Digital Sky Survey (SDSS, Blanton et al. 2003). There is also clear evidence that the global LF evolves with redshift, and that the LFs for different populations of galaxies evolve differently (Lilly et al. 1995; Zucca et al. 2006).

Measuring the LF evolution is nevertheless a challenge, as high-redshift galaxies are faint, and therefore generally unsuitable for spectroscopic redshift determination, which would require prohibitive exposure times. The current solution to this problem is to use the information contained within the fluxes of these sources in some broad-band filters, in order to estimate their redshift, known as photometric redshift. This procedure has a number of biases in its own right, because the precision of photometric redshifts relies on the templates and the training set used, assumed to be representative of the galaxy populations. These biases are described extensively in MacDonald & Bernstein (2010). In turn, redshift uncertainties typically result in an increase of the estimated number of low and high luminosity galaxies (Sheth 2007).

The forward-modeling approach to galaxy evolution. The traditional approach when comparing the results of models to data is sometimes referred to as backward modeling (e.g., Marzke 1998; Taghizadeh-Popp et al. 2015). In this scheme, physical quantities are derived from the observed data, and are then compared with the physical quantities predicted from simulations, semi-analytical models (SAM), or semi-empirical models. A more reliable technique is the forward modeling approach: a distribution of modeled galaxies are passed through a virtual telescope with all the observing process reproduced (filters, exposure time, telescope characteristics, seeing properties, as well as the cosmological and instrumental biases described above), and a direct comparison is made between simulated and observed datasets. The power of this approach comes from the fact that theory and observation are compared in the observational space: the same systematic errors and selection effects affect the simulated and observed data. Blaizot et al. (2005) were the first to introduce realistic mock telescope images from light cones generated by SAMs. Overzier et al. (2013) extended this idea by constructing synthetic images and catalogs from the Millenium Run cosmological simulation including detailed models of ground-based and space telescopes. More recently, Taghizadeh-Popp et al. (2015) used semi-empirical modeling to simulate *Hubble* Deep Field (HDF) images, from cutouts of real SDSS galaxies with modified sizes and fluxes, and compared them to observed HDF images. Here we make the case that forward modeling can be used to perform reliable inferences on the evolution of the galaxy luminosity and size functions.

Bayesian inference. Standard Bayesian techniques provide a framework to address any statistical inference problem. The goal of Bayesian inference is to infer the posterior probability density function (PDF) of a set of model parameters θ, given some observed data $\mathcal{D}$. This probability can be derived using Bayes' theorem:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}, \quad (1)$$

where $P(\mathcal{D}|\theta)$ is also called the likelihood (or the likelihood function) of the data, which gives the probability of the data given the model, $P(\theta)$ is the prior, or the probability of the model with the parameters θ, and $P(\mathcal{D})$ is the evidence, which acts as a normalization constant and is usually ignored in inference problems. The posterior PDF is approximated either analytically or via the use of sampling techniques, such as Markov chain Monte Carlo (MCMC).

However, there are multiple cases where the likelihood is intractable or unknown, for mathematical or computational reasons, which renders classical Bayesian approaches unfeasible. In our case, it is the modeling of the selection effects that is impractical to include in the likelihood. To tackle this issue, a new class of methods, called "likelihood-free", have been developed to infer posterior distributions without explicit computation of the likelihood.

Approximate Bayesian Computation. One of the "likelihood-free" techniques is called Approximate Bayesian Computation (ABC), and was introduced in the seminal article of Pritchard et al. (1999) for population genetics. ABC is based on repeated simulations of datasets generated by a forward model, and replaces the likelihood estimation by a comparison between the observed and synthetic data. Its ability to perform inference under arbitrarily complex stochastic models, as well as its well established theoretical grounds, have lead to its growing popularity in many fields, including ecology, epidemiology, and stereology (see Beaumont 2010, for an overview).

The classic ABC Rejection sampling algorithm, introduced in its modern form by Pritchard et al. (1999), is defined in Algorithm 1, where ρ is a distance metric built between the simulated and observed datasets, usually based on some summary statistics η, which are parameters that maximize the information contained within the datasets (for example, normally distributed datasets can be characterized using the mean and standard deviation of the underlying Gaussian distribution), and ϵ is a user-defined tolerance level >0. Using the ABC algorithm

for $t = 1$ *to* T **do**
  Repeat
    Generate θ^* from the prior distribution;
    Simulate data $\mathcal{D}^*$ from parameters θ^*;
  until $\rho(\eta(\mathcal{D}^*), \eta(\mathcal{D})) \leq \epsilon$;
  set $\theta_{(t)} = \theta^*$;
end

Algorithm 1: ABC Rejection sampling algorithm

with a good summary statistic and a small enough tolerance ultimately leads to a fair approximation of the posterior distribution (Sunnaker et al. 2013). The choices of ρ, η and ϵ are highly non-trivial though, and they constitute the fundamental difficulty in the application of ABC methods as they are problem-dependent (Marin et al. 2011). Moreover rejection sampling is notorious for its inherent inefficiency, as sampling directly from the prior distribution results in spending computing time simulating datasets in low-probability regions. Therefore, several classes of sampling algorithms have been developed to explore the parameter space more efficiently. Three of the most popular of them are outlined below.

- In the ABC-MCMC algorithm (Marjoram et al. 2003), a point in the parameter space called a particle performs a random walk (defined by a proposal distribution or transition kernel) across the parameter space, and is only moving if the simulated dataset generated by these parameters match better the observed dataset, until it converges to a stationary distribution. As in standard MCMC procedures, the efficiency of the algorithm is largely determined by the choice of the scale of the kernel.
- In the ABC Sequential Monte Carlo parallel algorithm (ABC-SMC, Toni et al. 2009), samples are drawn from the prior distribution until N particles are accepted, that is, those with a distance to the data $<\epsilon_0$. All accepted particles are attributed a statistical weight ω_0. The weighted particles then constitute an intermediate distribution from which another set of samples is drawn and perturbed with a fixed transition kernel, until N particles satisfy the acceptance criterion: $\rho < \epsilon_1$, with $\epsilon_1 < \epsilon_0$. They are then weighted with ω_1 and the process is repeated with a diminished tolerance at each step. After T iterations of this process, the particles are sampled from the approximate posterior distribution. The performance of ABC-SMC scales as N, where N is the number of particles. Different variations of ABC-SMC algorithms have been published, each with a different weighting scheme for particles.
- ABC Population Monte Carlo (ABC-PMC, Beaumont et al. 2009) is similar to ABC-SMC, but differs in its adaptive weighting scheme: its transition kernel is Gaussian and based on the variance of the accepted particles in the previous iteration. This scheme requires the fewest tuning parameters of the three algorithms discussed here (Turner & Van Zandt 2012). But ABC-PMC is also more computationally costly than ABC-SMC, as its performance scales as N^2 (caused by its adaptability).

The reader is referred to Csilléry et al. (2010), Marin et al. (2011), Turner & Van Zandt (2012), Sunnaker et al. (2013), and Gutmann & Corander (2016) for a set of historical, methodical, and theoretical reviews of this final approach, as well as a complete description of the algorithms mentioned above.

Parametric Bayesian indirect likelihood. Another class of likelihood-free techniques is called parametric Bayesian indirect likelihood (pBIL). First proposed by Reeves & Pettitt (2005) and Gallant & McCulloch (2009), pBIL transforms the intractable likelihood of complex inference problems into a tractable one using an auxiliary parametric model that describes the simulated datasets generated by the forward model. In this scheme, the resulting auxiliary likelihood function quantifies the discrepancy between the observed and simulated data. It is used in Bayes' theorem and the parameter space is explored using a user-defined sampling procedure, in an equivalent way to a classical Bayesian technique. While sharing similarities with the previous technique, pBIL is not an ABC method in the strict sense, as it does not require an appropriate choice of summary statistics and tolerance level to compare the observed and synthetic datasets. The accuracy of the inference in the pBIL scheme is determined by how well the auxiliary model describes the data (observed and simulated). The theoretical foundations of this scheme are described extensively in Drovandi et al. (2015).

Application of likelihood-free inference to astrophysics. The application of likelihood-free methods to astrophysics is still rare, as noted by Cameron & Pettitt (2012) in their review. Only lately has the potential of such techniques been considered. Schafer & Freeman (2012) praised the use of likelihood-free inference in the context of quasar luminosity function estimation. Cameron & Pettitt (2012) explored the morphological transformation of high-redshift galaxies and derived strong constraints on the evolution of the merger rate in the early Universe using an ABC-SMC approach. Weyant et al. (2013) also used SMC for the estimation of cosmological parameters from type Ia supernovae samples, and could still provide robust results when the data was contaminated by type IIP supernovae. Robin et al. (2014) constrained the shape and formation period of the thick disk of the Milky Way using MCMC as their sampling scheme, based on photometric data from the SDSS and the Two Micron All Sky Survey (2MASS). Finally Hahn et al. (2017) demonstrate the feasibility of using ABC to constrain the relationship between galaxies and their dark matter halo. The recent birth of Python packages providing sampling algorithms in an ABC framework, such as astroABC (Jennings & Madigan 2017) and ELFI (Kangasrääsiö et al. 2016), which implement SMC methods, and COSMOABC (Ishida et al. 2015) which implements the PMC algorithm, will probably facilitate the rise of likelihood-free inference techniques in the astronomical community.

Outline of the article. To the authors' knowledge, no likelihood-free inference approaches have yet included telescope image simulation in their forward modeling pipeline, because of the difficulty in implementation as well as a prohibitive computational cost. Prototypical implementations in a cosmological context have, however, been tested by Akeret et al. (2015) on a Gaussian toy model for the calibration of image simulations. In the present article we propose a new technique that combines the forward modeling approach with sampling techniques in the pBIL framework. In that regard, we use a stochastic semi-empirical model of evolving galaxy populations coupled to an image simulator to generate realistic synthetic images. Simulated images go through the same source extraction process and data analysis pipeline as real images. The observed and synthetic

data distributions are finally compared and used to infer the most probable models.

This article is organized as follows: Sects. 2 to 5 describe in detail the forward-modeling pipeline we propose, from model parameters to data analysis and sampling algorithm. Section 6 defines our convergence diagnostics. In Sect. 7, we demonstrate the validity, internal consistency and robustness of our approach by inferring the LF parameters and their evolution using one realization of our model as input data. We perform these tests in two situations: a configuration where the data is a mock Canada-France-Hawaii Telescope Legacy Survey (CFHTLS) Deep image containing two populations of ellipticals and lenticulars and late-type spirals, and where the parameters to infer are the evolving luminosity function parameters for each population (Sect. 7.4); and a configuration where the data is a mock CFHTLS Deep image with a single population of pure bulge elliptical galaxies, and in which the inference is performed on the evolving size and luminosity (Sect. 7.6). In Sect. 8, we compare the results of our forward modeling approach with those of the more traditional photometric redshift approach applied to the same situation. Finally, Sect. 9 provides suggestions to improve the speed and accuracy of this method.

Throughout this article, unless stated otherwise, we adopt the following cosmological parameters: $H_0 = 100\,h$ km s^{-1} Mpc^{-1} with $h = 1$, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$ (Spergel et al. 2003). Magnitudes are given in the AB system.

2. Model: from parameters to image generation

In order to infer the physical properties of galaxies from observed survey images without having to describe the complex selection effects the latter contain, we propose the following pipeline. We start from a set of physical input parameters, drawn from the prior distribution defined for each parameter. These parameters describe the luminosity and size distribution of the various populations of modeled galaxies. From this set of parameters, our forward model generates a catalog of galaxies modeled as the sum of their bulge and disk components, each with a different profile. The projected light profiles of the galaxies are determined by their inclination, the relative fraction of light contained within the bulge, and the galaxy redshift as well as the extinction of the bulge and disk components. The galaxies are randomly drawn from the luminosity function of their respective population. The catalog assumes that galaxies are randomly distributed on a virtual sky that includes the cosmological effects of an expanding universe with a cosmological constant. The survey image is simulated in every band covered by the observed survey, and reproduces all of its characteristics, such as filters transmission, exposure time, point spread function (PSF) model, and background noise model.

Then, a large number of "simulated" images are generated via an iterative process (a Markov chain) generating new sets of physical parameters at each iteration. Some basic flux and shape parameters are extracted in the same way from the observed and simulated images: after a pre-processing step (which is identical for observed and simulated data) where observables are decorrelated and their dynamic range reduced, the multidimensional distributions of simulated observables are directly compared to the observed distributions using a custom distance function on binned data.

The chain moves through the parameter space towards regions of high likelihood, that is, regions that minimize the distance between the modeled and observed datasets. The pathway of the chain is finally analyzed to reconstruct the multidimensional posterior probability distribution and infer the sets of parameters that most likely reproduce the observed catalogs, as well as the correlations between these parameters. The main steps of this approach are detailed in the sections below, and the whole pipeline is sketched in Fig. 1 of this article.

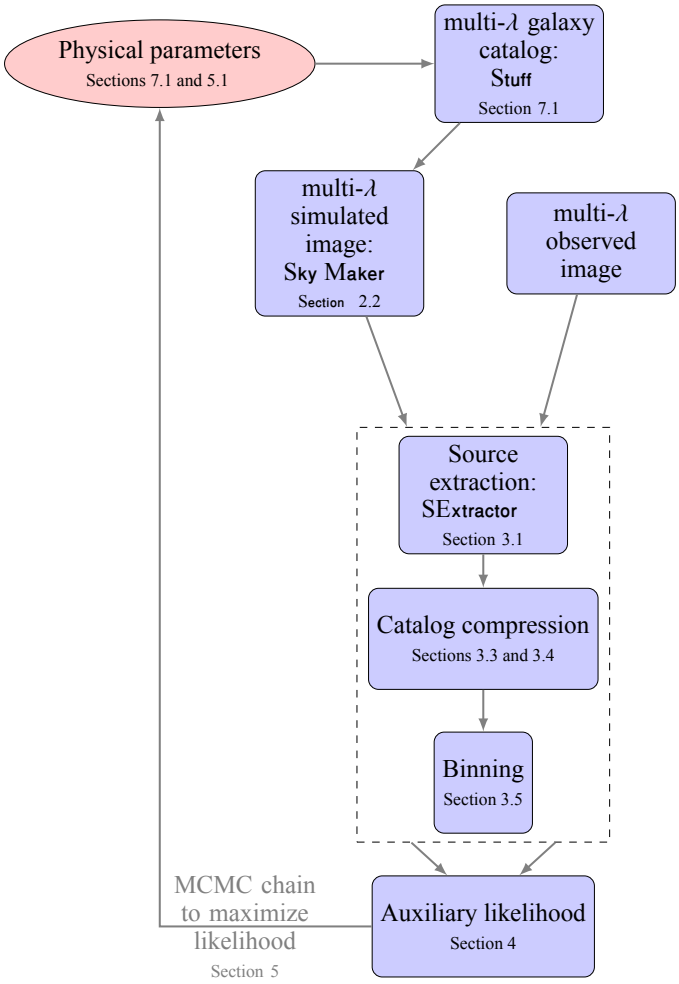


Fig. 1. Summary of the workflow.

2.1. Physical parameters and source catalog generation

Artificial catalogs are generated with the STUFF package (Bertin 2009) in fields of a given size. STUFF relies on empirical scaling laws applied to a set of galaxy "types", which it uses to draw galaxy samples with photometric properties computed in an arbitrary number of observation passbands. Each galaxy type is defined by its Schechter (1976) luminosity function parameters, its spectral energy distribution (SED), as well as the bulge-to-total luminosity ratio B/T and rest-frame extinction properties of each component of the galaxy through a "reference" passband.

The photometry of simulated galaxies is based on the composite SED templates of Coleman et al. (1980) extended by Arnouts et al. (1999). Any of the six "E", "S0", "Sab", "Sbc", "Scd", and "Irr" SEDs can be assigned to the bulge and disk components separately, for a given galaxy type. The version of STUFF used in this work does not allow the SEDs to evolve with redshift; instead, following Gabasch et al. (2004), galaxy evolution is modeled as a combination of density (Schechter's ϕ^*) and luminosity (Schechter's M^*) evolution with redshift z:

$$M^*(z) = M^*(0) + M_e \ln(1+z) \quad (2)$$

$$\phi^*(z) = \phi^*(0)(1+z)^{\phi_e}, \quad (3)$$

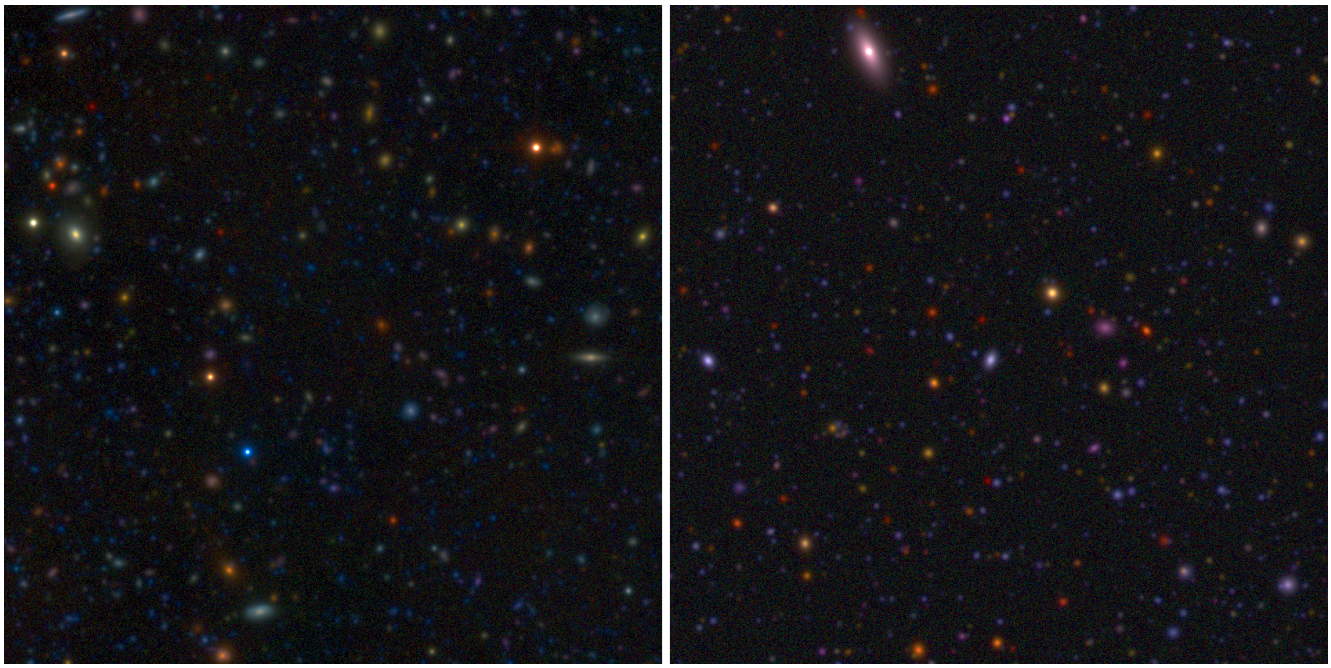


Fig. 2. Comparison between an observed survey image and a mock image generated by our model. *On the left*: a region of the CFHTLS D1 field (stack from the 85% best seeing exposures) built from the *gri* bands. *On the right*: a simulated image with STUFF+SKYMAKER with the same filters, exposure time, and telescope properties as the CFHTLS data. Both images are shown with the same color coding.

where M_e and ϕ_e are constants. The reference filter (i.e. the filter where the LF is measured) is set to the g-band in the present article.

Bulges and elliptical galaxies have a de Vaucouleurs (1953) profile:

$$\mu_b(r) = M_b + 8.3268\left(\frac{r}{r_b}\right)^{\frac{1}{4}} + 5\log r_b + 16.6337, \quad (4)$$

where $\mu_b(r)$ is the bulge surface brightness in mag pc^{-2}, $M_b = M - 2.5\log(B/T)$ is the absolute magnitude of the bulge component and M the total absolute magnitude of the galaxy, both in the reference passband. As a projection of the fundamental plane, the average effective radius $\langle r_b \rangle$ in pc follows an empirical relation we derive from the measurements of Binggeli et al. (1984):

$$\langle r_b \rangle = \begin{cases} r_{knee} 10^{-0.3(M_b - M_{knee})} & \text{if } M_b < M_{knee} \\ r_{knee} 10^{-0.1(M_b - M_{knee})} & \text{otherwise} \end{cases} \quad (5)$$

where $r_{knee} = 1.58\,h^{-1}$ kpc and $M_{knee} = -20.5$. The intrinsic flattening q of bulges follows a normal distribution with $\langle q \rangle = 0.65$ and $\sigma_q = 0.18$ (Sandage et al. 1970), which we convert to the apparent aspect-ratio $\sqrt{q^2 \sin^2 i + \cos^2 i}$, where i is the inclination of the galaxy with respect to the line of sight.

Disks have an exponential profile:

$$\mu_d(r) = M_d + 1.8222\left(\frac{r}{r_d}\right) + 5\log r_d + 0.8710, \quad (6)$$

where $\mu_d(r)$ is the disk surface brightness in mag pc^{-2}, $M_d = M - 2.5\log(1-(B/T))$ is the absolute magnitude of the disk in the reference passband, and r_d the effective radius. Semi-analytical models where disks originate from the collapse of the baryonic content of dark-matter-dominated halos (Dalcanton et al. 1997; Mo et al. 1998) predict useful scaling relations. Assuming that light traces mass and that there is negligible transport of angular momentum during collapse, one finds $r_d \propto \lambda L_d^{-\beta}$, where λ is the dimensionless spin parameter of the halo, $L_d = 10^{-0.4M_d}$ the total disk luminosity, and $\beta \simeq -1/3$ (de Jong & Lacey 2000). The distribution of λ, as seen in N-body simulations, can well be described by a log-normal distribution (Warren et al. 1992), and is very weakly dependent on cosmological parameters (Steinmetz & Bartelmann 1995), hence the distribution of r_d at a given M_d should behave as:

$$n(r_d|M_d) \propto \frac{1}{r_d} \exp\left[-\frac{\left(\ln(r_d/r_d^*) - 0.4\beta_d(M_d - M_d^*)\right)^2}{2\sigma_\lambda^2}\right]. \quad (7)$$

In de Jong & Lacey (2000), a convincing fit to I-band catalog data of late-type galaxies corrected for internal extinction is obtained, with $\beta_d = -0.214$, $\sigma_\lambda = 0.36$, $r_d^* = 5.93$ kpc, and $M_d^* = -22.3$ (for $H_0 = 65$ km s^{-1}). Both bulge and disk effective radii are allowed to evolve (separately) with redshift z using simple $(1+z)^\gamma$ scaling laws (see, e.g., Trujillo et al. 2006; Williams et al. 2010). The original values from Trujillo et al. (2006) are modified to those in Table 5 based on the *Hubble* Space Telescope Ultra Deep Field (UDF, Williams et al. 2010; Bertin, priv. comm.).

Internal extinction is applied (separately) to the bulge and disk SEDs $S(\lambda)$ using the extinction law from Calzetti et al. (1994), extended to the UV and the IR assuming an LMC law (Charlot, priv. comm.):

$$S(\lambda) = S_0(\lambda) e^{-\kappa\tau(\lambda)}, \quad (8)$$

where $S_0(\lambda)$ is the face-on, unextincted SED and $\tau(\lambda)$ the uncalibrated extinction law. The normalization factor κ is computed by integrating the effect of extinction A_{ref}, expressed in magnitudes, within the reference passband $p_{ref}(\lambda)$:

$$A_{ref} = -2.5\log_{10}\frac{\int p_{ref}(\lambda) S_0(\lambda) e^{-\kappa\tau(\lambda)} d\lambda}{\int p_{ref}(\lambda) S_0(\lambda) d\lambda}. \quad (9)$$

As the variation of $\tau(\lambda)$ is small within the reference passband, we take advantage of a second order Taylor expansion of both the exponential and the logarithm:

$$A_{\rm ref} \approx -2.5\log_{10}\left(1 - I_1\kappa + \frac{1}{2}I_2\kappa^2\right) \quad (10)$$

$$\approx 1.086\left(I_1\kappa + \frac{I_1^2 - I_2}{2}\kappa^2\right), \quad (11)$$

with

$$I_1 = \frac{\int p_{\rm ref}(\lambda)S_0(\lambda)\tau(\lambda)\mathrm{d}\lambda}{\int p_{\rm ref}(\lambda)S_0(\lambda)\mathrm{d}\lambda}, \quad I_2 = \frac{\int p_{\rm ref}(\lambda)S_0(\lambda)\tau^2(\lambda)\mathrm{d}\lambda}{\int p_{\rm ref}(\lambda)S_0(\lambda)\mathrm{d}\lambda}. \quad (12)$$

Solving the quadratic Eq. (11) we obtain:

$$\kappa \approx \frac{-2A_{\rm ref}}{1.086\left(I_1 + \sqrt{I_1^2 - \frac{2}{1.086}(I_1^2 - I_2)A_{\rm ref}}\right)}. \quad (13)$$

We adopt the parametrization of the extinction from the RC3 catalog (de Vaucouleurs et al. 1991):

$$A_{\rm ref} = -\alpha(\mathrm{T})\log_{10}(\cos i), \quad (14)$$

where i is the disk inclination with respect to the line-of-sight, and $\alpha(T)$ (not to be confused with Schechter's α) is a type-dependent "extinction coefficient" that quantifies the amount of extinction+diffusion in the blue passband. For simplicity we identify this passband with our reference g passband, although they do not exactly match. The extinction coefficient evolves with de Vaucouleurs (1959) revised morphological type as:

$$\alpha(T) = \begin{cases} 1.5 - 0.03(T-5)^2 & \text{for } T \geq 0 \\ 0 & \text{for } T \leq 0. \end{cases} \quad (15)$$

STUFF applies to SEDs the mean intergalactic extinction curve at the given redshift following Madau (1995) and Madau et al. (1996), using the list of Lyman wavelengths and absorption coefficients from the XSPEC code (Arnaud 1996). Galaxies are Poisson distributed in 5 h^{-1} Mpc redshift slices from $z = 20$ to $z = 0$. For now the model does not include clustering properties, therefore the galaxies positions are uniformly distributed over the field of view. Ultimately STUFF generates a set of mock catalogs (one per filter) to be read by the image simulation software, containing source position, apparent magnitude, B/T, bulge and disk axis ratios and position angles, and redshift. We note that for consistency, we kept most of the default values applied by STUFF to scaling parameters, although many of them come from slightly outdated observational constraints dating back to the mid-2000's (and even earlier). This of course does not affect the conclusions of this paper.

2.2. Image generation

STUFF catalogs are turned into images using the SKYMAKER package (Bertin 2009). Briefly, SKYMAKER renders simplified images of galaxy models as the sum of a Sérsic (1963) "bulge" and an exponential "disk" on a high resolution pixel grid. The models are convolved by a realistic PSF model generated internally, or derived from actual observations using the PSFEX tool (Bertin 2011a). Each convolved galaxy image – or point source for stars – is subsampled at the final image resolution using a Lanczos-3 kernel (Wolberg & George 1990) and placed on the pixel grid at its exact catalog coordinates. The next step involves large scale features: convolution by a PSF aureole (e.g., Racine 1996), addition of the sky background, and simulation of saturation features (bleed trails). Finally, photon (Poisson) and readout (Gaussian) noise are added according to the characteristics of the instrument being simulated, and the data are converted to ADUs (analog-to-digital units). An example of a simulated deep survey field is shown Fig. 2.

3. Compression of data: from source extraction to binning

3.1. Source extraction

The SEXTRACTOR package (Bertin & Arnouts 1996) produces photometric catalogs from astronomical images. Briefly, sources are detected in four main steps: first, a smooth model of the image background is computed and subtracted. Second, a convolution mask, acting as matched filter, is applied to the background-subtracted image for improving the detection of faint sources. Third, a segmentation algorithm identifies connected regions of pixels with a surface brightness in the filtered image higher than the detection threshold. Finally, the same segmentation process is repeated at increasing threshold levels to separate partially blended sources that may share light at the lowest level.

Once a source has been detected, SEXTRACTOR performs a series of measurements according to a user-defined parameter list. This includes various position, shape, and flux estimates. For this work we rely on FLUX_AUTO photometry. FLUX_AUTO is based on Kron's algorithm (Kron 1980) and gives reasonably robust photometric estimates for all types of galaxies. For object sizes we choose the half-light radius estimation provided by the FLUX_RADIUS parameter, which is the radius of the aperture that encloses half of the FLUX_AUTO source flux. We note that this size estimate includes the convolution of the galaxy light profile by the PSF. In order to retrieve properties such as color, SEXTRACTOR is run in the so-called double image mode, where detection is carried out in one image and measurements in another. By repeating source extraction with the same "detection image", but with "measurement images" in different filters, we ensure that the photometry is performed in the exact same object footprints in all filters.

SEXTRACTOR flags all issues occurring during the detection and measurements processes. In this work, we consider only detections with a SEXTRACTOR FLAG parameter less than four, which excludes sources that are saturated or truncated by the frame boundaries.

3.2. Parallelization

By construction, our sampling procedure based on MCMC (cf. Sect. 5) cannot be parallelized, because the knowledge of the $n-1$th iteration is required to compute the nth iteration. We can, however, parallelize the process of source extraction and, most importantly, image simulation. In fact, we find in performance tests that the pipeline runtime is largely dominated by the image generation process (cf. Fig. 4), and that the image generation time scales linearly with the area of the simulated image. Simulating a single image per band containing all the sources for every iteration would make this problem computationally unfeasible in terms of execution time. In order to limit the runtime of an iteration, the image making step is therefore split into $N_{\rm sub} \times N_{\rm f}$ parallel small square patches, as illustrated in Fig. 3, where $N_{\rm f}$ is the number of filters fixed by the observed data and

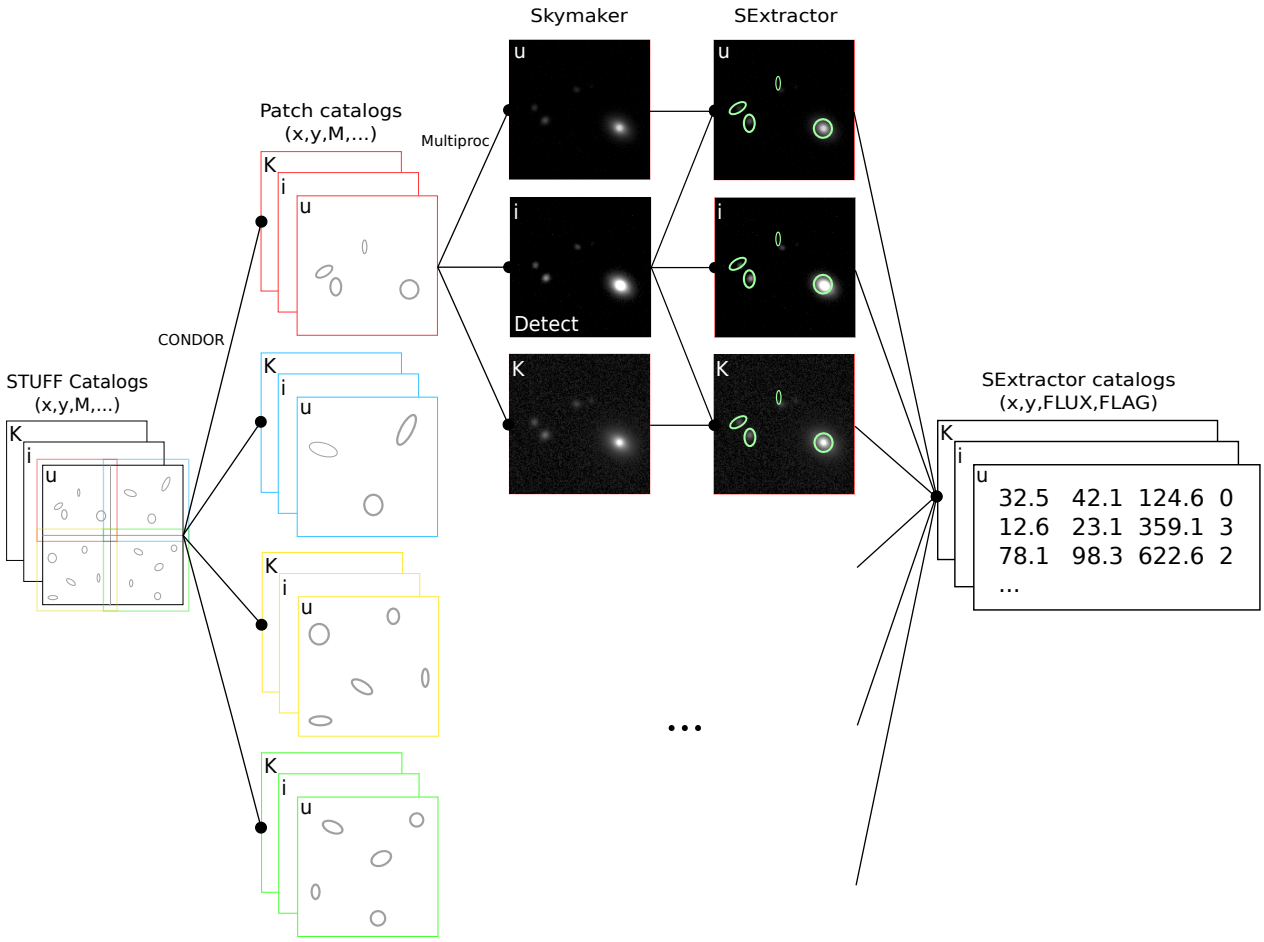


Fig. 3. Illustration of the parallelization process of our pipeline, described in detail in Sect. 3.2. STUFF generates a catalog, that is, a set of files containing the properties of simulated galaxies, such as inclination, bulge-to-disk ratio, apparent size, and luminosity. Each file lists the same galaxies in a different passband. The parallelization process is performed on two levels: first, the STUFF catalogs are split into sub-catalogs according to the positions of the sources on the image. These sub-catalogs are sent to the nodes of the computer cluster in all filters at the same time using the HTCondor framework. Each sub-catalog is then used to generate a multiband image corresponding to a fraction of the total field. This step is multiprocessed in order to generate the patches in every band simultaneously. SEXTRACTOR is then launched on every patch synchronously, also using multiprocessing. The source detection is done in one pre-defined band, and the photometry is done in every band. Finally, the SEXTRACTOR catalogs generated from all the patches are merged into one large catalog containing the photometric and size parameters of the extracted sources from the entire field.

$N_{\rm sub}$ the user-defined number of patches per band. Both quantities must be chosen so that their product optimizes the resources used by the computing cluster.

We start with $N_{\rm f}$ input catalogs generated from the model, each containing a list of sources' positions in a full-sized square field of size $L_{\rm f}$, as well as their photometric and size properties. The sources are then filtered according to their spatial coordinates and dispatched to their corresponding patch. Each patch has a size $L_{\rm f}/\sqrt{N_{\rm sub}}$, where $N_{\rm sub}$ is a square number. In practice, the sources are extracted from a box 150 pixels wider than the patch size in order to include the objects outside the frame that partially affect the simulated image. All the sources of position (x, y) are within a patch of coordinate $(i, j) \in [0,\ \sqrt{N_{\rm sub}} - 1] \times [0,\ \sqrt{N_{\rm sub}} - 1]$ if $x \in [i\frac{L_{\rm f}}{\sqrt{N_{\rm sub}}} - 150, (i+1)\frac{L_{\rm f}}{\sqrt{N_{\rm sub}}} + 150]$, and $y \in [j\frac{L_{\rm f}}{\sqrt{N_{\rm sub}}} - 150, (j+1)\frac{L_{\rm f}}{\sqrt{N_{\rm sub}}} + 150]$.

As a result, all the sources are scattered through $N_{\rm sub}$ catalog files per band. We then use the HTCondor distributed jobs scheduler on our computing cluster to generate and analyze all the patches at the same time. The flexibility of HTCondor offers many advantages to a pipeline that requires distributed computing over long periods of time. Thanks to its dynamic framework, jobs can be check pointed and resumed after being migrated if a node of the cluster becomes unavailable, and the scheduler efficiently provides an efficient match-making between the required and the available resources. This framework also has its drawbacks, in the form of inherent and uncontrollable latencies when jobs input files are sent to the various nodes.

In our case, each job corresponds to a single patch, and the $N_{\rm sub} \times N_{\rm f}$ resulting catalogs serve as input files for the jobs. We found that HTCondor latencies represent between 7% and 50% of the run time of each iteration, as illustrated in Fig. 4 in the context of the application described below (cf. Sect. 7).

For each job, the image generation and source extraction procedures are multiprocessed: SKYMAKER is first launched simultaneously in every band on the $L_{\rm f}/\sqrt{N_{\rm sub}}$-sized patch and, when all the images are available, SEXTRACTOR is launched in double image mode. Condor then waits until all jobs are completed. Finally, the catalog files generated from all the patches are merged

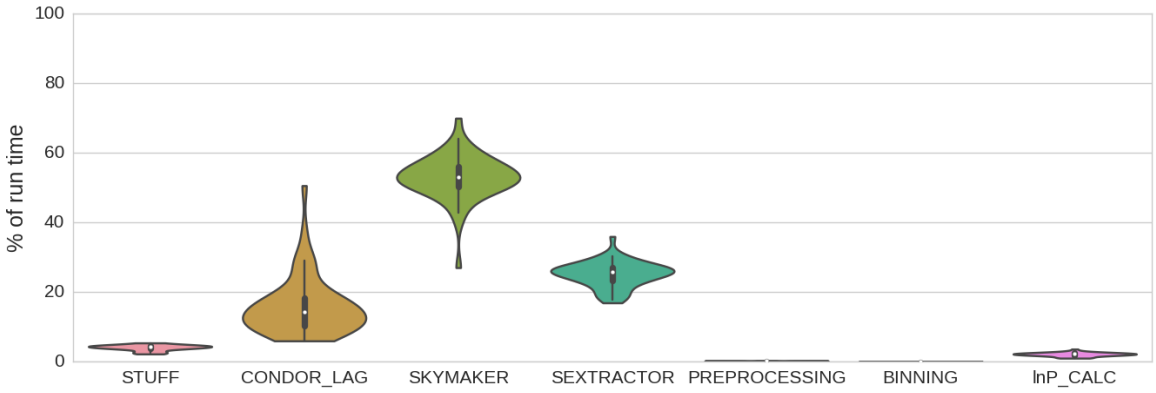


Fig. 4. Benchmarking of a full iteration of our pipeline, obtained with 50 realizations of the same iteration. An iteration starts with the STUFF catalog generation (here we consider a case where ~55 000 sources spread into two populations of galaxies are produced), and ends with the posterior density computation. The runtime of each subroutine called is analyzed in terms of the fraction of the total runtime of the iteration. In this scheme, the image simulation step clearly dominates the runtime, followed by the source extraction step and the HTCondor latencies. Source generation, pre-processing, binning and posterior density calculation (labeled lnP_CALC), however, account for a negligible fraction of the total runtime.

into one, so that in fine, a single catalog file per band contains all the extracted sources.

3.3. Reduction of the dynamic ranges

Observables such as fluxes may have a large dynamic range that goes up to the saturation level of the chosen survey. This can be problematic for the binning process of our pipeline, in the sense that it will create many sparsely populated bins. We must therefore reduce the dynamic range of the photometric properties of the sources. We cannot simply use the log of the flux arrays, because the noise properties of background-subtracted images can provide faint objects with negative fluxes. We therefore use the following transform $g(X)$, which has already been applied to model-fitting and machine learning applications (e.g., Bertin 2011b):

$$X_\mathrm{r} = g(X) = \begin{cases} \kappa_\mathrm{c}\sigma \ln\left(1 + \frac{X}{\kappa_\mathrm{c}\sigma}\right) \text{ if } X \geq 0, \\ -\kappa_\mathrm{c}\sigma \ln\left(1 - \frac{X}{\kappa_\mathrm{c}\sigma}\right) \text{ otherwise,} \end{cases} \tag{16}$$

where σ is the baseline standard deviation of X (i.e., the average lowest flux error), and κ_c a user-defined factor which can be chosen in the range from 1 to 100, typically. In all the test cases that we describe in Sect. 7, we set $\kappa_\mathrm{c} = 10$. In practice we apply this compression to each dimension of the observable space, with a different value of σ for each observable. We separate the σ values into two categories for each kind of observable: σ_f for flux-related observables and σ_r for size-related ones. These values are affected by the galaxy populations in the observed field as well as the photometric properties of the field itself, such as the bands used and the noise properties. For fluxes and colors, a root mean square error estimate of the flux measurement is given by SExtractor: FLUXERR_AUTO. We set σ_f to the median value of the distribution of FLUXERR_AUTO values for the sources extracted from input data, and this operation is repeated on each filter. However, SExtractor provides no such error estimate for FLUX_RADIUS. For this kind of observable we rely on the distribution of FLUX_RADIUS of the extracted sources with respect to the corresponding FLUX_AUTO. For each passband, the value of σ_r is set to the approximate FLUX_RADIUS of the extracted sources' distribution when FLUX_AUTO tends to 0. The exact values actually do not matter, because the same compression is applied on the observed and simulated data.

3.4. Decorrelation of the observables: whitening transformation

The choice of the nature and number of observables is a compromise between computational cost and informational content. In fact, memory limitations intrinsic to the computational cluster when binning observed and synthetic data (cf. Sect. 3.5) prevent us from using an arbitrary number of observables in the pipeline. Observables such as fluxes or magnitudes in different passbands also tend to be correlated with one another, as they originate from the same spectrum of a given galaxy from a given population. These correlations can be high if the passbands are too narrow, too close to each other, and not covering a large enough wavelength baseline. One must thoughtfully choose the appropriate set of filters a priori in order for the resulting set of observables to be able to disentangle the luminous properties of the different galaxy populations.

Strong correlations between input vector components can also make binning very inefficient, therefore an important preprocessing step is to decorrelate them. In that regard, we apply a linear transformation called principal component analysis whitening, or sphering (Friedman 1987; Hyvärinen et al. 2009; Shlens 2014; Kessy et al. 2017) to our reduced matrix of observables X_r of size $p \times N_\mathrm{s}$, where p is the number of observables and N_s is the number of sources. Principal component analysis (PCA) is an algorithm commonly used in the context of dimensionality reduction. Its goal is to find a set of orthogonal axes in a dataset called principal components that encapsulate most of the variance of the data. This can be performed via a singular values decomposition (SVD) of the covariance matrix of the data:

$$\left\langle X_\mathrm{r} X_\mathrm{r}^\mathrm{T} \right\rangle = \mathbf{U}\Lambda\mathbf{V}^\mathrm{T}, \tag{17}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and Λ the diagonal matrix containing the non-negative singular values of the covariance matrix, sorted by descending order.

PCA whitening is the combination of two operations: rotation and scaling. First the dataset (previously centered around zero by subtracting the mean in each dimension) is projected along the principal components, which removes linear correlations, and then each dimension is scaled so that its variance

equals to one. The whitening transform can therefore be summarized by:

$$X_{\rm w} = \Lambda^{-\frac{1}{2}} \mathbf{V}^{\rm T}(X_{\rm r} - \mu), \tag{18}$$

where $X_{\rm w}$ is the whitened version of the observables matrix $X_{\rm r}$ and μ is the average matrix. The PCA whitening transformation results in a set of new variables that are uncorrelated and have unit variance ($\langle X_{\rm w} X_{\rm w}^{\rm T} \rangle = I$). During the chain iterations, the observed and simulated data are centered, rotated, and scaled in the same way to ensure that both distributions can be well superposed and compared (cf. Sect. 4).

In practice, the simulated data is whitened using the Λ, $\mathbf{V}^{\rm T}$, and μ of the observed data. The number of principal components to keep is left to the choice of the user. Retaining only the components with the highest variance and therefore reducing the computational cost of the pipeline may be tempting. Nevertheless, subtle but important features can arise from low variance components, and deleting them comes at a price. In our application (cf. Sect. 7), we choose not to reduce the dimensionality of the problem.

3.5. Binning in the observational space

It remains to quantify the similarity between the two multivariate datasets, one containing preprocessed observables from the observations and the other from a simulation. Following the idea of Robin et al. (2014) and Rybizki & Just (2015), who grouped their data representing stellar photometry into bins of magnitude and color, we choose to bin our datasets, considering the relative simplicity and advantageous computational cost of this method. However, binning comes with some inevitable drawbacks: the number of bins increases exponentially with the number of dimensions. For a fixed-size dataset, multivariate histograms are also sparser than their univariate counterparts and display more complex shapes. Finally the choice of the binning scheme can significantly influence the information content of the dataset, and that choice is not trivial in high-dimensional spaces (Cadez et al. 2002). This class of problems is known as "the curse of dimensionality" (Bellman 1972).

Several binning schemes have been developed, like the Freedman & Diaconis (1981) rule extended to several dimensions, Knuth's rule (Knuth 2006), which uses Bayesian model selection to find the optimal number of bins, Hogg's rule (Hogg 2008), or Bayesian blocks (Scargle et al. 2013). But all these rules face the curse of dimensionality as the number of observables becomes high. Alternatives to binning for density estimation can also be used and are discussed in Sect. 9.

In our specific case, the dimensionality of the observable space is determined by the number p of photometric and size parameters in every passband extracted from the survey images. We use ten bins of constant width per dimension throughout the article. More bins per dimension would lead to memory issues caused by the limitations of our computing cluster in the applications that we propose in Sect. 7.6. The bin width for dimension $k \in [1, p]$ in this scheme is therefore given by:

$$W_k = \frac{\max(X_{{\rm w},k}) - \min(X_{{\rm w},k})}{10}, \tag{19}$$

where $X_{{\rm w},k}$ is the pre-processed observables matrix for the observed data.

In this pipeline, the binning pattern is only computed once and for the observed data only. The same binning is then directly applied to the simulated data to ensure better execution speed and comparability between histograms. Because the number of counts per bin is directly affected by the model parameters that rule the number density of galaxies, such as ϕ^* in our application (see Sect. 7), the resulting p-dimensional histograms are not normalized to prevent a loss of information in the minimization of distance between the synthetic and observed data.

4. Comparison between simulated and observed data

Estimating the discrepancy between the observed and simulated binned datasets in high-dimensional space is highly non-trivial, as the choice of a good distance metric is problem dependent. The observables' distributions may be multimodal and skewed, and many metrics rely on the assumption of normality. Others, such as the Kullback-Leibler divergence (Kullback & Leibler 1951) or the Jensen-Shannon distance (Lin 1991), cannot be used without estimating an analytical underlying PDF, which can be very computationally expensive in a high-dimensional observable space.

Here is a non-exhaustive list of non-parametric (i.e., distribution-free) distance metrics found in the literature that can be used on multivariate data in the ABC framework. A more complete review is available in Pardo & Menéndez (2006) and Palombo (2011); however, no study to quantify their relative power has been performed so far. These metrics include:

- The χ^2 test (Chardy et al. 1976) is a simple and widely used way of determining whether observed frequencies are significantly different from expected frequencies. The main drawback of this approach is that χ^2 test results are dependent on the binning choice (Aslan & Zech 2002). For example, Kurinsky & Sajina (2014) use the χ^2 distance to compare color-color histograms.
- The Kolmogorov-Smirnov (KS) test (Chakravarti et al. 1967) estimates the maximum absolute difference between the empirical distribution functions (EDF) of two samples. A generalization of this test for multivariate data has been proposed (Justel et al. 1997). However, as there is no unique way of ordering data points to compute a distance between two EDF, it is not as reliable as the one-dimensional version without the help of resampling methods such as bootstraping (Babu & Feigelson 2006).
- The Anderson-Darling (AD) test (Stephens 1974) is a modification of the KS test. This method uses a weight function that gives more weight to the tails of the distributions. It is therefore considered more sensitive than the KS test, but it also suffers from the same problems in the multivariate case.
- The Mahalanobis distance (Mahalanobis 1936) is similar to the Euclidean norm but has the advantage of taking into account the correlation structure of multivariate data. The Mahalanobis statistics, coupled with an univariate KS test, are used by Akeret et al. (2015) to compare photometric parameters for cosmological purposes. However, this distance only works for unimodal data distributions.
- The Bhattacharyya distance (Bhattacharyya 1946) is related to the Bhattacharyya coefficient, which measures the quantity of overlap between the two samples. It is considered more reliable than the Mahalanobis distance in the sense that its use is not limited to cases where the standard deviations of the distributions are identical.
- The Earth Mover's distance (EMD; Rubner et al. 1998) is based on a solution to the Transportation problem. The distributions are represented by a user-defined set of clusters

called signatures, where each cluster is described by its mean and by the fraction of the distribution encapsulated by it. The EMD is defined as the minimum cost of turning one signature into the other, the cost being linked to the distance between the two. A computationally fast approximate version of this distance using the Hilbert space-filling curve can be found in Bernton et al. (2017).

In the present article, we place ourselves within the pBIL framework to perform the inference process. In this context, the binning structure constructed in Sect. 3.5 and the assumption of a Poisson behavior of the number counts in each bin represent the auxiliary model that describes the data. The "auxiliary likelihood" derived from this structure is inspired from the maximum likelihood scheme of Cash (1979), a likelihood that has been used in previous studies like Robin et al. (2014), Bienayme et al. (1987), or Adye (1998):

$$\ln L = \sum_{i=1}^{b} (o_i \ln(s_i) - s_i) \tag{20}$$

where b is the total number of bins, s_i is the number count in bin i for the simulated data, and o_i is the number count in bin i for the observed data. The underlying assumptions for this choice of auxiliary likelihood can be found in Appendix A.

In that scheme, as the logarithm of s_i is used, empty bins cause a problem. In order to avoid singularities, a constant small value (that we set to 1) is added to every bin up to the edges of the observables space. This process is done in both modeled and observed data so that it does not bias our results.

5. Sampling procedure: Adaptive Proposal algorithm

Initialize parameters $\theta_{(0)}$ from prior distribution;
Initialize covariance matrix and temperature;
for $t = 0$ *to* T **do**
  Every S iterations:
    Update covariance matrix and temperature;
  Propose new state θ^* from proposal distribution;
  while θ^* *is outside the prior bounds* **do**
    Propose another state
  end
  Compute $\ln P(\theta^*|\mathcal{D})$ from proposed state (Eq. (24))
  if $\ln P(\theta^*|\mathcal{D}) \geq \ln P(\theta_{(t)}|\mathcal{D})$ **then**
    Accept the jump
  else
    Compute acceptance probability a;
    Draw uniformly distributed random number R_N in the interval $[0, 1]$;
    if $R_N < a$ **then**
      Accept the jump
    else
      Refuse the jump
    end
  end
end

Algorithm 2: Proposed sampling algorithm based on the AP algorithm (Haario et al. 1999).

MCMC methods are a set of iterative processes which perform a random walk in the parameter space to approximate the posterior distribution with the help of Markov chains. A Markov chain is a sequence of random variables $\{\theta_{(0)}, \theta_{(1)}, \theta_{(2)}, ..., \}$ in the parameter space (called states) that verifies the Markov property: the conditional distribution of $\theta_{(t+1)}$ given $\{\theta_{(0)}, ..., \theta_{(t)}\}$ (called transition probability or kernel) only depends on $\theta_{(t)}$. In other words, the probability distribution of the next state only depends on the current state.

After a period (whose length depends on the starting point and the random path taken by the chain) where the chain travels from low to high probability regions of the parameter space, the MCMC samples ultimately converge to a stationary distribution in such a way that the density of samples is proportional to the posterior PDF, also called target distribution. The portion of the chain which is not representative of the target distribution (i.e., the first iterations where the chain has not yet reached stationarity) is called burn-in, and is usually discarded from the analysis a posteriori. Well optimized MCMC methods provide an efficient tool to avoid wasting a lot of computing time sampling regions of very low probability. There is a great variety of MCMC algorithms, and the choice of a specific algorithm is problem-dependent. The reader is referred to Roberts & Rosenthal (2009) for a complete review of these methods.

To estimate the posterior distribution $P(\theta|D)$ defined in Eq. (1) in a reasonable amount of time, one must explore the parameter space in a fast and efficient way. For our purposes, we designed a custom sampling procedure, described in Algorithm 2, based on the MCMC Adaptive Proposal (AP) algorithm (Haario et al. 1999), which is itself built upon the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970). The Metropolis-Hastings algorithm is one of the most general MCMC methods. In this algorithm, given a state $\theta_{(t)}$ sampled from the target distribution $P(\theta)$, a proposed state θ^* is generated using a user-defined transition kernel $Q(\theta^*|\theta_{(t)})$, which represents the probability of moving from $\theta_{(t)}$ to θ^*. The proposition is accepted with probability:

$$a = \min\left\{\frac{P(\theta^*)}{P(\theta_{(t)})}\frac{Q(\theta_{(t)}|\theta^*)}{Q(\theta^*|\theta_{(t)})}, 1\right\}. \tag{21}$$

If the proposed sample is accepted, then $\theta_{(t+1)} = \theta^*$ and the chain jumps to the new state. Otherwise, $\theta_{(t+1)} = \theta_{(t)}$.

The choice of the transition kernel $Q(\theta^*|\theta_{(t)})$ is crucial to guarantee the rapid convergence of the chain. We opt for the popular choice of a multivariate Normal distribution $\mathcal{N}(0, \Sigma)$ centered on the current state and with a covariance matrix Σ which determines the size and orientation of the jumps, so that:

$$\theta^* = \theta_{(t)} + \zeta_{(t+1)}, \tag{22}$$

where $\zeta_{(t+1)}$ follows $\mathcal{N}(0, \Sigma)$.

A good way to assess convergence speed is to monitor the acceptance rate, that is, the fraction of accepted samples over previous iterations. The acceptance rate is mainly influenced by the covariance matrix of the transition kernel Σ. If the jump sizes are too high, the acceptance rate is too low, and the chain stays still for a large number of iterations. If the jump sizes are too small, the acceptance rate is very high but the chain needs a high number of iterations to move from one region of the parameter space to another. These situations are illustrated in Fig. 7. The desired acceptance rate depends on the target distribution, and there is no universal criterion for its optimization, but Roberts et al. (1997) proved that for any d-dimensional target distributions (with $d \geq 5$) with independent and identically distributed (i.i.d.) components, optimal performance of the Random Walk

Metropolis algorithm is attained for an asymptotic acceptance rate of 0.234.

As the modeling process is very time-consuming and the dimensionality of the problem may be high, we cannot afford to rely on trial and error to find the roughly optimal covariance matrix. We therefore opt for an adaptive MCMC scheme to limit user intervention as much as possible and achieve fast convergence. In the AP algorithm proposed by Haario et al. (1999), the covariance matrix of the Gaussian kernel Σ is tuned on-the-fly every fixed number of iterations using previously sampled states of the chain, and it therefore "learns" the target distribution covariance matrix. In our custom version of the algorithm, every S iterations the empirical covariance matrix from every different accepted state of the $N_{\rm last}$ iterations is computed. We then add a fixed diagonal matrix with elements very small relative to the empirical covariance matrix elements, set to 10^{-6}, to prevent it from becoming singular (Haario et al. 2001) while not impacting the results much (but to which extent remains presently an open question). The choice of S, also called the update frequency, is left to the user and weakly influences the performance of the algorithm, so we set it arbitrarily to 500. As for $N_{\rm last}$, we set it to 50 in order to minimize the chance of the covariance matrix being strongly influenced by a potential rapid evolution of the last few states.

In order to be able to converge in any case, a Markov chain must be ergodic. A stochastic process is said to be ergodic if its statistical properties can be retrieved by a finite random sample of the process. It is well known that adaptation can perturb ergodicity (see, e.g., Andrieu & Moulines 2006). In order to ensure that an adaptive sampling algorithm has the right ergodic properties, and hence converges to the right distribution, it must verify the Vanishing Adaption condition: the level of adaption must asymptotically depend less and less on previous states of the chain. Haario et al. (1999) showed that the AP algorithm is not ergodic in most cases. To tackle this issue, Haario et al. (2001) later released a revised version of their algorithm: the Adaptive Metropolis (AM) algorithm. In the AM algorithm, instead of using a fixed number of previous states, the proposal distribution covariance matrix is computed using all the previous states, which solves the ergodicity problem of the AP algorithm. However, we show in Sect. 7 that our custom implementation of the AP algorithm still yields robust results to our problem.

5.1. Prior

In any Bayesian inference problem, the choice of the prior distribution $P(\theta)$ is of crucial importance, because different prior choices can result in different posterior distributions from the same data. Without any information on what parameter values most probably explain our data, our choice by default is that of an uninformative prior, that is, a multivariate continuous uniform distribution whose boundaries are chosen according to the limits currently given for each parameter in the literature. The uniform prior is defined as:

$$P(\theta) = \begin{cases} \prod_{i=1}^{N_p} \frac{1}{d_i - c_i} & \text{if } d_i \leq \theta_i \leq c_i \ \forall i \in [1, N_p] \\ 0 & \text{else,} \end{cases} \tag{23}$$

where c_i and d_i are the lower and upper limit of the PDF for parameter i and $\theta = (\theta_1, \theta_2, ..., \theta_{N_p})$ is the parameter values vector.

If more precise information is available on a given subset of parameters, a convolution with a more informative PDF (e.g., Normal, Beta, ...) can be performed, but in any case a finite interval is needed in order to provide the source generation software with realistic input parameters. In fact, an infinite interval can result in situations in which no galaxies are generated by the model, or conversely when too many galaxies are generated, which would dramatically increase the computing time.

5.2. Acceptance probability

In practice, one uses the ratio of the posterior density at the proposed and current states to measure the acceptance probability. More specifically, we use the difference between the log of these quantities in order to avoid floating-point numbers precision problems when dealing with very small probabilities. In log probability space, Bayes' theorem (cf. Eq. (1)) becomes:

$$\ln P(\theta|\mathcal{D}) \propto \ln P(\theta) + \ln P(\mathcal{D}|\theta), \tag{24}$$

where $\mathcal{D}$ is the input data, $P(\theta|\mathcal{D})$ is the posterior, $P(\theta)$ is the prior defined in Sect. 5.1, and $P(\mathcal{D}|\theta)$ is the auxiliary likelihood defined in Eq. (20).

The target distribution can have a complex shape and if no particular precaution is taken, our sampling algorithm is not immune to getting stuck in a local maximum of likelihood. To tackle this issue, Kirkpatrick (1983) exploited the analogy between the way a heated metal cools and the search for a global optimum of a function. In the so-called simulated annealing algorithm, the acceptance probability a depends on a "temperature" parameter τ, initialized at high value and slowly decreasing over the iterations. In this scheme, the higher the temperature, the higher the algorithm is prone to accept large moves and to get away from a nearby local maximum:

$$a = \begin{cases} \exp - \frac{\ln P(\theta_{(t)}|\mathcal{D}) - \ln P(\theta^*|\mathcal{D})}{\tau} & \text{if } \ln P(\theta^*|\mathcal{D}) - \ln P(\theta_{(t)}|\mathcal{D}) < 0 \\ 1 & \text{if } \ln P(\theta^*|\mathcal{D}) - \ln P(\theta_{(t)}|\mathcal{D}) \geq 0 \end{cases} \tag{25}$$

where $\ln P(\theta_{(t)}|\mathcal{D})$ and $\ln P(\theta^*|\mathcal{D})$ are respectively the log of the posterior density at the current (i.e., at iteration t) and the proposed state. In other words, if a proposition is considered more probable, it is accepted. Otherwise, it is accepted with probability a (defined in Eq. (25)). To perform the latter operation in practice, a uniformly distributed random number R_N is drawn in the interval $[0, 1]$. If $R_N < a$, the jump is accepted. As expected, for $\tau = 1$, the acceptance probability is the same as that of the Metropolis-Hastings algorithm in Eq. (21) for the particular case of a symmetric proposal distribution, that is, when $Q(\theta_{(t)}|\theta^*) = Q(\theta^*|\theta_{(t)})$.

Because of the intrinsic stochasticity of our model, many realizations of the model at the same state $\theta_{(t)}$ can lead to many $\ln P(\theta_{(t)}|\mathcal{D})$ values. Therefore, artificial local maxima of the target distribution appear, because each iteration relies on a single realization of the model. The simulated annealing algorithm was designed to find the global maximum of the target distribution without knowing the posteriori distribution, and this requires us to lower τ in a user-defined scheme. But our goal is distinct as we need to freely explore the parameter space landscape in order to estimate the full posterior distribution. The main constraint for τ is to be comparable to the posterior density difference resulting from the jump. Here we define it as the root mean square (rms) of the current state, as suggested by Mehrotra et al. (1997). In that scheme, a high noise level or a small difference between the proposed and the current state leads to a higher probability of jumping to this state.

The temperature is computed every S iterations by running an empirically-defined number of realizations N_R of the model

at the current state $\theta_{(t)}$, storing every $\ln P(\theta_{(t)}|\mathcal{D})$ value returned in a vector, and computing the standard deviation of the resulting distribution. In the application below, we find that 20 realizations are sufficient to give a reasonable estimate of the rms (cf. Fig. 8) and that the temperature quickly reaches a stationary distribution at a relatively low level $\tau \simeq 30$, after the first few 10^3 iterations (cf. Fig. 9).

5.3. Initialization of the chain

The initial state $\theta_{(0)}$ is drawn randomly from the prior distribution (see Sect. 5.1). The initial position will only affect the speed of convergence, because the final distribution shall not depend on the initial position, if the chain converges. The initial temperature is then computed from this state. As for the proposal distribution, it is initialized so that no direction in the parameter space is preferred by the sampling algorithm at first. The initial covariance matrix is therefore diagonal, whose non-zero elements are set to:

$$C_{ii} = \frac{u_i - l_i}{E} \; \forall i \in [1, N_p], \tag{26}$$

where u_i and l_i are respectively the upper and lower bounds of the prior distribution for parameter i, N_p the number of parameters, and E a value set empirically to 200 in order to ensure reasonable acceptance rates at the beginning of the chain. According to Haario et al. (1999), the adaptive nature of the algorithm implies that the choice of E should not influence the output of the chain.

6. Convergence diagnostics

The goal of an MCMC chain is to reach a stationary distribution that is supposed to be representative of the target distribution. Unfortunately there is no theoretical criterion for convergence: in other words it is impossible from a finite MCMC chain to assess convergence with certainty. Many convergence diagnostics have been developed (the reader can find an extensive review of those and a comparison of their relative performances in, e.g., Cowles & Carlin 1996), but these diagnostics can only tell if a chain has not converged. So in order to have confidence in the convergence of the chains, we must perform multiple diagnostics.

The first check is carried out by visual inspection of the trace plot for each parameter. Trace plots are used to diagnose poor mixing, that is, when the chain is highly autocorrelated, or slow sampling caused by too small a step size, which suggests that the majority of the MCMC output is not representative of the target distribution (see Fig. 7). We also use trace plots to estimate the length of the burn-in phase. The latter is determined by eye, by a rough estimate of the minimum number of iterations D necessary for all the parameters to reach a seemingly stationary distribution. We then discard the D first iterations, where D depends on the chain.

Finally, one of the most popular convergence diagnostics is a test proposed by Gelman & Rubin (1992). Given m chains $\{\theta^j_{(t)}\}$ ($j = 1, ..., m$ and $m \geq 3$, and typically $\sim$10), each of length n after discarding burn-in ($t = 1, ..., n$) and with different starting points, the test compares the variance between the mean values of the m chains B and the mean of the m within-chain variances W:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}^j_{\cdot} - \bar{\theta}_{\cdot\cdot})^2, \tag{27}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} \left[\frac{1}{n-1} \sum_{i=1}^{n} (\theta^j_{(i)} - \bar{\theta}^j_{\cdot})^2 \right], \tag{28}$$

where $\bar{\theta}^j_{\cdot} = \frac{1}{n} \sum_{t=1}^{n} \theta^j_{(t)}$ is the mean value of chain j, and $\bar{\theta}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}^j_{\cdot}$ is the average value over the m chains.

An overestimate of the true marginal posterior variance is given by the unbiased estimator

$$\hat{V} = \frac{n-1}{n} W + \frac{1+m}{nm} B. \tag{29}$$

Finally convergence is estimated using the potential scale reduction factor (PSRF) $\hat{R}$:

$$\hat{R} = \frac{\hat{V}}{W}. \tag{30}$$

Here we use the Gelman Rubin diagnostic implemented in this form in the PyMC package (Patil et al. 2010) to perform our convergence tests, and we consider that convergence has been reached if $\sqrt{\hat{R}} < 1.1$ for all model parameters (Brooks & Gelman 1998); otherwise, more iterations are performed until the criterion is met.

7. Application to a toy model

As a proof-of-concept of the method, we apply our pipeline to a selection of idealized cases, where the "observed" data is a synthetic image containing one or two populations of galaxies generated by a set of known input parameters of the STUFF model. Our goal is to infer the values of the input parameters in this framework.

7.1. Simulated survey characteristics

As data image, we choose to reproduce a full-sized stack of the CFHTLS Deep field (e.g., Cuillandre & Bertin 2006). The CFHTLS Wide and Deep fields offer carefully calibrated stacks with excellent image quality. Covering 155 deg^2 on the sky in total, the Wide field allows for a detailed study of the large scale distribution of galaxies. As for the Deep field, which covers 4 deg^2 in total, it beneficits from long time exposures (33 to 132 h), which ensure reliable statistical samples of different populations of bright galaxies up to $z \sim 1$. Each stack of the CFHTLS Deep field is a 19 354 × 19 354 pixel image covering 1 deg^2 on the sky. We simulate one stack of the Deep field in three bands: Megacam u and i from the CFHTLS, and the WIRcam K_s infrared channel from the WIRcam Deep Survey (WIRDS) that covers part of the CFHTLS Deep fields. In accordance with CFHTLS product conventions, the image exposure time is normalized to one second and the AB magnitude zero-point is 30. The overall characteristics of the simulated images are summarized in Table 1.

The SKYMAKER PSF model for the CFHTLS image is generated within the software. The *aureole* simulation step is deactivated to speed up the image generation process. For the same reason, we exclude from the STUFF list all galaxies with apparent magnitudes in the reference band brighter than 19 or fainter than 30, in order to avoid simulating both very large and very numerous galaxies. There is no stellar contamination, as STUFF does not yet offer the possibility to simulate realistic star fields.

7.2. Source extraction configuration

SEXTRACTOR is configured according to the prescription of the T0007 CFHTLS release documentation (Hudelot et al. 2012). We use it in double image mode, with the i-band image as the

Table 1. Imaging characteristics of the CFHTLS+WIRDS surveys used for SkyMaker.

Passband	u	i	K_s
Image size [pixels]	19 354 × 19 354	19 354 × 19 354	19 354 × 19 354
Effective gain [e^-/ADU]	74 590	6807	2134
Well capacity [e^-]	∞	∞	∞
Saturation level [ADU]	6465	4230	110 884
Effective read-out noise [e^-]	4.2	4.2	30
Total exposure time [s]	1	1	1
Zero-point magnitude ["ADU/s"]	30	30	30
Effective wavelength [μm]	0.381	0.769	2.146
Sky level [AB mag/arcsec^2]	22.2	20.0	15.4
Seeing FWHM [arcsec]	0.87	0.76	0.73

Table 2. Uniform prior boundaries for the parameters of the luminosity and size functions, and their evolution with redshift.

Parameter	ϕ^*	M^*	α	ϕ_e	M_e	M_{knee}	r_{knee}	γ_b
Lower bound	10^{-7}	–22	–2.5	–3	–2.5	–21	0	–2
Upper bound	10^{-2}	–17	0	2	0	–19	3	0

Notes. All the parameters above are given for $H_0 = 100$ km s^{-1} Mpc^{-3}.

detection image, and the background is estimated and subtracted automatically with a 256 × 256-pixels background mesh size. In order to optimize the detectability of faint extended sources, detection is performed on the images convolved with a 7 × 7 pixels Gaussian mask having a full width at half maximum (FWHM) of three pixels, that approximates the size of the PSF and acts as a matched filter. Finally, the detection threshold is set to 1.2 times the (unfiltered) rms background noise above the local sky level.

In order for the results concerning faint sources near the detection limit not to depend too closely on the details of noise statistics, all negative fluxes and radii are clipped to 0 after extraction.

7.3. Pipeline configuration

We adopt non-informative, uniform priors for the free parameters of all the considered models, with boundaries defined in Table 2. The boundaries are chosen to prevent the pipeline from exploring non physical domains, such as a very steep LF faint end, which leads to an unreasonably high number of generated galaxies and dramatically increases the computing time. We select the least constraining prior possible, which corresponds to a large interval around generally accepted values, such as the values reviewed in de Lapparent et al. (2003) for example.

To perform the dynamic range compression as defined in Sect. 3.3, we need an estimate of the noise level in the conditions of a CFHTLS Deep field. To that end, we use the population of $\sim 10^4$ pure bulge elliptical galaxies described in Sect. 7.6 and apply the recipe described in Sect. 3.3. The resulting parameters for the dynamic range reduction function in the uiK_s filters are summarized in Table 3. For the various cases considered in this article, we use for all galaxy populations the σ_{FLUX_AUTO} and σ_{FLUX_RADIUS} values measured for the elliptical galaxies.

We consider two cases in the following sections: the first contains two types of galaxies, a mix between ellipticals and lenticulars, and late-type spirals, which undergo both luminosity and size evolution. But we limit the inference to the LF shape and evolution parameters for both populations. The second case focuses on a single population of pure bulge ellipticals, but this time the inference is performed on both the LF and the distribution of effective radii (both including the evolution parameters).

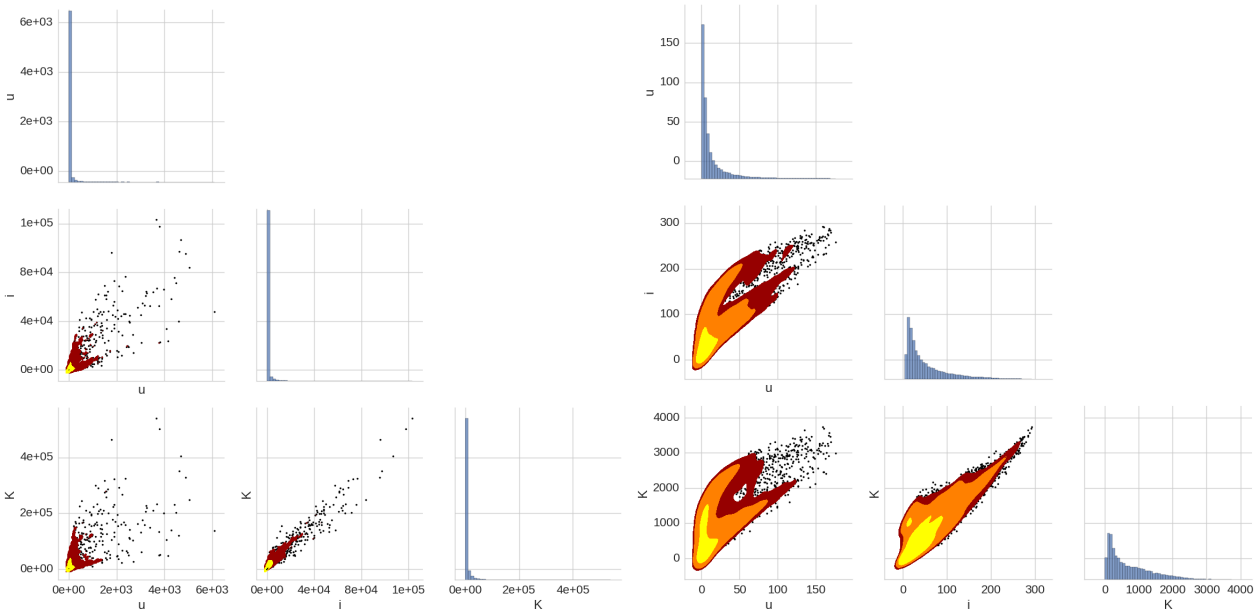
Table 3. Parameters of the dynamic range reduction function used in Eq. (16).

Filter	u	i	K_s
σ_{FLUX_AUTO}	3.4	3.6	54.0
σ_{FLUX_RADIUS}	3.5	2.7	2.6
κ_c	10	10	10

7.4. Multi-type configuration: luminosity evolution

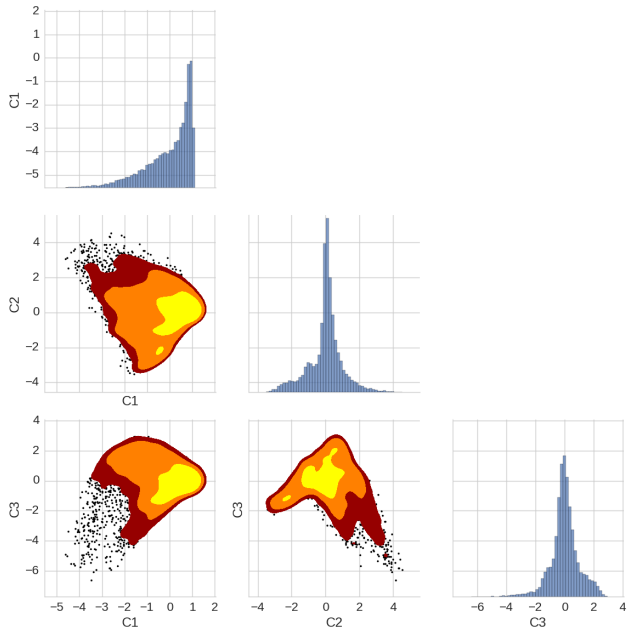
Astronomical survey images contain multiple galaxy populations. We need to emulate this situation in order to test the behavior of our pipeline in realistic conditions. To do so we use as input data a simulated CFHTLS Deep image in uiK_s containing two types of galaxies: a population of early-type galaxies (an average between E ans S0) of morphological type $T = -5$ and a population of late-type spirals (Sp) of morphological type $T = 6$. We rely on published results to define these populations. Using data from SDSS, the 2dF Galaxy Redshift Survey, COMBO-17, and DEEP2, Faber et al. (2007) split their distribution of galaxies into two populations by color, using the rest-frame M_B versus $U - B$ color−magnitude diagram: a blue population and a red population. We use their derived evolving LF parameters to build an E/S0 and Sp populations. The detailed conversion process from the LF parameters of Faber et al. (2007) to the values used in Stuff (which include a magnitude system conversion, a band transformation, and a cosmological correction) is provided in Appendix B. This provides us with values for M^* (LF characteristic magnitude) and the evolution parameters M_e and ϕ_e for both populations.

The B/T ratios in the g adopted reference band are determined using the distribution of B/T in g-band as a function of morphological type from EFIGI (Extraction of Idealized Forms of Galaxies in Image processing) data (Baillard et al. 2011; de Lapparent, priv. comm.). To limit run time, the ϕ^*



(a) Initial

(b) Reduced

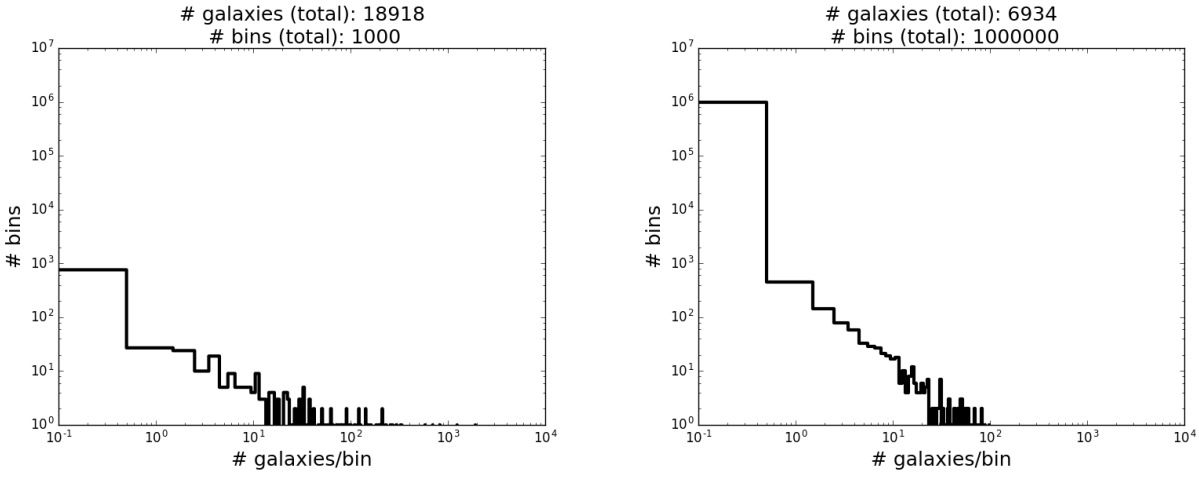


(c) Whitened

Fig. 5. Distribution of observables before and after each step of pre-processing from the mock input data with 2 populations of galaxies (Ellipticals+Spirals) described in Sect. 7.4. The dark red, orange and yellow areas in the contour plots are the regions that enclose 99%, 95% and 68% of the points respectively. *Top left panel*: scatter plot of the FLUX_AUTO of extracted sources (in ADUs) in filters uiK_s and their covariances. *Top right panel*: same plot, but with the dynamic range of the FLUX_AUTO distributions reduced via Eq. (16). *Bottom panel*: same plot, after whitening of the reduced distribution of observables. The latter distribution is uncorrelated, centered on the mean of the distribution and rescaled, allowing for a much more efficient binning process than on raw fluxes, and a more practical comparison with the simulated observables.

values for each population are set to have $\sim 4 \times 10^4$ galaxies in total generated quickly by STUFF in the field area. In this scheme, we have $\sim 10^4$ E/S0, and $\sim 3 \times 10^4$ Sp, which corresponds to a ϕ^* value for each population of ten times lower than the values given by Faber et al. (2007). We indeed do not match the number counts of a CFHTLS Deep field as it would lead to unreasonable computing time: reproducing realistic number counts over a full Deep field would actually imply STUFF generating a number of galaxies one order of magnitude higher for E/S0 and Sp, and also adding a population of $\sim 10^5$ Irr which dominates the number counts fainter than 22 to 24 mag, depending on the filter.

The input parameters used to generate both populations are listed in Table 4. The parameters to infer in this case are the five evolving LF parameters for each of the populations: ϕ^*, M^*, α, ϕ_e, and M_e, that is a total of ten parameters (we do not infer the size distribution and evolution parameters). The observables are the SEXTRACTOR FLUX_AUTO in each of the three passbands, which leads to a three-dimensional observable space. Using ten



(a) Multi-type case (b) Fattening E case

Fig. 6. Histogram of the number of sources extracted per bin for the pre-processed input data of the test cases presented in Sect. 7. In the *left panel*, three observables are considered: the FLUX_AUTO in uiK_s. In the *right panel*, six observables are considered: the FLUX_AUTO in uiK_s and the FLUX_RADIUS in uiK_s. With the binning rule described in Sect. 3.5, between the "Multi-type" case and the "Fatrtening E" case, the number of bins increases by a factor 10^3, and the number of empty bins is increased by roughly the same amount. This illustrates the curse of dimensionality we face in this method, and puts computational limits on the number of observables we can use.

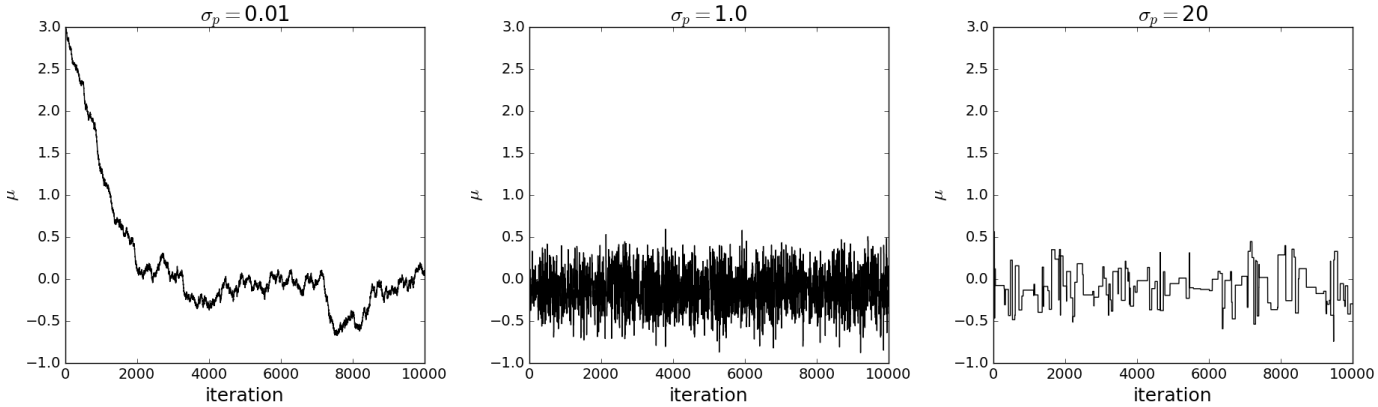


Fig. 7. Traceplots depicting three typical situations that can arise in a standard MCMC chain with the (non-adaptive) Metropolis-Hastings algorithm. The input data is a set of 20 points normally distributed with mean 0 and standard deviation 1. The parameter to infer is the mean μ of the input data distribution. The prior is a Normal distribution with mean 0 and standard deviation 1, and the transition kernel is a Normal distribution centered on the current state and width σ_p. In each case the chain starts from $\mu = 3$ and is run for 10 000 iterations. The target distribution sampled is the same, but the width of the proposal distribution, thatis, the jump size, is different for each case. *Left panel*: the jump size is too small. The burn-in phase is very long and a much longer chain is needed to sample the target distribution. *Central panel*: the jump size is optimal, therefore the target distribution is well sampled. *Right panel*: the jump size is too big. Hence the chain spends a lot of iterations in the same position, which makes the sampling of the target distribution inefficient.

Table 4. Characteristics of the galaxy test populations.

Population	SED_b[a]	SED_d[a]	ϕ^* [h^3 Mpc^{-3}]	M^*	α	ϕ_e	M_e	B/T	T[b]	$\alpha(T)$	Number[c]
Multi-type: E/S0	E	E	0.003	–19.97	–0.5	–1.53	–1.77	0.65	–5	0.0	10 447
Multi-type: Sp	E	Scd	1.4e–4	–19.84	–1.3	0.03	–1.95	0.2	6	1.47	28 281
Fattening E	E	E	0.0035	–19.97	–0.5	–1.53	–1.77	1.0	–5	0.0	11 353

Notes. The LF parameters are given for $H_0 = 100$ km s^{-1} Mpc^{-3}. [(a)] The disk and bulge SEDs are Coleman et al. (1980) templates. [(b)] de Vaucouleurs (1959) revised morphological type. [(c)] Number of sources generated by one realization of STUFF.

bins for each observable as indicated in Sect. 3.5, we obtain a total number of 10^3 bins in the observable space. Over the $\sim 4 \times 10^4$ galaxies generated by STUFF, we find that $\sim 2 \times 10^4$ are extracted with SEXTRACTOR. The number of extracted galaxies per bin is presented in Fig. 6.

7.5. Results of the "multi-type" configuration

We run the pipeline on a hybrid computing cluster of seven machines totaling 152 central processing unit (CPU) cores. We launched three chains in parallel for 18 357, 18 565, and

Table 5. Size parameters for the bulge and disk of each galaxy test population.

Disk				Bulge		
β_d	r_d^* [h^{-1} kpc]	γ_d	σ_λ	M_{knee}	r_{knee} [h^{-1} kpc]	γ_b
–0.214	3.85	–0.80	0.36	–20.0	1.58	–1.00

Notes. All the parameters above are given for $H_0 = 100$ km s^{-1} Mpc^{-3}, and "b" refers to bulge, and "d" to disk.

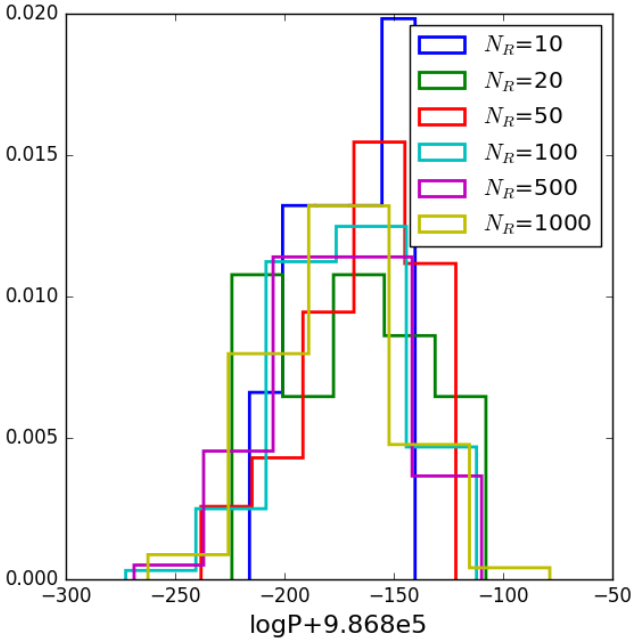


Fig. 8. Normed distribution of ln P for various numbers of realizations N_R of the model. Each distribution is generated in the conditions of the "Fattening E" case, at "true" input values (cf. Table 4) and with the same seed for galaxy generation in STUFF. Standard deviation of the distributions do not appear to differ significantly. We conclude that 20 realizations of the model are enough to characterize the order of magnitude of rms.

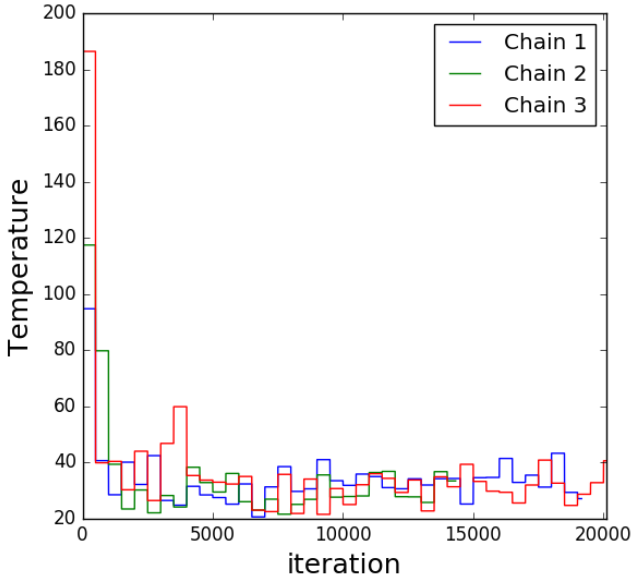


Fig. 9. Temperature evolution with the number of iterations of the MCMC process in the "Fattening E" case described in Sect. 7.6. Here the temperature is computed every 500 iterations at current state with 20 realizations of the model. We note that for each chain, the temperature values quickly converge to the level of noise of the model near input values.

16 211 iterations respectively, with randomly distributed starting points, using 50 400 CPU hours in total. The burn-in phase is estimated by visual examination of the trace plot. All the iterations before the upper and lower envelope of the trace becomes constant for all the chains and for all the parameters simultaneously are discarded as burn-in, which in the case under study corresponds to the first 10^4 iterations. Then convergence over the f last iterations of each chain is assessed based on the Gelman-Rubin test (cf. Table 6), where f is the minimum length over the three chains after burn-in, as the convergence test requires the same number of iterations for all the chains: $f = 6211$ iterations. Table 6 lists the results of the Gelman-Rubin test, which suggest that all the chains have converged to the same stationary distribution.

The final joint posterior distribution is the result of the combined accepted states of all the chains run after burn-in. The posterior PDF plot is shown in Fig. 10: it contains 3017 accepted iterations out of 23 132 propositions, corresponding to an overall 13% acceptance rate after burn-in. The graph shows that the "true" input values all lie within the 68% credible region, which in Bayesian terms means that there is a 68% probability that the model value falls within the credible region, given the data. Summary statistics of the posterior PDF are listed in Table 7. As the pipeline generates constraints that are consistent with the input parameters, we therefore conclude that our approach can be used to perform unbiased inference on the photometric parameters of galaxies using two broad classes of galaxy types given non-informative priors.

Moreover, we find in Fig. 10 some strong correlations or anti-correlations between various pairs of parameters, that are symptomatic of the degeneracies in the parameters for our specific set of observables (fluxes). For example, a strong anti-correlation is found between M^* and M_e in the two populations. This can be explained by the fact that a brighter (lower) M^* population at $z = 0$ can be partly compensated by a shallower (higher) redshift evolution.

7.6. Fattening ellipticals: size and luminosity evolution

We then test whether our pipeline can also infer the characteristic size evolution of galaxies. Because of memory limitations, we perform this test in a simplified framework. We use as input data a CFHTLS image in uiK_s containing $\sim 10^4$ E/S0 (pure bulge) galaxies generated with STUFF. The input photometric parameters are listed in Table 4 and those for bulge size are listed in Table 5. The parameters to infer are the five evolving LF parameters, as well as three parameters governing the bulge distribution and evolution: M_{knee}, r_{knee}, and γ_b (as defined in Sect. 7.4). That is a total of eight parameters. No extinction is included in this case. As the size evolution parameters cannot be retrieved with the photometric information only (FLUX_AUTO), the

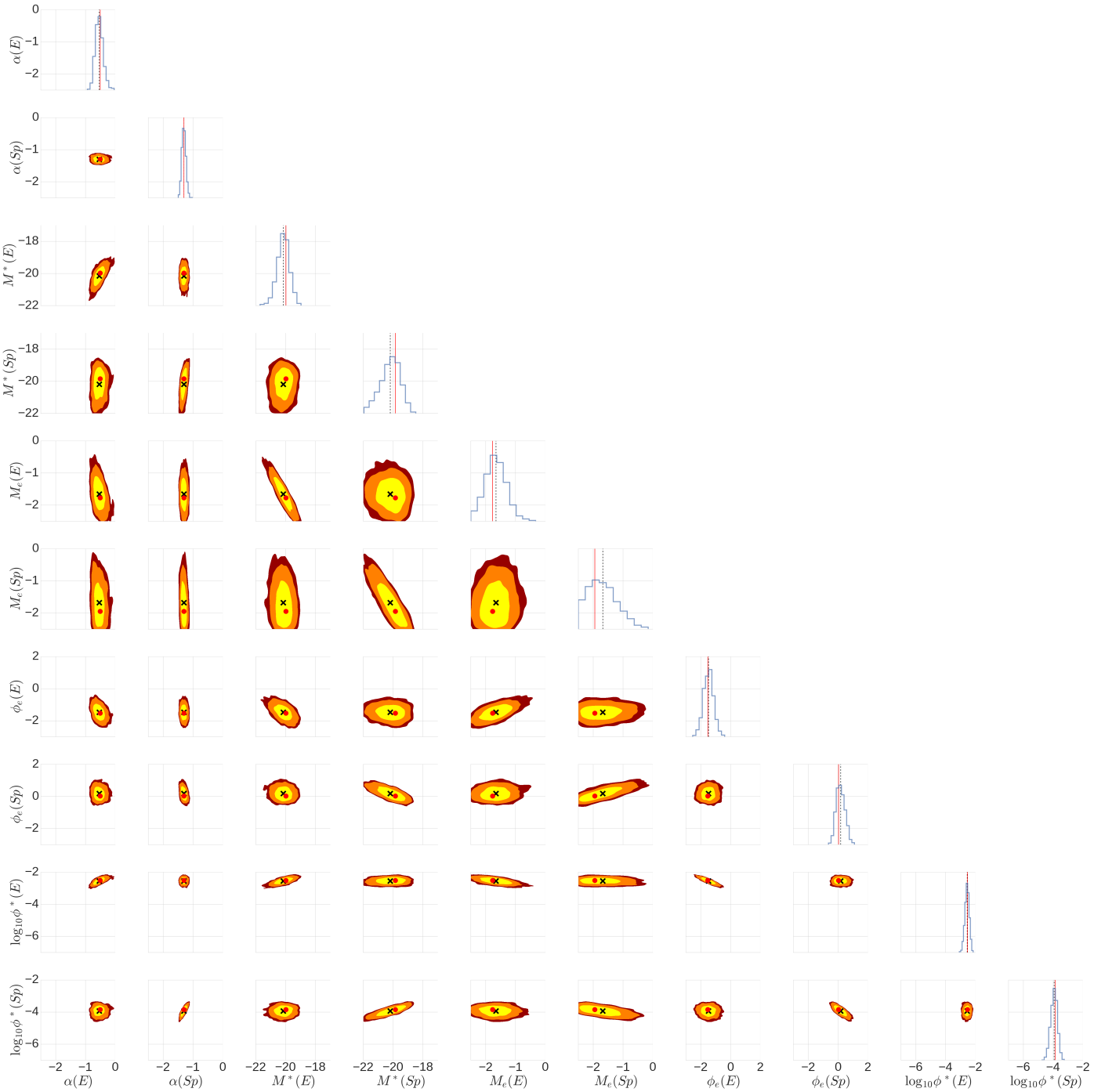


Fig. 10. Joint posterior distribution resulting from the "Multi-type" test described in Sect. 7.4. The diagonal plots show the marginal distribution for each parameter (the projection of the posterior onto that parameter). Each panel is bounded by the prior range values. The dark red, orange, and yellow areas in the contour plots represent the 99%, 95%, and 68% credible regions respectively. The black crosses and red dots are the mean of the posterior and input true value respectively. In the marginalized posterior panels, the black dotted and red lines represent the posterior mean and the true value respectively.

Table 6. Results of the Gelman-Rubin test.

Population	$\log_{10}(\phi^*)$	M^*	α	ϕ_e	M_e	M_{knee}	r_{knee} [h^{-1} kpc]	γ_b
Multi-type: E/S0	1.015	1.006	1.014	1.012	1.005	∅	∅	∅
Multi-type: Sp	1.020	1.013	1.028	1.007	1.012	∅	∅	∅
Fattening E	1.013	1.003	1.020	1.003	1.001	1.008	1.010	1.008

Notes. The values of $\sqrt{R}$ are obtained using 3 chains for each case, whose burn-in phase for each chain is determined by eye. All values are <1.1, which is a hint that in each case, all the chains have converged to the same distribution. The parameters above are given for $H_0 = 100$ km s^{-1} Mpc^{-3}.

Table 7. Summary statistics on the marginalized posterior distributions for the galaxy test populations and comparison with the input values.

Population	Parameters	Input value	Mean	MAP[a]	68% interval	95% interval	99% interval
Multi-type: E/S0	$\log_{10}(\phi^*)$	–2.52	–2.56	–2.57	[–2.72, –2.40]	[–2.85, –2.29]	[–2.95, –2.19]
	M^*	–19.97	–20.15	–20.12	[–20.54, –19.74]	[–20.95, –19.38]	[–21.20, –19.08]
	α	–0.5	–0.53	–0.54	[–0.68, –0.43]	[–0.77, –0.24]	[–0.82, –0.06]
	ϕ_e	–1.53	–1.48	–1.37	[–1.82, –1.15]	[–2.16, –0.81]	[–2.35, –0.49]
	M_e	–1.77	–1.66	–1.70	[–1.99, –1.29]	[–2.36, –0.96]	[–2.46, –0.75]
Multi-type: Sp	$\log_{10}(\phi^*)$	–3.85	–3.93	–3.96	[–4.15, –3.67]	[–4.35, –3.54]	[–4.53, –3.40]
	M^*	–19.84	–20.18	–19.81	[–20.68, –19.43]	[–21.59, –19.01]	[–21.93, –18.84]
	α	–1.3	–1.30	–1.33	[–1.37, –1.25]	[–1.42, –1.17]	[–1.44, –1.13]
	ϕ_e	0.03	0.18	0.15	[–0.16, 0.47]	[–0.34, 0.75]	[–0.48, 0.94]
	M_e	–1.95	–1.68	–1.81	[–2.39, –1.29]	[–2.49, –0.87]	[–2.49, –0.19]
Fattening E	$\log_{10}(\phi^*)$	–2.46	–2.46	–2.51	[–2.60, –2.31]	[–2.75, –2.18]	[–2.84, –2.08]
	M^*	–19.97	–20.04	–20.08	[–20.41, –19.63]	[–20.84, –19.23]	[–21.15, –19.12]
	α	–0.5	–0.49	–0.51	[–0.62, –0.40]	[–0.72, –0.25]	[–0.78, –0.13]
	ϕ_e	–1.53	–1.49	–1.56	[–1.84, –1.09]	[–2.22, –0.74]	[–2.41, –0.49]
	M_e	–1.77	–1.65	–1.61	[–2.04, –1.29]	[–2.35, –0.96]	[–2.47, –0.73]
	M_{knee}	–20.00	–20.10	–19.99	[–20.32, –19.79]	[–20.68, –19.58]	[–20.90, –19.46]
	r_{knee}	1.58	1.64	1.56	[1.47, 1.77]	[1.34, 1.97]	[1.27, 2.09]
	γ_b	–1.00	–1.03	–1.07	[–1.18, –0.87]	[–1.31, –0.71]	[–1.50, –0.65]

Notes. [(a)] Maximum A Posteriori. The LF parameters are given for $H_0 = 100$ km s^{-1} Mpc^{-3}.

FLUX_RADIUS parameters of SExtractor for all galaxies in each passband are added to the observables space. This leads to a six-dimensional observable space. Over the $\sim 10^4$ E generated by Stuff, we find that $\sim 7 \times 10^3$ are found by SExtractor. With ten bins as indicated in Sect. 3.5, this results in a total number of bins of 10^6. The number of extracted galaxies per bin is presented in Fig. 6.

7.7. Results of the "fattening ellipticals" configuration

We run our pipeline with three chains in parallel for 18 898, 14 056, and 20 110 iterations respectively, with uniformly distributed starting points, using 19 656 CPU hours in total. The first 10^4 iterations of each chain are discarded as burn-in. Convergence is reached over the $f = 4323$ last iterations of each chain, as assessed by the Gelman-Rubin test results displayed in Table 6. The resulting posterior distribution is shown in Fig. 11. It contains 6287 accepted iterations over 38 064, which leads to an acceptance rate of 16.5%.

Each marginalized posterior plot exhibits a main mode, with the peak and the mean almost indistinguishable from the input values. The joint posterior distribution shows that the input values all fall within the 68% credible region. Summary statistics of the posterior PDF are listed in Table 7. Here again, our pipeline produces constraints that are consistent with the true parameters. So we conclude that our pBIL method can reliably infer the luminosity and size distribution of one population of galaxies without any systematic bias.

The joint posterior PDF also reveals covariances between parameters. For instance, the ϕ^* and ϕ_e parameters are naturally anti-correlated because an increase of ϕ^* (at $z = 0$) can partially be compensated by a steeper decrease of the normalization with redshift, hence a smaller value of ϕ_e.

8. Comparison with SED fitting

As demonstrated above, our pBIL method is efficient at recovering the input parameters used to define the luminosity and size evolutions in the mock CFHTLS image. One may wonder how it compares with the classical, less CPU-expensive method for measuring LFs – SED fitting – which provides, from a multi-band photometric catalog, estimates of the photometric redshifts as well as rest-frame luminosities. Luminosity functions can then be derived using independent redshift bins.

The simulated field used for this comparison is the "Fattening E" sample, with a single population of pure bulges with the Coleman et al. (1980) "E" template. The Z-PEG code (Le Borgne & Rocca-Volmerange 2002) is applied to the u, i, and K_s photometric catalog obtained with SExtractor in the same configuration as described for the pBIL method in Sect. 7. Photometric redshifts are measured together with g-band luminosities for every i-band detected object down to $u_{AB} = 30$. The fits were performed using the whole range of SED templates from Coleman et al. (1980), from E to Irr galaxies. The discrete LFs obtained in each redshift bin were volume weighted with a V_{max} correction at the faint magnitude bins, and a Schechter (1976) function was fitted to the data independently in each redshift bin with a Levenberg-Marquardt algorithm with Φ^*, M^*, and α as free parameters.

Comparison of the evolution with redshift of the LF parameters between the pBIL approach (green dashed line for mean of posterior, and 68% light green shaded region) and the results from SED fitting (red symbols with error bars) are shown in Fig. 12. As expected, they both roughly follow the trends set by the evolution of the input parameters (blue solid line), with some offsets that can be explained by the fact that SED fitting is done on only three photometric bands. This is clearly a major limiting factor, albeit partly compensated by the choice of the SED templates: the input SED and the templates share a common SED (the "E" SED, even if all SEDs from Coleman et al. 1980 are also used for the SED fitting). We believe that this choice is fair because in the pBIL method, the same set of SEDs was also used for data generation and for the inference of LF parameters.

The significant systematic offset in α from SED fitting compared to the input and pBIL curves in Fig. 12 shows that the faint-end slope parameters α is poorly estimated, with a

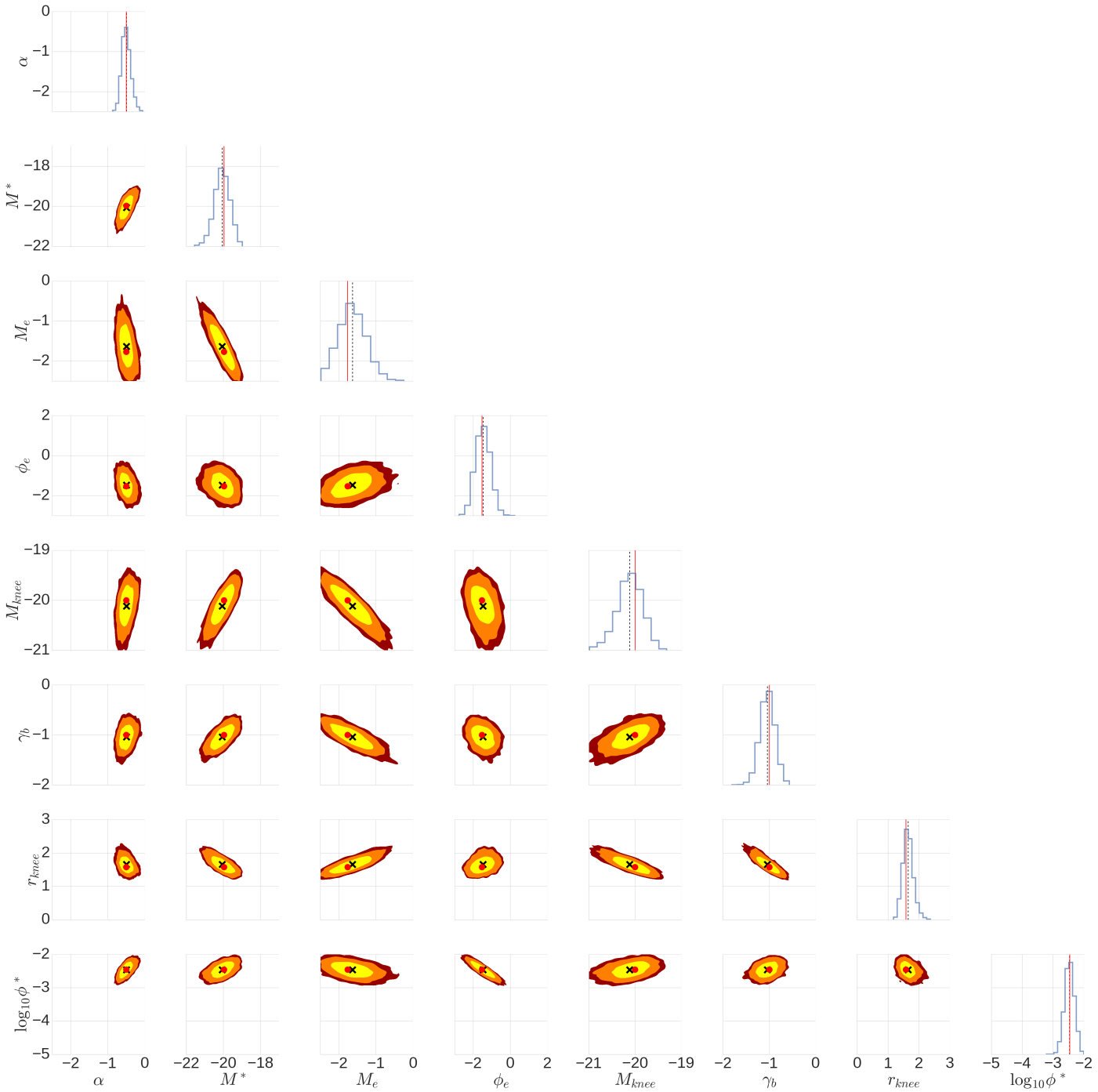


Fig. 11. Joint posterior distribution resulting from the "Fattening E" test described in Sect. 7.6. The diagonal plots show the marginal distribution for each parameter. Each panel is bounded by the prior range values. The dark red, orange, and yellow areas in the contour plots represent the 99%, 95%, and 68% credible regions respectively. The black crosses and red dots are the mean of the posterior and input true value respectively. In the marginalized posterior panels, the black dotted and red lines represent the posterior mean and the true value respectively.

significant systematic offset at $z \geq 0.7$. This is caused by the negative input slope (see Table 4), which yields few faint galaxies in the sample. Moreover, because of numerous catastrophic outliers in the photometric redshifts (caused by the u, i, K_s-only photometric catalog), there is a mismatch between the true redshift of many faint objects and the redshift bin to which they are assigned. This leads to an underestimate of the error bars on the individual points.

For this comparison of the LF parameters between the two approaches, we had to derive the envelop of the LF parameters as a function of redshift for the trace elements of the MCMC chains within the 68% credible region of the parameters space. Of course, the area appears in Fig. 12 as much smoother than the individual points derived from SED fitting because the chosen LF model for the inference evolves smoothly with redshift. Still, it is remarkable that the region is tight and almost centered on the

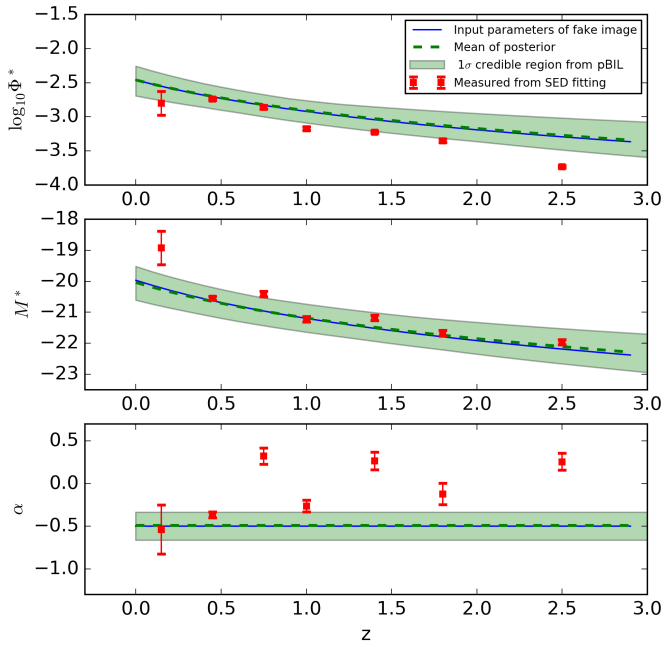


Fig. 12. Evolution of the LF parameters as defined in the mock data image (blue solid line) and inferred from the pBIL method in the "Fattening E" population described in Sect. 7.6 (the green dotted line is the mean of the posterior and the shaded area represents the 68% credible region), compared to the direct measurement of the LF obtained per redshift bin and estimated using a $V_{\rm max}$ weighting, after determination of the photometric redshifts from SED fitting (red dots).

values of the true parameters at all redshifts. This is because the various covariances between the five LF parameters of the model tend to narrow down the shaded areas in these graphs, therefore implying that galaxies at all redshifts in the images contribute to constrain the parameters of the model in the pBIL approach.

9. Discussion

One issue of concern in the posterior distributions that we derived with our pipeline (Figs. 10 and 11) is illustrated by the fact that in Fig. 12, the 68% shaded region is large compared to the distance between the input parameters (blue solid line) and the mean of the posterior (green dashed line), which are almost indistinguishable in all three graphs. We suspect that this results from an enlargement of the posterior because of the "temperature" term that we use in order to circumvent the stochastic nature of each model realization (see Sect. 5.2). In essence, the model's stochasticity itself (galaxies are randomly drawn from the distribution functions) inevitably contributes to the uncertainty in the posterior. We have no quantitative estimate of this enlargement and we suspect it might be a limiting factor on the precision of the parameter inference. Estimating this enlargement from simulated data would have required us to generate a very large number of realizations for each step of the chain (hence we could have turned off the "temperature" term). This would, however, be prohibitive in computing time, even in the considered simple tests performed in this article. We note that using surveys with large statistics in the number of characteristic population of galaxies is, as always, preferable, and should limit this bias.

Moreover, there is room for several technical improvements of our pipeline, in order to guarantee a faster convergence and a more accurate inference:

- As implemented in the present article, our method faces the inevitable curse of dimensionality. In fact, as we bin each observable over ten intervals, for each observable added the hyper-dimensional number of bin increases by one order of magnitude. This limits our approach to a restricted number of observables in order to prevent memory errors. In order to adapt this method to higher numbers of observables, we may have to change our strategy and bin projections of the datasets instead of binning the complete observable space, with the drawback of losing mutual information.
- Instead of binning the distribution of observables, whose results depend on the bin edges and bin width, a more reliable method for density estimation for multivariate data is Kernel Density Estimation (KDE). KDE transforms the data points into a smooth density field, and alleviates the dependence of the results on the bin edges by centering a unimodal function with mean 0, the kernel, on each data point. In practice, KDE is more computationally expensive than binning, and also requires some level of hand tuning, in the form of the right kernel function and the optimal bandwidth, which in KDE is the analog of bin width in histograms.
- The mean runtime of an MCMC chain in the context of the test cases described in Sect. 7 is approximately two weeks. Up to 50% of this runtime is currently spent in job scheduling latencies for each iteration (as shown in Fig. 4). A more integrated approach, based, for example, on Message Passing Interface (MPI) and operating only in memory might reduce those latencies. The next step would be to increase computational efficiency by offloading the most time-consuming image rendering and source extraction tasks to graphics processing units (GPUs), especially convolutions and rasterizations.
- We emphasize that on the order of 10^4 iterations is needed to attain convergence in the test cases studied. Considering the high computational cost of this approach, one may wonder how to attain faster convergence in realistic frameworks. In that regard, Gutmann & Corander (2016), who explored the computational difficulties arising from likelihood-free inference methods, proposed a strategy based on probabilistic modeling of the relation between the input parameters and the difference between observed and synthetic data. This approach would theoretically reduce the number of iterations needed to perform the inference.

Finally, more realistic mock astrophysical images are required before running our pipeline on real survey data:

- The addition of a likely stellar field to the simulated images would contaminate the source extraction process in a realistic way. This could be done via the use of photometric catalogs from real or simulated stellar surveys (e.g., Gaia Collaboration 2016; Robin et al. 2012).
- It is now well known that the contribution of clustering and environmental effects influence the colors (e.g., Madgwick et al. 2003; Bamford et al. 2009) and spectral types (Zehavi et al. 2002) of galaxies: red and quiescent galaxies are mostly distributed in regions of high density, such as the centers of clusters, whereas blue and star forming galaxies are less clustered. Galaxy clustering also has an impact on source blending and confusion. These effects are not implemented in STUFF, and this might bias our results in a way that is difficult to estimate. In order to limit this effect, one could select the areas of the analyzed survey that contain only field galaxies and use these areas as input data.
- The present application uses as a reference the CFHTLS Deep survey, which sensitivity is very homogeneous over the field of view. This is, however, not the case for many surveys.

A more general application of the method would require simulating each individual raw survey exposure, and performing the very same co-addition as with the observed data to generate stacks, hence reproducing all the observational effects affecting the reduced images. However, this dramatically increases computing time and is currently out-of-reach except for the shallower surveys.

Because the pipeline in this work makes it possible to constrain not only the galaxy luminosity evolution, but also the evolution of galaxy sizes, it opens interesting perspectives for addressing the current debate on the evolution of galaxy sizes with cosmic time. The contradictory results of, for example, Longhetti et al. (2007), Trujillo et al. (2007), Saracco et al. (2010), and Morishita et al. (2014) on the growth of massive early-type galaxies may be plagued with the varying selection effects in the surveys on which these analyses are based.

10. Conclusions

In the present article we lay the basis for a new method to infer robust constraints on galaxy evolution models. In this method, populations of synthetic galaxies are generated with the STUFF empirical model, sampled from luminosity functions for each galaxy type, and determined by the SEDs of the bulge and disk components, and the B/T ratio. In order to reproduce the selection effects affecting real catalogs, we use the SKYMAKER software to simulate realistic survey images with the appropriate instrumental characteristics. Real and mock images undergo the same source extraction, using SEXTRACTOR, and pre-processing pipeline. The distributions of extracted observables (fluxes and radii) are then compared, and we minimize their discrepancy using an adaptive MCMC sampling scheme in the parametric Bayesian indirect likelihood framework, designed for an efficient exploration of the parameter space.

This is the first attempt in the field of galaxy evolution to make image simulation a central part of the inference process. We have tested the self-consistency of this approach using a simulated image of a CFHTLS Deep field covering 1 deg^2 on the sky in three bands: u and i in the optical, and K_s in the near infrared, generated with the STUFF model containing E/S0 and spiral galaxies with evolving size and luminosity.

Starting from non-informative uniform priors, we find that our pipeline can reliably infer the input parameters governing the luminosity and size evolution of each of the galaxy populations in $\sim 10^4$ iterations, using few and disjointed observables, that is, the photometry (fluxes and radii) of the extracted sources in uiK_s. In each test performed, the input parameters lie within the 68% highest posterior density region.

We have also compared the results of our method with those of the classical photometric redshifts approach, with measurements from SED fitting on one of the mock sample, and found that when using the same set of observables (uiK_s photometry), our inference pipeline yields more accurate results.

Now that the validity of our pipeline is established on mock data, we intend to apply it to the observed CFHTLS Deep fields. We could also combine these data with several extragalactic surveys at various depths and with different instrumental setups simultaneously, such as UDF (Williams et al. 2010) at $z \sim 2$, and SDSS (Blanton et al. 2003) at $z \sim 0.1$, in order to better constrain galaxies in a wide range of redshifts. Nevertheless, this application will raise various modeling issues. In particular, real survey images display a continuum of galaxy populations, and our model only generates a discrete number of galaxy populations, defined by their bulge and disk SEDs and their B/T ratio. In practice, the number of modeled populations will be limited by computing time, as more populations lead to more free parameters to infer, hence to more iterations for the pipeline to find the high probability regions. This will certainly require a compromise between the desired accuracy of the modeled universe and convergence of the chains within a reasonable computing time.

Acknowledgements. The authors thank Erwan Cameron for his useful comments on this work. Based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/IRFU, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Science de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This work is based in part on data products produced at Terapix available at the Canadian Astronomy Data Centre as part of the Canada-France-Hawaii Telescope Legacy Survey, a collaborative project of NRC and CNRS. This work was partially supported by the ANR-13-BS05-002 SPIN(e) grant from the French Agence Nationale de la Recherche.

References

Adye, T. J. 1998, Ph.D. Thesis, Lincoln College, Oxford
Akeret, J., Refregier, A., Amara, A., Seehars, S., & Hasner, C. 2015, J. Cosmol. Astropart. Phys., 2015, 043
Andrieu, C., & Moulines, É. 2006, Am. App. Prob., 16, 1462
Arnaud, K. A. 1996, in Astronomical Data Analysis Software and Systems V, eds. G. H. Jacoby, & J. Barnes, ASP Conf. Ser., 101, 17
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Aslan, B., & Zech, G. 2002, ArXiv High Energy Physics – Experiment e-prints [arXiv:hep-ex/0203010]
Babu, G. J., & Feigelson, E. D. 2006, in Astronomical Data Analysis Software and Systems XV, eds. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, ASP Conf. Ser., 351, 127
Baillard, A., Bertin, E., de Lapparent, V., et al. 2011, A&A, 532, A74
Bamford, S. P., Nichol, R. C., Baldry, I. K., et al. 2009, MNRAS, 393, 1324
Beaumont, M. A. 2010, Annual Review of Ecology, Evolution, and Systematics, 41, 379
Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., & Robert, C. P. 2009, Biometrika, 96, 983
Bellman, R. 1972, Dynamic programming (Princeton University Press), 342
Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. 2017, ArXiv e-prints [arXiv:1701.05146]
Bertin, E. 2009, Mem. Soc. Astron. It., 80, 422
Bertin, E. 2011a, in Astronomical Data Analysis Software and Systems XX, eds. I. N. Evans, A. Accomazzi, D. J. Mink, & A. H. Rots, ASP Conf. Ser., 442, 435
Bertin, E. 2011b, STIFF: Converting Scientific FITS Images to TIFF, Astrophysics Source Code Library
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bhattacharyya, A. 1946, Sankhya: The Indian Journal of Statistics (1933–1960), 7, 401
Bienayme, O., Robin, A. C., & Creze, M. 1987, A&A, 180, 94
Binggeli, B., Sandage, A., & Tarenghi, M. 1984, AJ, 89, 64
Binggeli, B., Sandage, A., & Tammann, G. A. 1988, ARA&A, 26, 509
Blaizot, J., Wadadekar, Y., Guiderdoni, B., et al. 2005, MNRAS, 360, 159
Blanton, M. R., Hogg, D. W., Bahcall, N. A., et al. 2003, ApJ, 592, 819
Blanton, M. R., Lupton, R. H., Schlegel, D. J., et al. 2005, ApJ, 631, 208
Brooks, S., & Gelman, A. 1998, Computing Science and Statistics, 30
Cadez, I. V., Smyth, P., McLachlan, G. J., & McLaren, C. E. 2002, Machine Learning, 47, 7
Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582
Cameron, E., & Pettitt, A. N. 2012, MNRAS, 425, 44
Cash, W. 1979, ApJ, 228, 939
Chakravarti, M., Laha, R. G., & Roy, J. 1967, Handbook of Methods of Applied Statistics, Vol. I (John Wiley and Sons), 392
Chardy, P., Glemarec, M., & Laurec, A. 1976, Estuarine and Coastal Marine Science, 4, 179
Coleman, G. D., Wu, C.-C., & Weedman, D. W. 1980, ApJS, 43, 393
Condon, J. J. 1974, ApJ, 188, 279
Cowles, M. K., & Carlin, B. P. 1996, Source Journal of the American Statistical Association, 91, 883

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. 2010, Trends in Ecology & Evolution, 25, 410
Cuillandre, J.-C., & Bertin, E. 2006, in SF2A-2006: Semaine de l'Astrophysique Française, eds. D. Barret, F. Casoli, G. Lagache, A. Lecavelier, & L. Pagani, 265
Dalcanton, J. J., Spergel, D. N., Gunn, J. E., Schmidt, M., & Schneider, D. P. 1997, AJ, 114, 635
Davis, M., Faber, S. M., Newman, J., et al. 2003, in Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II, ed. P. Guhathakurta, Proc. SPIE, 4834, 161
de Jong, R. S., & Lacey, C. 2000, ApJ, 545, 781
de Lapparent, V., Galaz, G., Bardelli, S., & Arnouts, S. 2003, A&A, 404, 831
de Vaucouleurs, G. 1953, Astron. Soc. Pacific Leaflets, 6, 362
de Vaucouleurs, G. 1959, Handbuch der Physik, 53, 275
de Vaucouleurs, G., de Vaucouleurs, A., Corwin, Jr., H. G., et al. 1991, Third Reference Catalogue of Bright Galaxies, Vol. I: Explanations and references, Vol. II: Data for galaxies between 0^h and 12^h, Vol. III: Data for galaxies between 12^h and 24^h
Driver, S. P., & Phillipps, S. 1996, ApJ, 469, 529
Drovandi, C. C., Pettitt, A. N., & Lee, A. 2015, Stat. Sci., 30, 72
Eddington. 1913, MNRAS, 73, 359
Faber, S. M., Willmer, C. N. A., Wolf, C., et al. 2007, ApJ, 665, 265
Freedman, D., & Diaconis, P. 1981, Z. Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57, 453
Frei, Z., & Gunn, J. E. 1994, AJ, 108, 1476
Friedman, J. H. 1987, J. Am. Stat. Assoc., 82, 249
Gabasch, A., Bender, R., Seitz, S., et al. 2004, A&A, 421, 41
Gaia Collaboration (Brown, A. G. A., et al.) 2016, A&A, 595, A2
Gallant, A. R., & McCulloch, R. E. 2009, J. Am. Stat. Assoc., 104, 117
Gelman, A., & Rubin, D. B. 1992, Stat. Sci., 7, 457
Gutmann, M. U., & Corander, J. 2016, J. Mach. Learn. Res., 17, 1
Haario, H., Haario, H., Saksman, E., & Tamminen, J. 1999, Comput. Stat., 14, 1375
Haario, H., Saksman, E., & Tamminen, J. 2001, Bernoulli, 7, 223
Hahn, C., Vakili, M., Walsh, K., et al. 2017, MNRAS, 469, 2791
Hasinger, G., & Zamorani, G. 2000, in Exploring the Universe – A Festschrift in Honor of Ricardo Giacconi, Advanced Series in Astrophysics and Cosmology (Singapore: World Scientific Publishing Co. Pte. Ltd.), 119
Hastings. 1970, Biometrika, 57, 97
Helou, G., & Beichman, C. A. 1990, in Liege International Astrophysical Colloquia, 29, ed. B. Kaldeich
Hogg, D. W. 2008, ArXiv e-prints [arXiv:0807.4820]
Hogg, D. W., Baldry, I. K., Blanton, M. R., & Eisenstein, D. J. 2002, ArXiv e-prints [arXiv:astro-ph/0210394]
Hudelot, P., Goranova, Y., Yannick Mellier, Y., et al. 2012, T0007: The Final CFHTLS Release
Hyvärinen, A., Hurri, J., & Hoyer, P. O. 2009, in Natural Image Statistics, 93
Ishida, E. E. O., Vitenti, S. D. P., Penna-Lima, M., et al. 2015, Astron. Comput., 13, 1
Jennings, E., & Madigan, M. 2017, Astronomy and Computing, 19, 16
Jester, S., Schneider, D. P., Richards, G. T., et al. 2005, AJ, 130, 873
Justel, A., Peña, D., & Zamar, R. 1997, Statistics & Probability Letters, 35, 251
Kangasrääsiö, A., Lintusaari, J., Skytén, K., et al. 2016, in NIPS 2016 Workshop on Advances in Approximate Bayesian Inference
Kautsch, S. J., Grebel, E. K., Barazza, F. D., & Gallagher, J. S. 2006, A&A, 445, 765
Kessy, A., Lewin, A., & Strimmer, K. 2017, The American Statistician, in press, DOI: 10.1080/00031305.2016.1277159
Kirkpatrick, C. D. Gelatt, M. P. V. 1983, Science, 220, 671
Knuth, K. H. 2006, ArXiv Physics e-prints [arXiv:physics/0605197]
Kron, R. G. 1980, ApJS, 43, 305
Kullback, S., & Leibler, R. A. 1951, The Annals of Mathematical Statistics, 22, 79
Kurinsky, N., & Sajina, A. 2014, in Statistical Challenges in 21st Century Cosmology, eds. A. Heavens, J.-L. Starck, & A. Krone-Martins, IAU Symp., 306, 295
Le Borgne, D., & Rocca-Volmerange, B. 2002, A&A, 386, 446
Lilly, S. J., Tresse, L., Hammer, F., Crampton, D., & Le Fevre, O. 1995, ApJ, 455, 108
Lin, J. 1991, IEEE Transactions on Information Theory, 37
Loaring, N. S., Dwelly, T., Page, M. J., et al. 2005, MNRAS, 362, 1371
Longhetti, M., Saracco, P., Severgnini, P., et al. 2007, MNRAS, 374, 614
MacDonald, C. J., & Bernstein, G. 2010, PASP, 122, 485
Madau, P. 1995, ApJ, 441, 18
Madau, P., Ferguson, H. C., Dickinson, M. E., et al. 1996, MNRAS, 283, 1388
Madgwick, D. S., Hawkins, E., Lahav, O., et al. 2003, MNRAS, 344, 847
Mahalanobis, P. C. 1936, in Proceedings National Institute of Science, India, 2, 49
Malmquist. 1920, Meddelanden fran Lunds Astronomiska Observatorium Series II, 22, 3
Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. 2011, ArXiv e-prints [arXiv:1101.0955]
Marjoram, P., Molitor, J., Plagnol, V., & Tavare, S. 2003, Proceedings of the National Academy of Sciences of the United States of America, 100, 15324
Marzke, R. O. 1998, The Evolving Universe, 231, 23
Mehrotra, K., Mohan, C. K., & Ranka, S. 1997, Elements of artificial neural networks (MIT Press), 344
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087
Mo, H. J., Mao, S., & White, S. D. M. 1998, MNRAS, 295, 319
Morishita, T., Ichikawa, T., & Kajisawa, M. 2014, ApJ, 785, 18
Norberg, P., Cole, S., Baugh, C. M., et al. 2002, MNRAS, 336, 907
Overzier, R., Lemson, G., Angulo, R. E., et al. 2013, MNRAS, 428, 778
Palombo, G. 2011, ArXiv e-prints [arXiv:1102.2407]
Pardo, L., & Menéndez, M. L. 2006, Metrika, 64, 63
Patil, A., Huard, D., & Fonnesbeck, C. J. 2010, J. Stat. Software, 35, 1
Pearson, C. P., Serjeant, S., Oyabu, S., et al. 2014, MNRAS, 444, 846
Pritchard, J. K., Seielstad, M. T., & Perez-Lezaun, A. 1999, Mol. Biol. Evol., 16, 179
Racine, R. 1996, PASP, 108, 699
Ramos, B. H. F., Pellegrini, P. S., Benoist, C., et al. 2011, AJ, 142, 41
Reeves, R., & Pettitt, A. 2005, in Proc. 20th Int. Works. Stat. Mod. Australia, eds. A. R. Francis, K. M. Matawie, A. Oshlack, G. K. Smyth, 393
Roberts, G. O., & Rosenthal, J. S. 2009, J. Computational and Graphical Statistics, 18, 349
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Robin, A. C., Reylé, C., Fliri, J., et al. 2014, A&A, 569, A13
Rubner, Y., Tomasi, C., & Guibas, L. 1998, in Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271) (Narosa Publishing House), 59
Rybizki, J., & Just, A. 2015, MNRAS, 447, 3880
Sandage, A., Freeman, K. C., & Stokes, N. R. 1970, ApJ, 160, 831
Saracco, P., Longhetti, M., & Gargiulo, A. 2010, MNRAS, 408, L21
Scargle, J. D., Norris, J. P., Jackson, B., & Chiang, J. 2013, ApJ, 764, 167
Schafer, C. M., & Freeman, P. E. 2012, in Statistical Challenges in Modern Astronomy V, 3
Schechter, P. 1976, ApJ, 203, 297
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Sérsic, J. L. 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41
Sheth, R. K. 2007, MNRAS, 378, 709
Shlens, J. 2014, ArXiv e-prints [arXiv:1404.1100]
Singal, A. K., & Rajpurohit, K. 2014, MNRAS, 442, 1656
Spergel, D. N., Verde, L., Peiris, H. V., et al. 2003, ApJS, 148, 175
Steinmetz, M., & Bartelmann, M. 1995, MNRAS, 272, 570
Stephens, M. A. 1974, J. Am. Stat. Assoc., 69, 730
Sunnaker, M., Busetto, A. G., Numminen, E., et al. 2013, PLoS Comput. Biol., 9, 1
Taghizadeh-Popp, M., Fall, S. M., White, R. L., & Szalay, A. S. 2015, ApJ, 801, 14
Tolman, R., & Richard. 1934, Science, 80, 358
Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. 2009, J. Roy. Soc. Interf., 6
Trujillo, I., Förster Schreiber, N. M., Rudnick, G., et al. 2006, ApJ, 650, 18
Trujillo, I., Conselice, C. J., Bundy, K., et al. 2007, MNRAS, 382, 109
Turner, B. M., & Van Zandt, T. 2012, J. Math. Psych., 56, 69
Warren, M. S., Quinn, P. J., Salmon, J. K., & Zurek, W. H. 1992, ApJ, 399, 405
Weyant, A., Schafer, C., & Wood-Vasey, W. M. 2013, ApJ, 764, 116
Williams, R. J., Quadri, R. F., Franx, M., et al. 2010, ApJ, 713, 738
Willmer, C. N. A. 1997, AJ, 114, 898
Wolberg, G., & George. 1990, Digital image warping (IEEE Computer Society Press), 318
Wolf, C., Dye, S., Kleinheinrich, M., et al. 2001, A&A, 377, 442
Wolf, C., Meisenheimer, K., Rix, H.-W., et al. 2003, A&A, 401, 73
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579
Zehavi, I., Blanton, M. R., Frieman, J. A., et al. 2002, ApJ, 571, 172
Zucca, E., Ilbert, O., Bardelli, S., et al. 2006, A&A, 455, 879

Appendix A: Derivation of the auxiliary likelihood function used in the present article

If one assumes that the number count in each bin i is described by a Poisson distribution, the probability of o_i given the model s_i is:

$$l_i = \frac{\mathrm{e}^{-s_i} s_i^{o_i}}{o_i!}. \tag{A.1}$$

The likelihood function for the histogram is then:

$$L = \prod_{i=1}^{b} \frac{\mathrm{e}^{-s_i} s_i^{o_i}}{o_i!}. \tag{A.2}$$

Correlations between adjacent bins are neglected here. The log-likelihood is therefore given by:

$$\ln L = \sum_{i=1}^{b} (-s_i + o_i \ln(s_i) - \ln(o_i!)). \tag{A.3}$$

As we are interested in maximizing $\ln L$, $\ln(o_i!)$ is a constant that can be eliminated, so in fine, we obtain Eq. (20).

Appendix B: Conversion LF parameters from Faber et al. (2007) to STUFF parameters

In order to provide STUFF with realistic LF parameters, we use Faber et al. (2007), who used data from SDSS (York et al. 2000; Blanton et al. 2003), COMBO-17 (Wolf et al. 2001; Wolf et al. 2003), 2dF (Norberg et al. 2002), and DEEP2 (Davis et al. 2003) to derive the evolving LF parameters for two populations of red and blue galaxies. We associate the red and blue populations with our populations of E/S0 and spirals respectively. The LF parameters found by Faber et al. (2007) are listed in Table B.1, and apply to $z = 0.5$. In order to obtain the LF parameters for $z = 0$, we use the fitted functions provided by Faber et al. (2007) for each population:

$$M_B^*(z=0) = M_B^*(z) - \frac{Q \log_{10}(1+z)}{\log_{10}(2)} \tag{B.1}$$

$$\log_{10} \phi^*(z=0) = \log_{10} \phi^*(z) - \frac{P \log_{10}(1+z)}{\log_{10}(2)}. \tag{B.2}$$

The absolute magnitude in Eq. (B.1) is given in the Johnson system. Because in our simulation STUFF operates in the AB system, we use the AB offset calculated by Frei & Gunn (1994):

$$B_{AB} = B_{\mathrm{Johnson}} - 0.163. \tag{B.3}$$

We then apply the transformation equations of Jester et al. (2005) for stars with $R_c - I_c < 1.15$ and $U - B > 0$,

$$B_{AB} = g + 0.39(g - r) + 0.21, \tag{B.4}$$

in order to derive g-band magnitudes:

$$M^*(z=0)_g = M_B^*(z=0) - 0.39(g-r) - 0.21 - 0.163. \tag{B.5}$$

We subsequently adopt average colors of $(g-r)_{\mathrm{E/S0}} = 0.75$ and $(g-r)_{\mathrm{Sp}} = 0.5$ from EFIGI data (de Lapparent, priv. comm.) to derive the value of $M^*(z=0)_g$ for each population.

In STUFF, the input LF parameters are provided assuming $H_0 = 100\ h$ km s^{-1} Mpc^{-3} with $h = 1$. As Faber et al. (2007) provide their results assuming $h = 0.7$, an additional conversion is needed:

$$M^*_{\mathrm{STUFF}} = M^*(z=0)_g - 5 \log_{10} h \tag{B.6}$$

$$\phi^*_{\mathrm{STUFF}} = \phi^* h^{-3}. \tag{B.7}$$

In STUFF, the LF evolution parameters are defined as:

$$M^*(z) = M^*(z=0) + M_e \ln(1+z) \tag{B.8}$$

$$\log_{10} \phi^*(z) = \log_{10} \phi^*(z=0) + \phi_e \log_{10}(1+z). \tag{B.9}$$

Combining Eqs. (B.1) and (B.2) with Eqs. (B.8) and (B.9) respectively, we obtain:

$$M_e = \frac{Q}{\ln(10) \log_{10}(2)} \tag{B.10}$$

$$\phi_e = \frac{P}{\log_{10}(2)}. \tag{B.11}$$

The values of P and Q listed in Table B.1 are used to derive the LF parameters for the populations of E/S0 and Sp. In fine, the $\phi^*(z=0)$ of each population is reduced by a factor ten to limit computation time. The final LF parameters are listed in Table 4 (Sect. 7.4).

Table B.1. Luminosity function parameters of the blue and red populations of galaxies at $z = 0.5$ inferred from SDSS, 2dF, COMBO-17, and DEEP2 data, adapted from Tables 3, 4, and 6 of Faber et al. (2007).

Population	$M_B^*(z=0.5)$	$\log_{10}(\phi^*[\mathrm{Mpc}^{-3}])(z=0.5)$	P	Q	α
Red	–20.80	–2.72	–0.46	–1.23	–0.5
Blue	–20.84	–2.55	0.01	–1.35	–1.3

Notes. The redshift evolution is fitted by $y = a_0(z=0.5) + a_1[\log_{10}(1+z) - \log_{10}(1+0.5)]/\log_{10}(2)$, where M_B^* and $\log_{10}(\phi^*)$ are the zero points and Q and P are the slopes resp. The LF parameters in Faber et al. (2007) are given for $H_0 = 70$ km s^{-1} Mpc^{-1}